

UNIVERSITÄT HAMBURG

Fakultät für Erziehungswissenschaft, Psychologie und Bewegungswissenschaft
Fachbereich Psychologie

Gewissenhaftigkeit und Ausbildungserfolg: Wie Fleiß, Ordnung und Selbstwirksamkeitsüberzeugungen IHK-Noten beeinflussen

Dissertation

zur Erlangung des akademischen Grades
Doktor der Philosophie
im Fachbereich Psychologie der Universität Hamburg.

vorgelegt von:

Kirsten Dreier

aus Pinneberg

Hamburg, 2012

**Angenommen vom
Fachbereich der Psychologie der Universität Hamburg am: 13.02.2013**

**Veröffentlicht mit Genehmigung des
Fachbereichs Psychologie der Universität Hamburg.**

**Vorsitzender des
Prüfungsausschusses: Prof. Dr. Erich H. Witte**

Erstgutachter: Prof. Dr. Burghard Andresen

Zweitgutachter: Prof. Dr. André Beauducel

Danksagung

Bei der Durchführung der vorliegenden Untersuchung und der Fertigstellung dieser Arbeit wurde ich durch eine Reihe von Menschen bzw. Institutionen unterstützt, denen ich an dieser Stelle meinen Dank aussprechen möchte.

An erster Stelle möchte ich meinem Doktorvater und Betreuer Prof. Dr. Burghard Andresen für die Freiheiten bei der Durchführung und Erstellung meiner Arbeit, den jederzeit freundlichen, hilfreichen sowie praktischen Rat und interessante Fachgespräche im Verlaufe meines Dissertationsprozesses danken. Mein Dank gilt auch dem Zweitgutachter dieser Arbeit – Prof. Dr. André Beauducel.

Die Arbeit entstand im Rahmen meiner Tätigkeit als Assistenz der Geschäftsführung und Wissenschaftlicher Mitarbeiterin bei Moldzio & Partner – Institut für Personalauswahl. Ich möchte dem Geschäftsführenden Partner Thomas Moldzio für seine Mithilfe bei der unternehmerischen Akquise und Betreuung des von mir untersuchten Kundenprojektes danken. Natürlich gilt mein Dank auch den ehemaligen Kollegen: Matthias Mickleit danke ich für hilfreiche Anregungen und die wertvolle Unterstützung bei der Erstellung der vorliegenden Arbeit. Marion Wende bin ich für ihre ausdauernde, sorgfältige und launige Assistenz bei der Dateneingabe sowie ihre Mithilfe bei der Datenerhebung zu Dank verpflichtet. Aus dem Kreise der ehemaligen Kollegen danke ich außerdem allen Praktikantinnen und Praktikanten von Moldzio & Partner, die mich oft unterstützt haben, damit ich an der Dissertation weiter arbeiten konnte.

Ein Dank geht auch an alle Teilnehmer meiner Untersuchung, ohne deren Bereitschaft – teilweise auch mehrfach – an meiner Untersuchung teilzunehmen, eine Realisierung des Projekts nicht möglich gewesen wäre. Bei der Durchführung der Untersuchung konnte ich mich darüber hinaus auf die Unterstützung der Ausbildungsleiter sowie Lehrerinnen und Lehrer der von mir befragten Auszubildenden sowie Schulabsolventen verlassen, die ich aus Gründen der Anonymität allerdings nicht namentlich nennen kann.

Danken möchte ich auch meinen guten Freunden Moritz Godau und Ines Hinrichs, auf deren moralische und auch praktische Unterstützung ich in der Entstehungsphase dieser Arbeit stets setzen konnte. Besonders Joelle Vollertsen bin ich zu großem Dank für wertvolle methodische Hinweise, praktisch-hilfreiche Ratschläge sowie ihre Ausdauer und ihr Interesse an den Ergebnissen dieser Arbeit verpflichtet.

Nicht zuletzt möchte ich meiner Familie – allen voran meinen Eltern Christa und Peter Dreier – ganz herzlich dafür danken, dass ich mich immer und in allen Lebenslagen auf sie verlassen konnte und kann.

Kirsten Dreier

Inhaltsverzeichnis

Abbildungsverzeichnis	ix
Tabellenverzeichnis.....	xi
Abkürzungsverzeichnis	xvi
1. Einleitung	1
1.1 Nutzen der Arbeit	3
1.2 Struktur der Arbeit	4
2. Theoretischer und empirischer Hintergrund	6
2.1 Berufliche Eignungsdiagnostik	6
2.2 Allgemeine kognitive Fähigkeiten in der Eignungsdiagnostik	11
2.3 Persönlichkeitsmerkmale in der Eignungsdiagnostik	13
2.3.1 Gewissenhaftigkeit	25
2.3.2 Selbstwirksamkeit	36
2.3.3 Wahrgenommene eigene Fähigkeiten	43
2.3.4 Zur Abgrenzung der Selbsteinschätzungsstrukture	45
2.4 Kritische Bewertung von Persönlichkeitsmaßen in der Eignungsdiagnostik.....	46
2.4.1 Vorteile.....	46
2.4.2 Nachteile	48
2.4.3 Sozial erwünschtes Antwortverhalten.....	52
2.5 Berufsbezogenes Testen in der Eignungsdiagnostik.....	58
2.6 Vorliegende Instrumente zur Erhebung von Gewissenhaftigkeit.....	66
2.6.1 NEO-Persönlichkeitsinventare (NEO-PI-R und NEO-FFI).....	67
2.6.2 Personality Research Form (PRF).....	68
2.6.3 Big Five Aspect Scales (BFAS).....	69
2.6.4 Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP)	71
2.6.5 Freiburger Persönlichkeitsinventar (FPI).....	72
2.6.6 Big Five Inventory (BFI)	73
2.6.7 START-P.....	75
2.6.8 Zusammenfassung und Fazit zu den Gewissenhaftigkeitsmaßen	76
2.7 Leistungs- und Erfolgskriterien in der Eignungsdiagnostik.....	77
2.7.1 Leistung im Beruf	80
2.7.2 Schulleistung	82
2.7.3 Berufsschulleistung	90
2.8 Theoretische Vorhersagemodelle in der Eignungsdiagnostik	96
2.8.1 Existierende Vorhersagemodelle	96
2.8.2 Modellbildung zur Vorhersage von Ausbildungserfolg.....	100
3. Forschungsfragen	106

3.1 Ableitung der Ziele und inhaltlichen Hypothesen.....	106
3.1.1 Validierung eines kontextbezogenen Gewissenhaftigkeitsmaßes.....	106
3.1.2 Entwicklung und Überprüfung eines Vorhersagemodells	108
3.1.3 Gegenüberstellung eines allgemeinen Gewissenhaftigkeitsmaßes mit einer kontextbezogenen Gewissenhaftigkeitsskala	110
3.2 Mögliche konfundierende Variablen.....	111
3.2.1 Demografische Merkmale	112
3.2.1.2 Alter.....	112
3.2.1.2 Geschlecht	114
3.2.1.3 Schulbildung	114
3.2.2 Berufliche Informationen	115
3.2.3 Sozial Erwünschtes Antwortverhalten	115
4. Entwicklung der AGS	118
4.1 Definition der Konstrukte Ordnung und Fleiß	118
4.2 Testkonstruktion.....	119
4.2.1 Allgemeine Vorüberlegungen zur Testkonstruktion	119
4.2.2 Vorüberlegungen zur Itemauswahl	121
4.2.3 Itemgenerierung	123
4.2.4 Itemauswahl	125
4.2.5 Antwortformat.....	126
4.2.5.1 Anzahl der Stufen.....	126
4.2.5.2 Benennung der Pole/ Stufen.....	127
4.2.6 Instruktionen	127
4.3 Voruntersuchungen	127
4.3.1 Skalenüberprüfung	128
4.3.1.1 Überprüfung der deutschsprachigen AGS-Vorversion	128
4.3.1.2 Überprüfung der englischsprachigen AGS-Vorversion	129
4.3.2 Zusammenhänge mit weiteren psychologischen Konstrukten	129
4.3.2.1 Soziale Erwünschtheit.....	129
4.3.2.2 Selbstwirksamkeit und wahrgenommene eigene Fähigkeiten	130
4.3.2.3 Karriereinteressen	131
4.3.2.4 Copingstrategien	133
4.3.2.5 Selbstführungsstrategien	134
4.3.2.6 Schulische Leistungskriterien	134
4.3.2.7 Berufliche Leistungskriterien.....	135
4.4 Itemreanalyse	135
4.5 Beispielitems	137
5. Untersuchungsplanung/ Operationalisierung	138

5.1 Operationalisierung der Variablen auf Prädiktorebene	138
5.1.1 Gewissenhaftigkeit	138
5.1.2 Selbstwirksamkeit	139
5.1.3 Wahrgenommene eigene Fähigkeiten	140
5.1.4 Allgemeine kognitive Fähigkeiten	141
5.1.4.1 Leistungsprüfsystem	141
5.1.4.2 Intelligenz-Struktur-Test	143
5.2 Operationalisierung der Kriteriumsvariablen	144
5.2.1 Zeugnisnoten	144
5.2.2 Berufsschul- und IHK-Noten	144
5.2.3 Vorgesetztenbeurteilung	145
5.3 Erhebung weiterer Variablen	146
5.3.1 Demografische Merkmale der Untersuchungsteilnehmer.....	146
5.3.2 Kontrollvariablen	146
5.3.2.1 Sozial erwünschtes Antwortverhalten.....	146
5.3.2.2 Berufliche Vorerfahrungen	147
5.3.2.3 Weitere Einflüsse	147
6. Statistische Auswertung der Daten.....	149
6.1 Analyse und Behandlung fehlender Werte.....	149
6.1.1 Analyse der Fehlendmechanismen.....	149
6.1.2 Elimination von Fällen mit zu hohem Anteil fehlender Werte	150
6.1.3 Mittelwertbasierte Ersetzung fehlender Werte.....	150
6.2 Überprüfung der Skaleneigenschaften der AGS	151
6.2.1 Analysen auf Itemebene	151
6.2.1.1 Deskriptiva	151
6.2.1.2 Itemschwierigkeiten	152
6.2.1.3 Trennschärfen.....	152
6.2.2 Messgüte der Gewissenhaftigkeitsaspekte Fleiß und Ordnung.....	153
6.2.2.1 Reliabilitätsschätzungen der Subskalen	153
6.2.2.2 Retest-Reliabilität.....	154
6.2.2.3 Skalenhomogenität der Subskalen Fleiß und Ordnung.....	156
6.2.3 Konstruktvalidierung der AGS	160
6.3 Vorhersage von Ausbildungs- bzw. Schulleistung	162
6.3.1 Korrelationsanalysen.....	163
6.3.2 Kriterien der Dateneignung.....	164
6.3.2.1 Range Restriction	164
6.3.2.2 Minderungskorrektur.....	165
6.3.3 Regressionsanalysen	166

6.3.4 Moderationsanalysen.....	167
6.3.5 Mediationsanalysen.....	168
6.4 Vergleich der AGS mit der Gewissenhaftigkeitsskala des NEO-FFI	170
6.5 Weitere angewandte Testverfahren	170
6.5.1 t-Tests.....	170
6.5.2 Levene-Test.....	171
6.5.3 Tests auf Normalverteilung.....	171
6.5.4 Multikollinearitätsanalysen	171
6.5.5 Teststärke	172
6.6 Nicht angewandte Testverfahren.....	172
7. Durchführung	173
7.1 Bewerberstichprobe.....	173
7.1.1 Rekrutierung der Bewerberstichprobe	173
7.1.2 Beschreibung der Bewerberstichprobe	173
7.1.3 Ablauf der Bewerbertestungen.....	175
7.1.4 Probleme bei der Rekrutierung und Testdurchführung.....	176
7.2 Retest-Stichprobe	176
7.2.2 Beschreibung der Retest-Stichprobe	176
7.2.3 Ablauf der Testung.....	178
7.2.4 Probleme bei der Rekrutierung, Testdurchführung und -auswertung.....	179
7.3 Kontrollgruppe	179
7.3.1 Rekrutierung der Kontrollgruppe	179
7.3.2 Beschreibung der Kontrollgruppe	180
7.3.3 Ablauf der Kontrollgruppentestung	181
7.3.4 Probleme bei der Rekrutierung und Testdurchführung.....	182
8. Ergebnisse	184
8.1 Vorbereitende Datenanalysen	184
8.1.1 Analyse und Behandlung fehlender Werte	184
8.1.1.1 Elimination von Fällen mit zu hohem Anteil fehlender Werte	184
8.1.1.2 Analyse des Fehlendmechanismus.....	186
8.1.1.3 Ersetzungen fehlender Werte	186
8.2 Überprüfung der Skaleneigenschaften der AGS	186
8.2.1 Analysen auf Itemebene	186
8.2.1.1 Deskriptiva	186
8.2.1.2 Itemschwierigkeiten	189
8.2.1.3 Trennschärfen.....	190
8.2.2 Messgüte der Subskalen Fleiß und Ordnung	191
8.2.2.1 Reliabilitätsschätzungen der Subskalen	191

8.2.2.2 Retest-Reliabilität.....	191
8.2.2.3 Skalenhomogenität der Subskalen Fleiß und Ordnung.....	194
8.2.3 Konstruktvalidierung der AGS	197
8.3 Vorhersage von Schul- bzw. Berufsschulnoten	200
8.3.1 Überprüfung der Voraussetzungen der Dateneignung.....	200
8.3.1.1 Normalverteilung der Daten.....	200
8.3.1.2 Varianzhomogenität	201
8.3.1.3 Stichprobengröße	201
8.3.1.4 Multikollinearität.....	201
8.3.2 Ergebnisse der Korrelationsanalysen	201
8.3.3 Ergebnisse der Moderationsanalysen zur Modellüberprüfung.....	209
8.3.4 Ergebnisse der Mediationsanalysen zur Modellüberprüfung.....	217
8.4 Inkrementelle Validität.....	224
8.4.1 Vergleich der Vorhersagekraft der AGS mit dem NEO-FFI	225
8.4.2 Inkrementelle Validität der AGS	225
8.4.3 Intrapersonelle Antwortvariabilität	229
8.5 Ergebnisse der Überprüfung der Kontrollvariablen	230
8.5.1 Alter.....	231
8.5.2 Geschlecht.....	232
8.5.3 Schulbildung	233
8.5.4 Berufliche Erfahrungen	234
8.5.5 Sozial erwünschtes Antwortverhalten.....	235
8.6 Zusammenfassung der Befunde	236
9. Diskussion	238
9.1 Bewertung des methodischen Vorgehens	238
9.1.1 Interne Validität der Untersuchung.....	238
9.1.1.1 Drop-out	238
9.1.1.2 Reifung.....	239
9.1.1.3 Zeiteinflüsse	239
9.1.1.4 Testeffekte.....	240
9.1.1.5 Statistische Regression.....	240
9.1.2 Externe Validität der Untersuchung.....	240
9.1.2.1 Repräsentativität der Stichprobe	241
9.1.2.2 Operationalisierung	242
9.1.3 Statistische Validität der Untersuchung	242
9.1.4 Zusammenfassende Bewertung der Validität der Untersuchung	244
9.2 Diskussion der Befunde	244
9.2.1 Skalenüberprüfung der AGS	244

9.2.1.1 Interne Konsistenz der Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS) ...	244
9.2.1.2 Zeitliche Stabilität der Arbeitsbezogenen Gewissenhaftigkeitsskalen.....	245
9.2.1.3 Überprüfung des latenten Konstrukts der AGS	249
9.2.1.4 Überprüfung der Konstruktvalidität der AGS	250
9.2.2 Überprüfung der Modellpassung.....	251
9.2.2.1 Vorhersagekraft von Fleiß, Ordnung, Selbstwirksamkeit und wahrgenommenen eigenen Fähigkeiten.....	251
9.2.2.2 Überprüfung des Vorhersagemodells.....	256
9.2.3 Vergleich der AGS mit einem allgemeinen Gewissenhaftigkeitsmaß	259
9.2.3.1 Vergleiche zur Vorhersagekraft der verschiedenen Gewissenhaftigkeitsmaße ..	259
9.2.3.2 Vergleich der intra- und interpersonellen Antwortvariabilitäten	261
9.2.4 Überprüfung der Kontrollvariablen.....	262
9.3 Zusammenfassende Bewertung der Ergebnisse	264
9.4 Ausblick und Implikationen	265
9.4.1 Implikationen für die Forschung	265
9.4.2 Implikationen für die Praxis	272
9.4.3 Fazit.....	275
10. Literaturverzeichnis.....	276
Zusammenfassung der Untersuchung	295
Summary	297
Appendix	299
A1. Ergebnisse der MVA und Mittelwertimputation	299
A2. Ergebnisse der Konsistenzanalysen.....	303
A3. Ergebnisse der Multikollinearitätsüberprüfung der AGS.....	304
A4. Ergebnisse der konfirmatorischen Faktorenanalyse zur Skalenüberprüfung der AGS	307
A5. Weitere Ergebnisse zur Überprüfung des Vorhersagemodells.....	312
B1. Anschreiben an die Retest Stichprobe	314
B2. Beispiel-Anschreiben zur Rekrutierung der Kontrollstichprobe	315
B3. Weitergehende Informationen zur Rekrutierung der Kontrollstichprobe (2 Seiten)...	316
B4. Instruktionen der Kontrollgruppentestung.....	318
B5. Codierungsliste	319

Abbildungsverzeichnis

Abbildung 2-1	Die Ansätze der Eignungsdiagnostik und ihre Validierungslogik (Schuler & Höft, 2006, S. 103)	S. 8
Abbildung 2-2	Hierarchisches Modell der Persönlichkeit (Quelle: DeYoung, 2012)	S. 18
Abbildung 2-3	Grund- und Globaldimensionen des START-P (nach Beauducel & Kersting, 2010)	S. 76
Abbildung 2-4	Verlauf der operationalen Validitäten der Gewissenhaftigkeits-skala und -facetten zur Vorhersage von Studienerfolg über einen Zeitraum von sieben Jahren (nach Lievens et al., 2009)	S. 94
Abbildung 2-5	Schematisches Modell zur Vorhersage von Ausbildungsleistung	S. 101
Abbildung 3-1	Konzeptuelles Modell zur Vorhersage von Ausbildungserfolg.	S. 109
Abbildung 6-1	Schematische Darstellung des Messmodells zur Prüfung der Skalenhomogenität des Gewissenhaftigkeitsaspekts Fleiß bzw. Ordnung.	S. 157
Abbildung 6-2	Schema des zweifaktoriellen Modells, das die latenten Variablen Fleiß und Ordnung enthält.	S. 159
Abbildung 6-3	Schematische Darstellung des Modells zur Überprüfung der konvergenten Validität mit der Gewissenhaftigkeitsdimension aus dem NEO-FFI.	S. 162
Abbildung 6-4	Darstellung möglicher Moderatorbeziehung/en im Untersuchungsmodell	S. 167
Abbildung 6-5	Darstellung möglicher Mediationsbeziehung/en im Untersuchungsmodell	S. 169
Abbildung 7-1	Vergleich der Altersverteilung der kaufmännischen und technischen Bewerber	S. 174
Abbildung 7-2	Vergleich der Altersverteilung der kaufmännischen und technischen Auszubildenden (Retest-Stichprobe)	S. 177
Abbildung 7-3	Vergleich der Altersverteilung in der Kontrollgruppe	S. 180
Abbildung 8-1	Häufigkeitsverteilung der Skalenmittelwerte der Fleißskala für die Bewerberstichprobe ($N_{Bew}=740$) und die Kontrollgruppe ($N_{KG}=97$)	S. 188
Abbildung 8-2	Häufigkeitsverteilung der Skalenmittelwerte der Ordnungsskala für die Bewerberstichprobe ($N_{Bew}=740$) und die Kontrollgruppe ($N_{KG}=97$)	S. 188
Abbildung 8-3	Streudiagramm der beiden Fleiß-Messungen (X-Achse: T1; Y-Achse: T2).	S. 192
Abbildung 8-4	Streudiagramm der beiden Ordnungs-Messungen (X-Achse: T1; Y-Achse: T2).	S. 193
Abbildung 8-5	Streudiagramm der beiden Selbstwirksamkeits-Messungen (X-Achse: T1; Y-Achse: T2).	S. 193
Abbildung A-1	AMOS-Ausgabe des einfaktoriellen CFA-Modells des Fleißaspekts für die Bewerberstichprobe ($N_{Bew}=740$).	S. 307
Abbildung A-2	AMOS-Ausgabe des einfaktoriellen CFA-Modells des Ordnungsaspekts für die Bewerberstichprobe ($N_{Bew}=740$).	S. 307

Abbildung A-3	AMOS-Ausgabe des einfaktoriellen CFA-Modells des Fleißaspekts für die Kontrollgruppe ($N_{KG}=97$).	S. 308
Abbildung A-4	AMOS-Ausgabe des einfaktoriellen CFA-Modells des Ordnungsaspekts für die Kontrollgruppe ($N_{KG}=97$).	S. 308

Tabellenverzeichnis

Tabelle 2-1	Anwendungshäufigkeit und prädiktive Validität von Personalauswahlverfahren	S. 10
Tabelle 2-2	Einsatzhäufigkeit von Auswahlverfahren für Auszubildende (Auszug aus Schuler et al., 2007)	S. 22
Tabelle 2-3	Verwendungshäufigkeit von Auswahlverfahren in deutschen KMU bzw. Konzernen (aus Nachtwei & Schermuly, 2009)	S. 23
Tabelle 2-4	Vorbehalte von Personalern gegenüber Eignungstests (aus Nachtwei & Schermuly, 2009)	S. 24
Tabelle 2-5	Kriteriumsvaliditäten und interne Konsistenzen von Gewissenhaftigkeit bzw. deren Facetten (aus Schmit et al., 1995)	S. 28
Tabelle 2-6	Ladungen der Gewissenhaftigkeitsfacetten auf die Gewissenhaftigkeitsdimension (nach Lord, 2011)	S. 32
Tabelle 2-7	Interkorrelation der Gewissenhaftigkeitsfacetten (nach Lord, 2011)	S. 33
Tabelle 2-8	Ergebnisse zum Vergleich der generischen vs. kontextualisierten Gewissenhaftigkeitsdimension	S. 61
Tabelle 2-9	Interne Konsistenzen und Kriteriumsvaliditäten unterteilt nach Referenzrahmen (nach Lievens et al., 2008)	S. 64
Tabelle 2-10	Deskriptive Statistiken der BFAS (nach DeYoung et al., 2007)	S. 70
Tabelle 2-11	Kennwerte der Skalen des BFI und K-BFI (nach Lang & Lüdtke, 2005)	S. 74
Tabelle 2-12	Prädiktive und inkrementelle Validität verschiedener Auswahlmethoden bzw. -variablen zur Vorhersage von beruflicher Leistung (nach Schmidt & Hunter, 1998)	S. 81
Tabelle 2-13	Zusammenhänge zwischen den Big 5 und Intelligenz mit GPA (nach Poropat, 2009)	S. 86
Tabelle 2-14	Zusammenhänge zwischen Gewissenhaftigkeit bzw. deren Facetten und GPA (nach Nofle & Robins, 2007)	S. 89
Tabelle 2-15	Prädiktive und inkrementelle Validität verschiedener Auswahlmethoden bzw. -variablen zur Vorhersage von beruflichem Lernerfolg (nach Schmidt & Hunter, 1998)	S. 93
Tabelle 3-1	Inhaltliche Hypothesen zur Skalengüte der Arbeitsbezogenen Gewissenhaftigkeitsskalen	S. 107
Tabelle 3-2	Inhaltliche Hypothesen zur Überprüfung des Vorhersagemodells	S. 109
Tabelle 3-3	Inhaltliche Hypothesen zum Vergleich der AGS mit einem allgemeinen, generischen Gewissenhaftigkeitsmaß	S. 111
Tabelle 4-1	Ergebnisse der Skalenüberprüfung der AGS-Vorversion in Deutschland (nach Struckmeier, 2009)	S. 128
Tabelle 4-2	Ergebnisse der Skalenüberprüfung der englischen Vorversion der AGS (nach Dreier, 2008)	S. 129
Tabelle 4-3	Zusammenhänge der AGS mit Selbstwirksamkeit und wahrgenommenen eigenen Fähigkeiten (nach Dreier, 2008)	S. 130
Tabelle 4-4	Zusammenhänge der AGS mit den Karriereankern (nach Dreier, Moldzio & Kasper, 2010)	S. 133

Tabelle 4-5	Korrelationen der AGS mit Selbstführungsstrategien (nach Struckmeier, 2009)	S. 134
Tabelle 5-1	Beispiel-Items für die Skalen des NEO-FFI (nach Borkenau & Ostendorf, 1993, 2008)	S. 139
Tabelle 5-2	Notenspiegel/ Bewertungsschlüssel der IHK (Quelle: IHK-Ostwestfalen, 2012)	S. 145
Tabelle 5-3	Bewertungsschlüssel der Auszubildenden	S. 145
Tabelle 7-1	Geschlechterverteilung in der Bewerberstichprobe	S. 174
Tabelle 7-2	Schulbildungsniveau in der Bewerberstichprobe	S. 174
Tabelle 7-3	Geschlechterverteilung in der Retest-Stichprobe	S. 176
Tabelle 7-4	Bildungsniveau der Retest-Stichprobe	S. 177
Tabelle 7-5	Berufliche Vorerfahrung in der Retest-Stichprobe	S. 178
Tabelle 7-6	Geschlechterverteilung in der Schüler-Kontrollgruppe	S. 180
Tabelle 7-7	Berufsvorstellungen in der Schüler-Kontrollgruppe	S. 181
Tabelle 7-8	Überblick über die verfügbaren Kriteriumsdaten in der Schüler-Kontrollgruppe	S. 181
Tabelle 8-1	Fehlende Werte Bewerberstichprobe	S. 185
Tabelle 8-2	Fehlende Werte Kontrollgruppe	S. 185
Tabelle 8-3	Fehlende Werte Retest-Stichprobe	S. 185
Tabelle 8-4	Itemkennwerte der AGS	S. 187
Tabelle 8-5	Itemschwierigkeiten der AGS	S. 189
Tabelle 8-6	Korrigierte Trennschärfekoeffizienten der Items für die beiden AGS-Skalen Ordnung und Fleiß	S. 190
Tabelle 8-7	Retestreliabilität AGS und Selbstwirksamkeit	S. 191
Tabelle 8-8	Korrelationen der MTMM-Analyse	S. 198
Tabelle 8-9	Interkorrelationstabelle kaufmännische Auszubildende	S. 202
Tabelle 8-10	Interkorrelationstabelle technische Auszubildende	S. 203
Tabelle 8-11	Interkorrelationstabelle Kontrollgruppe	S. 204
Tabelle 8-12	Ergebnisse der Korrekturberechnungen der Korrelationen in der Bewerberstichprobe	S. 207
Tabelle 8-13	Ergebnisse der Powerberechnungen für alle Stichproben	S. 208
Tabelle 8-14	Ergebnisse der Moderationsanalyse für die Gewissenhaftigkeitsaspekte Fleiß und Ordnung zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden	S. 209
Tabelle 8-15	Ergebnisse der Moderationsanalyse für die Gewissenhaftigkeitsaspekte Fleiß und Ordnung zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden	S. 210
Tabelle 8-16	Ergebnisse der Moderationsanalyse für die Gewissenhaftigkeitsaspekte Fleiß und Ordnung zur Vorhersage von Schulleistung in der Kontrollgruppe	S. 210
Tabelle 8-17	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und das Selbstwirksamkeitskonstrukt zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden	S. 211

Tabelle 8-18	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und das Selbstwirksamkeitskonstrukt zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden	S. 211
Tabelle 8-19	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und das Selbstwirksamkeitskonstrukt zur Vorhersage von Schulleistung in der Kontrollgruppe	S. 212
Tabelle 8-20	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und die wahrgenommenen eigenen Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung in der Subgruppe der kaufmännischen Auszubildenden	S. 212
Tabelle 8-21	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und die wahrgenommenen eigenen Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung in der Subgruppe der technischen Auszubildenden	S. 213
Tabelle 8-22	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und die wahrgenommenen eigenen Fähigkeiten zur Vorhersage von Schulleistung in der Kontrollgruppe	S. 214
Tabelle 8-23	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und das Selbstwirksamkeitskonstrukt zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden	S. 214
Tabelle 8-24	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und das Selbstwirksamkeitskonstrukt zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden	S. 215
Tabelle 8-25	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und das Selbstwirksamkeitskonstrukt zur Vorhersage von Schulleistung in der Kontrollgruppe	S. 215
Tabelle 8-26	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und die wahrgenommenen eigenen Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung in der Subgruppe der kaufmännischen Auszubildenden	S. 216
Tabelle 8-27	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und die wahrgenommenen eigenen Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung in der Subgruppe der technischen Auszubildenden	S. 216
Tabelle 8-28	Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und wahrgenommene eigene Fähigkeiten zur Vorhersage von Schulleistung in der Kontrollgruppe	S. 217
Tabelle 8-29	Ergebnisse der multiplen Regression für Fleiß und Selbstwirksamkeit zur Vorhersage der Vorgesetztenbeurteilung bei den kaufmännischen Auszubildenden	S. 218
Tabelle 8-30	Ergebnisse der multiplen Regression für Fleiß und Selbstwirksamkeit zur Vorhersage der Vorgesetztenbeurteilung bei den technischen Auszubildenden	S. 218
Tabelle 8-31	Ergebnisse der multiplen Regression für Fleiß und Selbstwirksamkeit zur Vorhersage von Schulleistung in der Kontrollgruppe	S. 219

Tabelle 8-32	Ergebnisse der multiplen Regression für Ordnung und Selbstwirksamkeit zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden	S. 220
Tabelle 8-33	Ergebnisse der multiplen Regression für Ordnung und Selbstwirksamkeit zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden	S. 220
Tabelle 8-34	Ergebnisse der multiplen Regression für Ordnung und Selbstwirksamkeit zur Vorhersage der Abschlussnote in der Kontrollgruppe	S. 221
Tabelle 8-35	Ergebnisse der multiplen Regression für Fleiß und wahrgenommene eigene Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden	S. 221
Tabelle 8-36	Ergebnisse der multiplen Regression für Fleiß und wahrgenommene eigene Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden	S. 222
Tabelle 8-37	Ergebnisse der multiplen Regression für Ordnung und wahrgenommene eigene Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden	S. 223
Tabelle 8-38	Ergebnisse der Regressionsanalysen für Ordnung und wahrgenommene eigene Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden	S. 223
Tabelle 8-39	Ergebnisse der Mediationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und wahrgenommene eigene Fähigkeiten zur Vorhersage von Schulleistung in der Kontrollgruppe	S. 224
Tabelle 8-40	Korrelationen der AGS bzw. Gewissenhaftigkeitsskala (NEO-FFI) mit den Leistungskriterien in der Auszubildendenstichprobe	S. 225
Tabelle 8-41	Regressionsmodelle zur Vorhersage des Kriteriums der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden	S. 226
Tabelle 8-42	Regressionsmodelle zur Vorhersage des Kriteriums der allgemeinen Berufsschulleistung für die kaufmännischen Auszubildenden	S. 226
Tabelle 8-43	Regressionsmodelle zur Vorhersage des Kriteriums der Vorgesetztenbeurteilung für die technischen Auszubildenden	S. 227
Tabelle 8-44	Regressionsmodelle zur Vorhersage des Kriteriums der allgemeinen Berufsschulleistung für die technischen Auszubildenden	S. 227
Tabelle 8-45	Regressionsmodelle zur Vorhersage des Kriteriums der praktischen IHK-Zwischenprüfungsergebnisse für die technischen Auszubildenden	S. 228
Tabelle 8-46	Regressionsmodelle zur Vorhersage des Kriteriums der schriftlichen IHK-Zwischenprüfungsergebnisse für die technischen Auszubildenden	S. 228
Tabelle 8-47	Vergleich der Reliabilität und Standardabweichungen der Gewissenhaftigkeitsmaße	S. 229
Tabelle 8-48	Ergebnisse der t-Tests zur Überprüfung der Gruppenunterschiede von Auszubildenden mit und ohne berufliche Vorerfahrungen	S. 234
Tabelle A-1	Fehlende Werte in den AGS	S. 299
Tabelle A-2	Fehlende Werte im NEO-FFI und Selbstwirksamkeitsskala	S. 299
Tabelle A-3	Fehlende Werte Retest AGS (Auszubildende)	S. 300
Tabelle A-4	Fehlende Werte Selbstwirksamkeit T2 Azubis und Kontrollgruppe	S. 300

Tabelle A-5	Fehlende Werte wahrgenommene eigene Fähigkeiten	S. 300
Tabelle A-6	Ausgeschlossene Fälle aufgrund zu hohem Anteil fehlender Werte	S. 301
Tabelle A-7	Mittelwerte vor und nach der Mittelwertimputation in der Bewerberstichprobe	S. 301
Tabelle A-8	Mittelwerte vor und nach der Mittelwertimputation in der Kontrollgruppe	S. 302
Tabelle A-9	Mittelwerte vor und nach der Mittelwertimputation in der Reteststichprobe	S. 302
Tabelle A-10	Steigerungsmöglichkeiten von Cronbachs Alpha	S. 303
Tabelle A-11	Überprüfung der multivariaten Normalverteilung auf Itemebene	S. 304
Tabelle A-12	Iteminterkorrelationen der Skala Fleiß	S. 305
Tabelle A-13	Iteminterkorrelationen der Skala Ordnung	S. 306
Tabelle A-14	Standardisierte Ladungen und Fehlervarianzen der AGS	S. 309
Tabelle A-15	Quadrierte multiple Korrelationen	S. 310
Tabelle A-16	Interkorrelationstabelle Retest-Stichprobe gesamt	S. 311
Tabelle A-17	Ergebnisse der Mediationsanalyse für Fleiß und Ordnung zur Vorhersage der Vorgesetztenbeurteilung bei den kaufmännischen Auszubildenden	S. 312
Tabelle A-18	Ergebnisse der Regressionsanalyse für Fleiß und Ordnung zur Vorhersage der Vorgesetztenbeurteilung bei den technischen Auszubildenden	S. 312

Abkürzungsverzeichnis

α	Cronbachs Alpha
Abb.	Abbildung
AB5C	Abrigded Big-Five Circumplex
ADF	Asymptotically Distribution-Free
AG	Arbeitsgruppe
AGG	Antidiskriminierungsgesetz
AGS	Arbeitsbezogene Gewissenhaftigkeitsskalen
al.	alii (et alii, lateinisch für: und andere)
AO-Psychologie	Arbeits- und Organisationspsychologie
Aufl.	Auflage
β	Pfadkoeffizient
B	Regressionskoeffizient
BAI	Beliefs About Intelligence
BDSG	Gesetz zur Regelung des Beschäftigtendatenschutzes
BEW	Bewerberstichprobe
BFAS	Big Five Aspect Scales
BFI (K-BFI)	Big Five Inventory (Kurzversion)
BIP	Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung
BPI	Business Personality Indicator
BSW	Berufliche Selbstwirksamkeitserwartungen
bzgl.	Bezüglich
bzw.	Beziehungsweise
ca.	Circa
CFA	konfirmatorische Faktorenanalyse
CMSDS	Crowne Marlowe Social Desirability Scale
<i>c.r.</i>	critical ratio
d	Effektstärke
df	degrees of freedom, Freiheitsgrade
d.h.	das heißt
DH	Dienst oder Hingabe für eine Idee oder Sache
DIHK	Deutsche Industrie- und Handelskammer
ε	Messfehler
E	Exzess (Kurtosis)
EDV	elektronische Datenverarbeitung
F	F-Wert
F	Fleiß
f./ ff.	und folgende Seite/ -n
FFM	Fünf-Faktoren-Modell
FOR	Frame-of-Reference
FPI	Freiburger Persönlichkeitsinventar
ggf.	gegebenenfalls
GH	Gewissenhaftigkeit
GM	Befähigung zum General Managment
GPA	Grade Point Average (Durchschnittsnote)
GSE	General Self-Efficacy
λ	Ladung
h^2	Kommunalität
H1, H2, ...	Hypothese 1, Hypothese 2, ...
Hrsg.	Herausgeber

HTHM	Heterotrait-Heteromethod
HTMM	Heterotrait-Monomethod
<i>i</i>	Personenindex
IBA	International Business Administration
i.d.R.	in der Regel
IHK	Industrie- und Handelskammer
IPV	Intrapersonelle Antwortvariabilität
IST 2000	Intelligenz-Struktur-Test 2000
<i>k</i>	Studienanzahl
KA	Karriere Anker
KG	Kontrollgruppe
KMU	kleine und mittelständische Unternehmen
korr.	korrigiert
Kov	Kovarianz
KTt	Klassische Testtheorie
LPS	Leistungsprüfsystem
LS	Lebensstilintegration
M	Mittelwert
MAR	Missing At Random
Max	Maximum
MBA	Master of Business Administration
MCAR	Missing Completely At Random
Med.	Mediator
Min	Minimum
ML	Maximum Likelihood
MNAR	Missing Not At Random
Mod.	Moderator
MRT	Magnetresonanztomographie
MTMM	Multi-Trait-Multi-Method
MVA	Missing Value Analysis
<i>N, n</i>	Stichprobengröße
NEO-FFI	NEO-Fünf-Faktoren-Inventar
NEO-PI-R	NEO-Personality Inventory revidierte Version
Nr.	Nummer
n.s.	nicht signifikant
O	Ordnung
o.ä.	oder ähnliche/s
OAR	Observed At Random
o.g.	oben genannte/n
OPQ	Occupational Personality Questionnaire
<i>p</i>	Signifikanzniveau
<i>P</i>	Itemschwierigkeit, Popularität
PA	Perceived Abilities (wahrgenommene eigene Fähigkeiten)
PC	Personal Computer
PPIK-Theorie	Process, Personality, Interests and Knowledge-Theorie
Präd.	Prädiktor
PRF	Personality Research Form
<i>r</i>	Korrelationskoeffizient, Validitätskoeffizient
R^2	Determinationskoeffizient (aufgeklärte Varianz)
RA	Rechenaufgaben
RCI	Reliable Change Index

<i>RMSEA</i>	Root-Mean-Square-Error of Approximation
s.	siehe
S.	Seite
SAT	Scholastic Aptitude Test
SB	Sicherheit/ Beständigkeit
SBI	Selbstbeschreibungsinventar
<i>SD</i>	Standardabweichung
SDS-17	Social-Desirability-Scale-17
SE	Selbstwirksamkeit
S.E.	Standardfehler
s.o.	siehe oben
sog.	sogenannte/r
SPSS	Statistical Software Package for the Social Sciences
<i>SRMR</i>	Standardized-Root-Mean-Residual
START-P	Testbatterie für Berufsanfänger – Persönlichkeitstest
s.u.	siehe unten
SU	Selbstständigkeit/ Unabhängigkeit
<i>t</i>	t-Wert
T1/ T2	Testzeitpunkt 1/ Testzeitpunkt 2
Tab.	Tabelle
TF	Technische/ funktionale Kompetenz
TH	Totale Herausforderung
u.a.	unten angegeben/e
u.ä.	und ähnliche/s
UK	Unternehmerische Kreativität
USA	United States of America
u.U.	unter Umständen
<i>v</i>	Schiefe
<i>v</i>	Variablenindex
V	Varianz
v.a.	vor allem
Val.	Validität
vgl.	vergleiche
vollst.	vollständige
vs.	versus
WB5	Workplace Big Five
z.B.	zum Beispiel

1. Einleitung

*„People see the extraordinary feats of others,
but not the unweaving commitment
and countless hours of perserverant effort
that produce them.”*

Bandura, 1997, S. 119

Ein zentrales Interesse der Arbeits- und Organisationspsychologie ist eine möglichst genaue Vorhersage von Arbeitsleistung mittels wissenschaftlich entwickelter und validierter Testverfahren sowie die Entwicklung von Theorien hierzu. Auch in der pädagogischen Psychologie ist die Vorhersage von Schulerfolg sowie das Verhindern akademischen Misserfolgs ein zentrales Anliegen der Forschung. Seit mehr als hundert Jahren gibt es Untersuchungen und verschiedenste theoretische Ansätze zur Erklärung der Leistung und Erfolg bedingenden Faktoren.

Eine Kombination der beiden Forschungsrichtungen bildet die Erforschung des dualen Ausbildungssystems, das in dieser Form in Deutschland, Österreich und der Schweiz existiert. Dieser Bereich ist im Hinblick auf die Tatsache, dass rund die Hälfte der deutschen Bevölkerung über 15 Jahren einen beruflichen Abschluss besitzt, den sie durch eine Lehre bzw. eine Berufsausbildung im dualen System erworben hat (Statistisches Bundesamt, 2011), gesamtwirtschaftlich gesehen sehr bedeutend. Mehr als 50 % der Auszubildenden verbleiben nach dem erfolgreichen Abschluss ihrer Berufsausbildung in dem von ihnen erlernten Beruf. Allerdings gibt es bezüglich der Leistungs- und Erfolgsvorhersage von Auszubildenden in Deutschland vergleichsweise wenige Forschungsaktivitäten. Bislang existiert kein Vorhersagemodell, das die Zusammenhänge von Ausbildungserfolg mit verschiedensten Prädiktoren in Verbindung bringt und weitere Einflussvariablen mit berücksichtigt (Dudley, Orvis, Lebiecki & Cortina, 2006).

Schmidt und Hunter (1998) empfahlen in ihrer Metaanalyse für die Auswahl von Mitarbeitern¹ ohne Berufserfahrung, Testverfahren zur Erfassung allgemeiner kognitiver Fähigkeiten zur Vorhersage von Berufserfolg zu verwenden. Kombiniert mit Integritätstests, strukturierten Interviews sowie Gewissenhaftigkeitsmaßen würden diese Leistung am besten vorhersagen. Leider wird dies in der Praxis nicht so gehandhabt, da viele Praktiker dazu tendieren, Verfahren anzuwenden, die weniger valide sind und die Forschungsergebnisse regelmäßig ignorieren (König, Klehe, Berchtold & Kleinmann, 2010; Schuler, Hell, Trapmann, Schaar & Borarmir, 2007). In der Praxis der Personalauswahl werden weiterhin vor allem Bewerbungsunterlagen von Auszubildenden gesichtet und unstrukturierte Interviews zur Bewerberauswahl bzw. Leistungsprognose geführt. Allerdings zeigte sich, dass für gleiche Leistungen in unterschiedlichen Schulen auch unterschiedliche Noten vergeben werden. Die Qualität von Schulnoten als Vorhersagekriterium wird von vielen Forschern aus diesem Grund bemängelt (Görlich & Schuler, 2007). Daher sollten diese für die Auswahl von Auszubildenden nicht das einzige oder Haupteinstellungskriterium sein.

In Deutschland ist der Einsatz von Intelligenztests in der Praxis der beruflichen Eignungsdiagnostik im internationalen Vergleich nicht nur gering, sondern sogar rückgängig. Nur 30 % der Firmen nutzen Verfahren zur Erfassung allgemeiner kognitiver Fähigkeiten zur Personalauswahl (Schuler et al., 2007).

Persönlichkeitstests werden sogar noch weniger verwendet, obwohl in letzter Zeit die Erforschung der Persönlichkeit wieder verstärkt im Fokus der Psychologie steht (Ortner et al.,

¹ Werden Personenbezeichnungen aus Gründen der besseren Lesbarkeit lediglich in der männlichen oder weiblichen Form verwendet, so schließt dies das jeweils andere Geschlecht mit ein.

2007). Neuere Forschungsergebnisse deuten zudem an, dass Gewissenhaftigkeit sowie deren Facetten nicht nur zur Vorhersage von Ausbildungserfolg geeignet sind, sondern über Intelligenzmaße hinweg auch inkrementelle Validität besitzen (Lievens, Ones & Dilchert, 2009; Spengler & Fintrup, 2011; Sutin, Costa, Miech & Eaton, 2009). Vor allem Gewissenhaftigkeit bzw. deren Facetten haben sich als valide Prädiktoren von Ausbildungs- bzw. schulischer und beruflicher Leistung erwiesen (Lievens et al., 2009; O'Connor & Paunonen, 2007). Insbesondere bei der Vorhersage von akademischer Leistung, v.a. für höhere Bildungsabschlüsse, scheinen Persönlichkeitsfragebögen ähnlich effektiv zu sein wie Intelligenztests.

Im Gegensatz zu Ländern wie den Niederlanden, Großbritannien oder Frankreich wird in Deutschland bei der Personalauswahl der systematischen Erfassung von Persönlichkeitsaspekten mittels standardisierter Testverfahren aber weiterhin vergleichsweise wenig Beachtung geschenkt. Dies widerspricht nicht nur der empirischen Befundlage zur Validität von allgemeiner kognitiver Fähigkeit und Persönlichkeitseigenschaften, sondern auch der in vielen Unternehmensauftritten getroffenen Aussage, dass neben Ausbildung und Berufserfahrung vor allem auf die „Persönlichkeit“ der Mitarbeiter besonderer Wert gelegt werde. Zudem werden Persönlichkeitseigenschaften auch gerade durch die Umwandlung der Markt- in eine Dienstleistungsgesellschaft immer wichtiger (Hülshager & Maier, 2008).

Als Gründe für die Nicht-Nutzung von Persönlichkeitstests für die Auswahl von Mitarbeitern geben die Personaler vor allem Ängste vor negativen Bewerberrückmeldungen an. Aber auch als zu hoch eingeschätzte Kosten bei der Anwendung dieser Verfahren sowie der (geringe) Verbreitungsgrad von Persönlichkeitstests in der Praxis sind Hindernisse bei deren Anwendung. Die prädiktive Validität der Verfahren spielt bei der Beurteilung der Nützlichkeit dieser Tests eine eher untergeordnete Rolle. Vielmehr ist die wahrgenommene Validität ein signifikanter Prädiktor für die Nutzung bzw. Nicht-Nutzung von Persönlichkeitstests (König et al., 2010). Rechtliche Überlegungen scheinen den Autoren zufolge bei der Debatte ebenfalls keine Bedeutung für die Personaler zu haben. Als weiteren Grund gegen die Verwendung von Persönlichkeitstests im Rahmen der Personalauswahl wurde ihren Ergebnissen zufolge auch vielfach sowohl seitens der Bewerber als auch auf Seiten der Personaler der scheinbar fehlende Bezug zur Arbeitstätigkeit genannt.

Da die Forschung im Bereich der Auswahlmethoden derzeit außerdem dahin tendiert, ein Plateau zu erreichen, was die Vorhersagekraft der Verfahren betrifft, gilt es nun Möglichkeiten auszumachen, wie man die Vorhersagekraft der bestehenden Verfahren erhöhen kann. Eine solche Möglichkeit bietet die Kontextualisierung von Persönlichkeitsfragebögen für die Anwendung in der beruflichen Eignungsdiagnostik. Die Annahme dahinter ist, dass es zu einer Interaktion dispositioneller Eigenschaften mit situationalen Faktoren kommt (Sutherland, de Bruin & Crous, 2007). Dispositionelle Eigenschaften beziehen sich auf Faktoren oder Variablen wie Persönlichkeitseigenschaften, Bedürfnisse, Einstellungen, Präferenzen, geistige Fähigkeiten, emotionale Intelligenz und Motive. Situationelle Faktoren beinhalten beispielsweise organisationale Vorgaben und Prozeduren, Management-Praktiken, Autonomie oder Teamarbeit, sind aber nicht begrenzt auf diese.

Auch um die Kritik des fehlenden Anwendungsbezugs von Persönlichkeitstests aufzugreifen, wurden daher die Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS) entwickelt, deren Evaluation hinsichtlich ihrer Konstruktvalidität sowie prädiktiven Validität in dieser Arbeit erfolgen soll.

Neben den Gewissenhaftigkeitsaspekten soll zudem das Selbstwirksamkeitskonstrukt von Bandura (1977) als weiterer Prädiktor bzw. beeinflussende Variable für die Leistungsvorhersage berücksichtigt werden. Selbstwirksamkeit hat sich sowohl im beruflichen (Judge & Bono, 2001; Stajkovic & Luthans, 1998) als auch im schulischen Bereich (Gore, 2006; Okech & Harrington, 2002) als valider Prädiktor von Leistung gezeigt und weist mit der Gewissen-

haftigkeitsdimension u.a. in Bezug auf Lernen Zusammenhänge auf (Barrick & Mount, 1991, Chen, Casper & Cortina, 2001). Vor allem für Berufsanfänger ist berufliche Selbstwirksamkeit somit ein wichtiger Prädiktor von Ausbildungserfolg (Abele, Stief & Andrä, 2000). Es wird daher angenommen, dass das Konstrukt bei der Vorhersage von Schulabschluss- und IHK-Noten einen wichtigen Beitrag leisten kann.

Darüber hinaus wird das noch eher wenig beforschte und bekannte Konstrukt der wahrgenommenen eigenen Fähigkeiten (Noftle & Robins, 2007) als beeinflussende Variable in das Vorhersagemodell mit aufgenommen und evaluiert.

Diese Arbeit zur Vorhersage von Ausbildungserfolg verfolgt somit das Ziel, eine arbeitsbezogene Gewissenhaftigkeitsskala gemäß der geltenden Gütekriterien der Klassischen Testtheorie umfassend zu evaluieren, diese mit einem allgemeinen, häufig in der Eignungsdiagnostik verwendeten Gewissenhaftigkeitsmaß zu vergleichen und ein komplexeres Modell zur Vorhersage von IHK-Noten zu überprüfen, das u.a. das Selbstwirksamkeitskonstrukt (Bandura, 1977) als weiteren Prädiktor bzw. beeinflussende Variable mit berücksichtigt.

1.1 Nutzen der Arbeit

Die Attraktivität dieses Forschungsvorhabens ist vielfältig. Zum einen kann mit der Weiterentwicklung und Validierung der Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS) das Repertoire an eignungsdiagnostischen Verfahren im Kontext der Personalauswahl und -entwicklung um ein sensibles, reliables und ökonomisches Instrument erweitert werden. Aufgrund der angestrebten hohen psychometrischen Güte sowie der ausschließlichen Erfassung relevanter Persönlichkeitseigenschaften liegt mit den AGS ein Instrument vor, das vor allem für die Praxis ein interessantes Testverfahren darstellen sollte. Die Überprüfung der Vorhersagekraft des Instruments in Bezug auf die Vorhersage von beruflichem bzw. schulischem Erfolg in einer Feldstichprobe ermöglicht die Ableitung von Implikationen für die Anwendung der Skala in der Praxis der Personalauswahl im Bereich der Vorauswahl von Auszubildenden. Die praxisnahe Studie ermöglicht somit konkrete Aussagen über den Nutzen der AGS in realen Auswahl-situationen. Dies sollte auch der Fall sein, wenn eine Testperson noch keine Berufserfahrung verfügt, da die Kontextualisierung auch den Zweck einer einheitlicheren Itembeantwortung und somit eine Reduktion der interpersonellen sowie intrapersonellen Variabilität erwirken soll (Reddock, Biderman & Nguyen, 2011).

Zudem liegt mit den AGS ein Verfahren vor, das kulturübergreifend bzw. mehrsprachig entwickelt wird. Kersting (2011) forderte für die Gruppe der psychometrischen Verfahren die Entwicklung mehrsprachiger Versionen, deren Äquivalenz eindeutig nachgewiesen ist. Vor allem in Anbetracht des zunehmenden kulturellen Wandels und der zunehmenden Internationalisierung vieler Unternehmen ist dies ein Aspekt, der im Rahmen der Testkonstruktion bisher oftmals vernachlässigt wurde.

Die Analyse bzw. Antizipation bisheriger Problemfelder der Persönlichkeitstestung verstärkt den praktischen Nutzen dieser Studie, z.B. durch die Kontextualisierung des Verfahrens und die Analyse des Einflusses sozial erwünschten Antwortverhaltens. So sollte der klar erkennbare Arbeitsbezug des neu entwickelten Fragebogens nicht nur die prädiktive, sondern auch die Augenscheinvalidität erhöhen und somit letztlich zu einer erhöhten Akzeptanz des Persönlichkeitsmaßes beitragen.

Das längsschnittliche Design der Studie spricht darüber hinaus für die Qualität und Aussagekraft der abgeleiteten Aussagen.

Hervorzuheben ist auch der erwartete Nutzen aus der Überprüfung des komplexen Vorhersagemodells. Persönlichkeitsmaße könnten in der eignungsdiagnostischen Praxis um weitere Selbstbeschreibungsinventare ergänzt werden und so in Kombination zu präziseren Vorhersa-

gen von Berufserfolg führen. Dies ist sowohl im Sinne der Bewerber als auch der Unternehmen relevant, da beide Seiten an einer möglichst hohen Passung zwischen Bewerber und vakanter Position interessiert sein sollten.

Darüber hinaus ist besonders in Deutschland aufgrund des derzeit noch dreigliedrigen Schulsystems und der damit einhergehenden starken Vorselektion der Schüler eine Berücksichtigung weiterer, nicht-kognitiver Einflussfaktoren bzw. Prädiktoren notwendig. Intelligenzmaße hängen in Deutschland im internationalen Vergleich aufgrund der bestehenden schulischen Vorselektion weniger mit Leistungsmaßen zusammen (Salgado, Anderson, Moscoso, Bertua & De Fruyt, 2003). Zusätzliche Auswahlfaktoren wie Persönlichkeitsvariablen könnten diesbezüglich helfen, die Zuverlässigkeit von Auswahlentscheidungen zu erhöhen. Zudem werden seitens der Wissenschaft weitere Validierungsstudien von Persönlichkeitstests im Kontext der praktischen Personalauswahl in Deutschland gefordert (Hülshager & Maier, 2008).

Mit der Validierung der AGS erfüllt die vorliegende Studie die Forderung nach einem Brückenschlag zwischen Wissenschaft und Praxis. Einerseits soll ein Beitrag zur weiteren Bekanntmachung von Gewissenhaftigkeits- bzw. Persönlichkeitsmessungen in der Praxis der Personalauswahl und -entwicklung geleistet werden, da diese Verfahren in der eignungsdiagnostischen Praxis in Deutschland bisher immer noch zu Unrecht unterrepräsentiert sind (vgl. Abschnitt 2.3). Andererseits soll die praktische Anwendung der Skala dazu beitragen, anwendungsnahe Normen zu erstellen, mit deren Hilfe Testdaten aus Personalauswahlverfahren besser eingeordnet werden können. Die Bereitstellung von Normen, die anhand repräsentativer Stichproben arbeitstätiger Personen verschiedener Berufsgruppen ermittelt wurden, ist nämlich eine notwendige Voraussetzung für die Verwendung eines Verfahrens in der eignungsdiagnostischen Praxis (Hülshager & Maier, 2008). Hierzu zählt auch die Erhebung von Normen in Bewerberstichproben sowie relevanten Vergleichsgruppen. Entsprechend kritisierte auch Kersting (2011), dass Tests bislang noch zu häufig an Gelegenheits- und Zufallsstichproben, wie beispielsweise Schülern und Studenten, validiert werden. Benötigt werden Daten aus dem Feld, insbesondere auch von älteren Mitarbeitern und Bewerbern, vor allem auch mit Migrationshintergrund. Auch im Sinne der Einhaltung bzw. Erfüllung der Anforderungen der DIN 33430, welche die Anforderungen an eignungsdiagnostische Verfahren und deren Einsatz in der Personalpsychologie beschreibt, sind anwendungsbezogene Normtabellen von großem Wert (DIN, 2002).

Aus der Verschmelzung von Praxis und Wissenschaft können wiederum neue Fragestellungen entstehen. Somit dient diese Arbeit auch dem allgemeinen wissenschaftlichen Fortschritt in der eignungsdiagnostischen Forschung und Praxis.

1.2 Struktur der Arbeit

Im Folgenden wird der Aufbau der vorliegenden Arbeit skizziert.

Zunächst werden in Kapitel zwei der theoretische und empirische Hintergrund der Untersuchung vorgestellt. Ausgehend von einer Skizzierung der Forschungsgeschichte der beruflichen Eignungsdiagnostik, wird vor allem auf die Bedeutung von Persönlichkeitseigenschaften in dieser eingegangen. Der Fokus liegt hierbei auf den Merkmalen Gewissenhaftigkeit und Selbstwirksamkeit sowie der Definition von Leistung im Beruf, der Schule sowie der Berufsschule. Am Ende wird der aktuelle Forschungsstand zu Vorhersagemodellen zur Ausbildungsleistung skizziert, bevor verschiedene Modelle zur Vorhersage von Ausbildungsleistung mittels der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie Selbstwirksamkeit und wahrgenommener eigener Fähigkeiten im Rahmen der Arbeit hinsichtlich ihrer Varianzaufklärung getestet werden.

Im dritten Kapitel werden die Ziele der vorliegenden Arbeit präzisiert. Die wesentlichen inhaltlichen Hypothesen beziehen sich einerseits auf die Überprüfung der Gütekriterien Reliabilität und Validität der weiter entwickelten Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS), den Vergleich der AGS mit einem allgemeinen Gewissenhaftigkeitsmaß sowie die Überprüfung des theoretischen Vorhersagemodells. Auch auf mögliche konfundierende Variablen wird in diesem Kapitel eingegangen.

Die Weiterentwicklung der AGS ist Bestandteil des vierten Kapitels. Es werden Informationen zur Testkonstruktion und Weiterentwicklung der Skala gegeben sowie bisher veröffentlichte Studien zu den AGS beschrieben. Auch auf Zusammenhänge mit anderen Konstrukten und Kriterien wird eingegangen.

Kapitel fünf ist der Schilderung der Untersuchungsplanung bzw. Operationalisierung gewidmet. Die AGS sowie die weiteren Einflussvariablen werden im Rahmen einer Feldstudie in einer Bewerberstichprobe sowie in einer Kontrollgruppe von Schülern erhoben. Als abhängige Variablen dienen die Ausbildungsleistung der Auszubildenden, d.h. die Noten der Berufsschule bzw. der Industrie- und Handelskammer (IHK) sowie eine Vorgesetztenbeurteilung. In der Kontrollgruppe werden als Kriteriumsvariable die Abschlussnoten der Schüler verwendet.

Um die angenommenen Hypothesen empirisch zu überprüfen, kommen verschiedene statistische Verfahren zum Einsatz, deren Voraussetzungen und Gütekriterien in Kapitel sechs erläutert werden. Im ersten Teil werden die Methoden der Datenaufbereitung geschildert. Anschließend werden Reliabilitäts- und Multi-Trait-Multi-Method-Analysen sowie konfirmatorische Faktorenanalysen vorgestellt, mit denen die AGS hinsichtlich ihrer Skaleneigenschaften überprüft werden sollen. Der Modell-Fit wird mit Hilfe von Moderations- bzw. Mediationsanalysen untersucht. Die Basis für diese Ansätze sowie die Überprüfung der inkrementellen Validität bilden Regressionsanalysen. Darüber hinaus werden am Ende des Kapitels einige weitere Verfahren beschrieben, die u.a. zur Überprüfung der Effekte der Kontrollvariablen sowie Voraussetzungsprüfung benötigt werden.

Die Durchführung der Studie wird in Kapitel sieben dokumentiert. Hierzu werden Charakteristika der Stichproben, der jeweiligen Versuchsumgebungen sowie die Abläufe der Testungen beschrieben.

In Kapitel acht folgt die Darstellung der Untersuchungsergebnisse. Dabei kommen die in Kapitel sechs beschriebenen statistischen Verfahren zum Einsatz, um die angenommenen Hypothesen zu überprüfen.

Anhand der Ergebnisse folgt im neunten Kapitel eine Diskussion, inwiefern die Ziele der vorliegenden Arbeit erreicht wurden. Es werden Vorteile und Einschränkungen dieser Studie sowie der darin gefundenen Ergebnisse diskutiert. Implikationen für die weitere Forschung sowie die Anwendung in der Praxis werden ebenfalls besprochen.

2. Theoretischer und empirischer Hintergrund

Das zentrale Anliegen der vorliegenden Arbeit besteht in der Überprüfung der prognostischen Validität der Persönlichkeitsmerkmale Fleiß und Ordnung sowie des Selbstwirksamkeitskonstrukts zur Vorhersage von Schulleistung bzw. IHK-Noten und Ausbildungserfolg. Der Rahmen dieser Untersuchung wird durch das breite Forschungsfeld der beruflichen Eignungsdiagnostik gestellt, der im folgenden Abschnitt 2.1 umrissen wird und als Einleitung in die Thematik dient. In Abschnitt 2.2 wird kurz auf allgemeine kognitive Fähigkeiten in der Eignungsdiagnostik eingegangen, bevor in Abschnitt 2.3 ein besonderes Augenmerk auf Geschichte, Relevanz und Vorhersagekraft von Persönlichkeitsmerkmalen in der beruflichen Eignungsdiagnostik gelegt wird. Anschließend wird eine kritische Bewertung von Persönlichkeitsmaßen bzw. Selbstbeschreibungsmaßen in der beruflichen Eignungsdiagnostik vorgenommen (2.4), bevor in Abschnitt 2.5 auf berufsbezogenes Testen eingegangen wird. Anschließend werden in Abschnitt 2.6 verschiedene, bereits vorliegende Instrumente zur Erfassung von Gewissenhaftigkeit unter besonderer Berücksichtigung ihrer Nützlichkeit sowie ihrer Aussage- und Vorhersagekraft in der beruflichen Eignungsdiagnostik vorgestellt. Abschnitt 2.7 widmet sich den Leistungs- bzw. Erfolgskriterien, die im Rahmen der Eignungsdiagnostik vorhergesagt werden können bzw. sollen. Abschließend wird in Abschnitt 2.8 der aktuelle Stand bzgl. der bestehenden Vorhersagemodelle in der beruflichen Eignungsdiagnostik berichtet und ein konzeptuelles Vorhersagemodell vorgestellt, in dem die Persönlichkeitseigenschaften Fleiß, Ordnung und Selbstwirksamkeit zu einem weiteren Erfolgsfaktor sowie den Leistungs- und Erfolgskriterien in Beziehung gesetzt werden.

2.1 Berufliche Eignungsdiagnostik

Nach Schuler und Höft (2006) wird mit dem Begriff *Berufseignungsdiagnostik* die „*Methodologie der Entwicklung, Prüfung und Anwendung psychologischer Testverfahren zur eignungsbezogenen Erfolgsprognose und als Entscheidungshilfe im beruflichen Kontext*“ bezeichnet (S. 102). Die wissenschaftliche Basis der beruflichen Eignungsdiagnostik stellen den Autoren zufolge vor allem Theorien über Anforderungen, Fähigkeiten und Leistungen sowie die zugehörigen Methoden zu deren Messung und Modelle der Klassifikation dar. Allgemein sind also immer die Anforderungen einer Tätigkeit oder Position zu ermitteln, die an die (zukünftigen) Inhaber gestellt werden. Daraus können dann die erforderlichen und zu messenden Eignungsmerkmale sowie relevante Leistungs- und Erfolgskriterien abgeleitet werden.

Traditionell wurden und werden bei der Auswahl von Mitarbeitern vor allem Fachkenntnisse als notwendige Voraussetzung für die erfolgreiche Ausübung einer Tätigkeit erachtet. Diese können u.a. durch eine Analyse der Bewerbungsunterlagen bzw. der beigefügten Zeugnisse, Zertifikate und des Lebenslaufs nachgewiesen werden. Vor allem bei jungen Bewerbern, die gerade erst in das Berufsleben einsteigen, fallen diese Indikatoren zur Leistungs- und Erfolgsvorhersage jedoch aus. Hier liegen in der Regel lediglich Abschlusszeugnisse vor, die zur Leistungs- und Erfolgsprognose genutzt werden können. Von den meisten Firmen werden zudem Einstellungsinterviews genutzt, um die Eignung und Motivation einer Person für eine bestimmte Tätigkeit oder Position zu erheben (s.u. Tabelle 2-1). Darüber hinaus erfreuen sich vor allem in handwerklichen Berufen, wie z.B. Zahntechnikunternehmen, Überprüfungen der handwerklichen Fähigkeiten der Bewerber zunehmender Beliebtheit. Die praktischen und in der Regel sehr anforderungsnahen Fähigkeitsüberprüfungen werden als *Simulationen* bezeichnet. Zunehmend wird in Einstellungsverfahren auch auf psychologische Konstrukte zurückgegriffen. So werden beispielsweise Intelligenz und Persönlichkeitseigenschaften in die Praxis der beruflichen Eignungsdiagnostik mit übernommen.

Eine Befragung von mehr als 2.000 Unternehmen durch den Deutschen Industrie- und Handelskammertag (DIHK) zeigte, dass Unternehmen neben Fachwissen auch an der „Persön-

lichkeit“ ihrer Bewerber interessiert sind (Backhaus, 2004a). Es geht den Unternehmen zunehmend nicht nur darum, hoch qualifizierte Mitarbeiter zu rekrutieren, sondern auch, ob ein Bewerber ins Team passt, leistungs- und lernwillig ist sowie sich flexibel und offen sich verändernden Rahmenbedingungen anpassen kann. Das CRF Deutschland (2009) ermittelte in seiner Studie „Top Arbeitgeber Deutschland 2009“ die zehn entscheidenden Einstellungskriterien bei Absolventen und Young Professionals in deutschen Firmen. Hier rangierten „Persönlichkeitsmerkmale“ auf dem ersten Platz, sogar noch vor Fachkenntnissen. Der Studie zufolge sehen 88 % der befragten deutschen Unternehmen die „Persönlichkeit“ des Bewerbers als relevantestes Kriterium an. Schul- und Abschlussnoten befinden sich mit nur 10 % Relevanz lediglich auf einem hinteren neunten Platz.

Sarges (2010) zufolge entscheiden insbesondere in der heutigen Zeit vor allem Persönlichkeitseigenschaften wie z.B. Lern-, Kommunikations- und Innovationsfähigkeit über Erfolg und Misserfolg im Beruf. Je komplexer eine Tätigkeit bzw. deren Anforderungen, desto wichtiger werden dem Autor zufolge Persönlichkeitsmerkmale. Insbesondere durch eine anforderungsbezogene und methodisch saubere Auswahl von Personal können Unternehmen ihre Wettbewerbsfähigkeit in der zunehmend dynamischeren Wirtschaft erhalten und ausbauen. Denn gerade in einer „Wissens- und Dienstleistungsgesellschaft“ wie Deutschland sind die Mitarbeiter die kritische Unternehmensressource, die im internationalen Wettbewerb am wenigsten Gefahr läuft, imitiert zu werden.

Somit spielen die sog. „Softskills“ eine wichtige Rolle für Ausbildungs- und Berufserfolg. Sie sind über Berufsgruppen hinweg bedeutsam und werden auch oft als *Schlüsselqualifikationen* bezeichnet (Backhaus, 2004b). *Persönlichkeitseigenschaften* werden als zeitlich überdauernde Bereitschaften oder Verhaltensdispositionen definiert, die bestimmte Verhaltensweisen oder -aspekte einer Person beschreiben und vorhersagen (Schuler & Höft, 2006). Im Gegensatz zum Fachwissen, das schnell veraltet sein kann, bleiben die Persönlichkeitsmerkmale eines Menschen im Laufe seines Lebens gleich ausgeprägt, da es sich hierbei um *Traits*, d.h. zeitlich stabile Verhaltensdispositionen, handelt.

Zur allgemeinen Klassifikation eignungsdiagnostischer Instrumente unterscheiden Schuler und Höft (2006) drei Ansätze: 1. den Eigenschaftsansatz, 2. den Simulationsansatz und 3. den Biografieansatz. Der Eigenschafts- oder Konstruktansatz beschreibt die Erfassung von Merkmalen, die als zeitlich stabil und überdauernd angenommen werden, wie z.B. Intelligenz und Gewissenhaftigkeit. Diese Eigenschaften werden in der Regel mit Hilfe von psychologischen Testverfahren erfasst. Die Testgüte dieser Verfahren wird durch die Konstruktvalidität (s. Abschnitt 6.2.3) erfasst. Der Simulationsansatz beinhaltet die Verfahren, die der anforderungsnahen Erfassung von Arbeitsverhalten dienen, wozu beispielsweise Arbeitsproben zählen. Die Güte von Simulationen wird mittels der Kontentvalidität, d.h. bezüglich ihrer inhaltlichen Repräsentativität bestimmt. Der Biografieansatz beinhaltet alle Verfahren, die entweder in schriftlicher oder mündlicher Form, beispielsweise im Einstellungsinterview, eignungsdiagnostisch relevante Informationen erfassen. Die Validitätslogik dieser Verfahren ist die der prognostischen Validität (s. Abschnitt 6.3 ff). Abbildung 2-1 veranschaulicht diesen sog. *Trimodalen Ansatz*.

Da sich die vorliegende Arbeit ausschließlich mit der Überprüfung der Vorhersagekraft von Persönlichkeitseigenschaften beschäftigt, wird im Folgenden lediglich der Eigenschaftsansatz weiter vertieft. Eine anschauliche Darstellung aller drei Ansätze ist bei Schuler und Höft (2006) nachzulesen. Risavy und Hausdorf (2011) geben zudem eine gute Übersicht, welche Arten von Personalauswahlstrategien es gibt und wie diese durchgeführt werden.

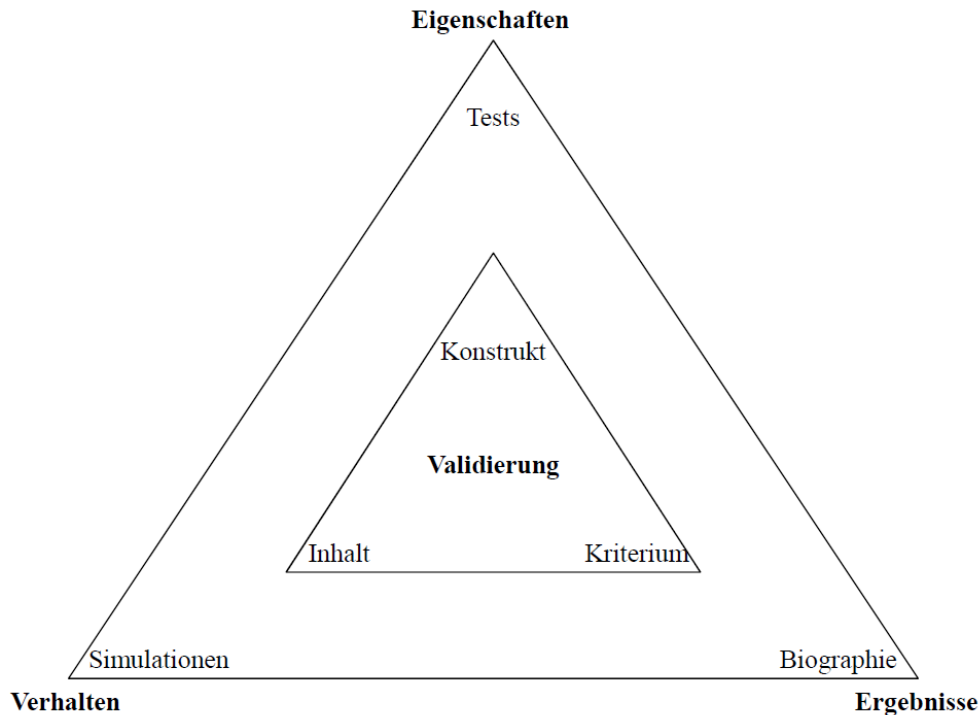


Abbildung 2-1: Die Ansätze der Eignungsdiagnostik und ihre Validierungslogik (Schuler & Höft, 2006, S. 103)

Der Konstruktansatz bedient sich vor allem psychologischer Testverfahren zur Ermittlung der interessierenden Eigenschaften. Tests sind wissenschaftlich begründete Hilfsmittel und für viele Laien das typische Werkzeug psychologisch-diagnostischer Arbeit. Es gibt verschiedene Arten von Testverfahren in der Berufseignungsdiagnostik. Hierzu zählen einerseits Leistungs- und Persönlichkeitstests, aber auch standardisierte Interviews oder Beobachtungen kann man gemäß Bühner (2011) allgemein in diese Methoden einsortieren. Tests erfassen dem Autor zufolge sowohl Fähigkeiten, Fertigkeiten und Eigenschaften als auch aktuelle Zustände einer Person. Wenn Tests sorgfältig konstruiert sind, korrekt genutzt und kompetent interpretiert werden, sind sie wertvolle Informationsquellen für die Arbeit von Psychologen und Personalern in der beruflichen Eignungsdiagnostik (Hagemeister, Lang & Kersting, 2010).

Spengler und Fintrup (2011) untersuchten in einer praxisnahen Studie, ob sich das Bestehen des anspruchsvollen Examens zum Wirtschaftsprüfer besser anhand von psychologischen Tests oder mittels Noten vorhersagen lässt. Schul- und Studiennoten sind wie oben erwähnt klassische erste Selektionsfilter für die Auswahl von Berufseinsteigern. Gemeinsam mit anderen Kriterien, wie z.B. dem Aussehen der Bewerbungsunterlagen, entscheiden sie in der Regel über die Zulassung oder Ablehnung eines Kandidaten zu weiteren Schritten in Personalauswahlprozessen, obwohl die prognostischen Mängel von Bewerbungsunterlagen wie beispielsweise mangelnde Vergleichbarkeit von Schulabschlüssen oder die Wohlwollenspflicht bei der Zeugniserstellung hinreichend bekannt sind. Die Autoren untersuchten 69 Hochschulabsolventen, die sich bei einem großen Wirtschaftsprüfungs- und Steuerberatungsunternehmen in der Schweiz berufsbegleitend zum Wirtschaftsprüfer ausbilden lassen wollten. Das von dem Unternehmen durchgeführte Auswahlverfahren war vierstufig, mit einer stufenweisen Selektionsstrategie. Aufgrund der hohen kognitiven Anforderungen im Wirtschaftsprüferexamen wurden nach der Sichtung der Bewerbungsunterlagen und neben Einstellungsinterviews auch Verfahren zur Erfassung der kognitiven Fähigkeiten der Bewerber eingesetzt. Darüber hinaus wurde auch die berufliche Leistungsmotivation erfasst. Die Ergebnisse zeigten, dass die Abschlussnoten der Bewerber insgesamt 27 % der Leistungsunterschiede der Kandidaten voraussagten. Ohne den zusätzlichen Einsatz der psychologischen Testverfahren hätten allerdings 37 % aller Kandidaten wegen ihrer geringen Abschlussnoten

eine falsche Leistungsprognose für ihr Wirtschaftsprüferexamen erhalten. Bei Hinzunahme der psychometrischen Daten verringerte sich dieser Wert bzw. stieg die Prognosekraft von 63 % auf 83 %. Der Anteil ungeeigneter Mitarbeiter, die fälschlicherweise als geeignet klassifiziert worden wären, verringerte sich den Analysen zufolge von 15 % auf 6 %. Auch der Anteil geeigneter Kandidaten, die fälschlicherweise als ungeeignet klassifiziert worden wären, sank von 22 % auf 11 %.

Insgesamt kann aus den Ergebnissen von Spengler und Fintrup (2011) somit geschlossen werden, dass psychologische Verfahren trotz der Varianzeinschränkung aufgrund der Selektionsentscheidung eine hohe Vorhersagekraft und inkrementelle Validität besitzen. Für die Unternehmen bedeutet dies auch, dass beispielsweise leistungsstärkere Kandidaten höhere Vertriebs- bzw. Verkaufszahlen erreichen und durch den Einsatz von Testverfahren somit deutlich zum Unternehmenserfolg beitragen. Die geringe Übereinstimmung der Uni-Note mit dem Bestehen des Wirtschaftsprüferexamens zeigt zudem auf, dass die Vorhersagekraft von Abschlussnoten bisher v.a. den hohen Basisraten in den allgemeinen Untersuchungen geschuldet war und die Vorhersagekraft von Noten vermutlich noch geringer ausfallen würde, wenn bei der Einstellungsentscheidung keine zusätzlichen, validen Testverfahren verwendet würden.

Allerdings besteht eine große Diskrepanz zwischen wissenschaftlich gesicherten Erkenntnissen zur Vorhersagequalität von eignungsdiagnostischen Methoden und der tatsächlichen Praxis der Personalauswahl in den Unternehmen. König und Kollegen (2010) ermittelten, welche Auswahlverfahren in deutschen Unternehmen verwendet bzw. wie häufig welche Verfahren eingesetzt werden. Den Autoren zufolge werden in nahezu allen Auswahlverfahren (99,6 %) die Unterlagen der Bewerber analysiert und strukturierte, semi-strukturierte oder unstrukturierte Auswahlgespräche (99,4 %) geführt. Ebenfalls überprüft die Mehrheit der Unternehmen die Referenzen der Bewerber (89,1 %). Ungefähr ein Drittel der befragten Firmen (32,0 %) gab an, Persönlichkeitseigenschaften zu erfassen. Ca. ein Viertel der Unternehmen führt zur Personalauswahl Assessment Center durch (26,3 %). Auch Arbeitsproben (23,5 %) und Fähigkeitstests (18,6 %) werden bei den befragten Unternehmen teilweise verwendet. Trotz der geringen prognostischen Validität des Verfahrens lassen vergleichsweise viele Unternehmen (15,8 %) graphologische Gutachten zu ihren Bewerbern erstellen. Biografische Fragebögen gehören bei 12,7 % der befragten Unternehmen zum Auswahlportfolio. Teilweise lassen Unternehmen auch medizinische Gutachten über ihre Bewerber erstellen (5,3 %), sofern dies tätigkeitsrelevant ist. Tabelle 2-1 gibt einen Überblick über die Verwendungshäufigkeit sowie die durch Schmidt und Hunter (1998) metaanalytisch ermittelte prognostische Validität der o.g. Auswahlverfahren. Die Ergebnisse der Studien zeigen deutlich die Kluft zwischen wissenschaftlich gesicherten Erkenntnissen und der Praxis der Personalauswahl in den Unternehmen.

In der wissenschaftlichen Eignungsdiagnostik werden demgegenüber berufs- und hierarchieübergreifend am häufigsten Testverfahren beforscht bzw. verwendet. Vor allem Intelligenz- und Persönlichkeitstests kommen hier zum Einsatz (Schuler & Höft, 2006). Schuler (2000, zitiert nach Schuler & Höft, 2006) verglich den Einsatz von Persönlichkeits- und Intelligenztests in großen Unternehmen in Europa. Seinen Ergebnissen zufolge verwendeten 35 % der Unternehmen Leistungstests, während nur 10 % der befragten Unternehmen auch Persönlichkeitstests einsetzten. In Frankreich und Spanien werden demgegenüber sowohl Leistungstests von 40 % bzw. 62 % als auch Persönlichkeitstests von 55 % bzw. 58 % der Unternehmen eingesetzt und somit deutlich mehr genutzt. Allerdings stammen die Daten dem Autor zufolge bereits aus dem Jahr 1990. Die Verwendung von Leistungs- und Persönlichkeitstests hat sich Schuler und Höft (2006) zufolge bei deutschen Unternehmen in den letzten Jahren deutlich gesteigert.

Tabelle 2-1: Anwendungshäufigkeit und prädiktive Validität von Personalauswahlverfahren

Auswahlprozedur	Verwendungshäufigkeit in der Praxis (König et al., 2010)	Prognostische Validität (r) (Schmidt & Hunter; 1998)
Analyse der Bewerbungsunterlagen	99,6 %	keine Angaben
Interviews	99,4 % (davon mind. 71,3 % semi-strukturiert)	.51 (strukturiert) .38 (unstrukturiert)
Überprüfung der Referenzen	89,1 %	.26
Persönlichkeitstests	32,0 %	.31 (nur Gewissenhaftigkeits- dimension)
Assessment Center	26,3 %	.37
Arbeitsproben	23,5 %	.54
Fähigkeitstests	18,6 %	.48
Graphologische Gutachten	15,8 %	.02
Biografische Fragebögen	12,7 %	.35
Medizinische Untersuchungen	5,3 %	keine Angaben

Ein weiteres Problem bei der Auswahl von Bewerbern, insbesondere im Rahmen der Auszubildendenauswahl, betrifft die schulische Grund- bzw. Vorbildung der Bewerber. Vor allem von Ausbildungsbeauftragten wird zunehmend eine Abnahme der Allgemeinbildung der Auszubildenden bemängelt. Eine tiefergehende Analyse dieser Problematik steht nicht im Fokus dieser Arbeit und wird an dieser Stelle daher nicht vorgenommen. Festzuhalten ist jedoch, dass die mangelhafte Bildungsplanung und -politik auch eine beeinflussende Rolle bei der steigenden Nachfrage nach qualifizierten Arbeitskräften in dem Sinne spielt, dass diese Nachfrage nach Fachkräften immer weniger bedient werden kann. Dazu kommt, dass der Anteil Geringqualifizierter, für die kein adäquater Bedarf besteht, sehr hoch ist (Kersting, 2011). In Zeiten des zunehmenden Personalmangels und der subjektiv abnehmenden Aussagekraft der Abschlusszeugnisse sind für die Unternehmen Auswahlverfahren, die sich nur an den laut Schulnoten „besten“ Absolventen orientieren, somit nicht mehr zielführend. Zu viele potenziell geeignete Talente werden so aufgrund mäßiger schulischer Leistungen zu früh aussortiert. Unternehmen wollen in der Regel nur die wenigen gut benoteten Bewerber haben. Unter Fachleuten wird diesbezüglich auch vom „War of Talents“ gesprochen. Insbesondere die Vorselektion von Bewerbern sollte Kersting (2011) zufolge daher in Bezug auf die Schulleistungen weniger streng ausfallen und um weitere aussagekräftige Screeningverfahren ergänzt werden. Auf diese Weise wird die Grundquote erhöht und die Validität der verwendeten Auswahlverfahren gewinnt an Bedeutung. In Zeiten des wachsenden Personalmangels kommt es nämlich nicht nur darauf an, die „besten“ Bewerber zu rekrutieren, sondern die langfristig „richtigen“ zu finden. Bewerber sollten allgemein intelligent und lernfähig sein (Horn, 1983). Zudem sollten sie passende Ausprägungen in relevanten Persönlichkeitsdimensionen wie z.B. Extraversion, emotionaler Stabilität und Gewissenhaftigkeit aufweisen (vgl. Abschnitt 2.3).

Allgemein sollte es bei der Auswahl von Auszubildenden und Mitarbeitern grundsätzlich nicht um eine reine „Bestenauslese“ gehen, sondern der am meisten passende Kandidat ausgewählt werden. Schon Horn (1983) postulierte, dass eine zu hohe Begabung für einen gewählten Beruf ähnlich nachteilig sein dürfte wie eine zu niedrige Leistung, auch wenn dies in der Regel erst viel später bemerkt würde. Konkret warnt der Autor: „*Wer überqualifiziert ist, wird entweder bei nächster Gelegenheit abwandern oder sich bald langweilen, sehr unzufrieden*“.

den werden und vielleicht dem Alkohol verfallen!“ und „Bei einer für die ergriffene Schul- oder Berufslaufbahn zu knappen Begabung kann fleißiges ‚Pauken‘ in vielen Fällen über Prüfungen oder Hürden ähnlicher Art hinweghelfen. Derartig eingepauktes, jedoch nur mangelhaft verstandenes Wissen ist nur schlecht anzuwenden und bei dem schnellen Wechsel und dem Anwachsen des Wissens von fragwürdigem Wert.“ (S. 55).

In den letzten Jahren wurde in der Arbeits- und Organisationspsychologie v.a. intensiv zur Vorhersagekraft von Persönlichkeitseigenschaften und allgemeinen kognitiven Fähigkeiten geforscht. Die Ergebnisse der wissenschaftlichen Bemühungen haben auch zur Theoriebildung im Bereich der Eignungsdiagnostik beigetragen. Andere Personalauswahlmethoden wie beispielsweise strukturierte Interviewtechniken, Biografische Fragebögen oder Assessment Center sind seitens der Forschung weniger stark beachtet worden, da diesbezüglich auch weniger klar ist, welche Theorien diesen Methoden zugrunde liegen (Schmidt, Shafer & Oh, 2008) und die Erforschung dieser Methoden zudem vergleichsweise aufwendig zu realisieren ist. Im Folgenden werden daher nur die Bedeutung allgemeiner kognitiver Fähigkeiten (s. Abschnitt 2.2) sowie von Persönlichkeitseigenschaften (s. Abschnitt 2.3 ff) im Kontext der Berufseignungsdiagnostik dargestellt.

2.2 Allgemeine kognitive Fähigkeiten in der Eignungsdiagnostik

Allgemeine kognitive Fähigkeiten werden auch als *Intelligenz* bezeichnet und umschreiben „die Gesamtheit aller kognitiven oder geistigen Fähigkeiten“ (Schuler & Höft, 2006, S. 105). Es gibt eine Vielzahl verschiedener Intelligenztheorien, die an dieser Stelle jedoch aus Gründen der Ökonomie nicht erläutert werden sollen. Für die vorliegende Fragestellung reicht es aus festzuhalten, dass alle Intelligenzkonzepte im Kern die Qualität und Geschwindigkeit der Lösung neuartiger, nicht routineartiger Aufgaben erfassen.

Intelligenztests als Maße der kognitiven Leistungsfähigkeit haben eine lange Tradition in der beruflichen Eignungsdiagnostik und wurden bereits frühzeitig für eignungsdiagnostische Zwecke eingesetzt, wie beispielsweise der „Army Alpha-Test“ zur Auswahl von Rekruten in Amerika Anfang des vergangenen Jahrhunderts (Schuler & Höft, 2001). Der Einsatz von Leistungstests in der beruflichen Eignungsdiagnostik unterlag jedoch – wie auch die Verwendung von Persönlichkeitstests – in der Mitte des vergangenen Jahrhunderts einem starken Rückgang. Aufgrund einiger narrativer Zusammenfassungen und der Beurteilung der Kriteriumsvalidität dieser Verfahren als zu niedrig, wurden Leistungstests in der Mitte des vergangenen Jahrhunderts immer weniger verwendet. Zudem gab es eine allgemeine Tendenz dahingehend, dass eignungsdiagnostische Verfahren allgemein aus ideologischen Gründen abgelehnt wurden. Darüber hinaus wurde bekannt, dass der Vererbungsforscher Burt seine Daten gefälscht hatte, was den gesamten Forschungszweig und dessen Ergebnisse in Verruf brachte (Schuler & Höft, 2006). Erst durch das Aufkommen der metaanalytischen Techniken gegen Ende des vergangenen Jahrhunderts – und damit einhergehend der Relativierung der unzureichenden bzw. falschen Ergebnisse – begann die Renaissance der Leistungs- und Persönlichkeitstests. Hunter und Hunter (1984) ermittelten in ihrer Metaanalyse auf diese Weise beispielsweise einen mittleren Validitätskoeffizienten von $r = .53$ für allgemeine Intelligenz, der in ihrer Untersuchung durch kein anderes eignungsdiagnostisches Einzelverfahren übertroffen wurde. Vor allem auch durch die von Hunter und Schmidt (2004) entwickelte Validitätsgeneralisierungstechnik sind mittlerweile zuverlässige Aussagen über die prognostische Validität von Leistungs- und Persönlichkeitsindikatoren möglich. So fanden Salgado et al. (2003) in einer entsprechenden Metaanalyse mit Studien aus dem europäischen Raum mittlere Validitätskoeffizienten von $r = .24$ (für Polizisten) bis $r = .67$ (für Manager).

Zwischen den verschiedenen Berufsgruppen unterscheiden sich die Intelligenzwerte jedoch teilweise deutlich. Je höher die Anforderungen einer Tätigkeit, desto geringer werden die

Streuungen und desto höher die Intelligenzwerte (Schuler & Höft, 2006). Es ist daher notwendig, auf eine gute Passung zwischen kognitiven Anforderungen einer Stelle und den Fähigkeiten einer Person zu achten. So werden Personen, die die kognitiven Anforderungen einer Stelle nicht erfüllen, dauerhaft erfolglos sein und somit wieder aus dem Beruf ausscheiden. Auch Personen, die die Anforderungen einer Stelle deutlich übererfüllen und somit auf Dauer unterfordert sind, werden ein Unternehmen wieder verlassen bzw. eine Tätigkeit wieder aufgeben (Horn, 1983).

Ein Problem bei der Anwendung von Intelligenztests in der beruflichen Eignungsdiagnostik ist, dass diese in großem Maße abhängig von den Rahmenbedingungen sind, unter denen die Verfahren durchgeführt werden. Zwischen den Angehörigen verschiedener Berufsgruppen gibt es zwar erhebliche Unterschiede in der kognitiven Leistungsfähigkeit, innerhalb der Berufsgruppen ist diese Streuung jedoch erheblich geringer (Schuler & Höft, 2001). Ist eine Bewerbergruppe beispielsweise bereits stark hinsichtlich ihrer kognitiven Fähigkeiten vorselektiert, ergeben sich geringere Vorhersagewerte. Dies konnten Funke, Krauss, Schuler und Stapf (1987) in einer Metaanalyse nachweisen, in der sich für die Leistungsvorhersage einer stark vorselegierten Stichprobe von wissenschaftlich-technischen Mitarbeitern trotz Korrekturformeln lediglich eine prognostische Validität von $r = .16$ ergab. Nicht-kognitive Verfahren wiesen in dieser Stichprobe höhere Vorhersagewerte auf.

Des Weiteren messen die in der beruflichen Eignungsdiagnostik verwendeten Intelligenzmaße alle die maximale Leistungsfähigkeit einer Person in Bezug auf einen bestimmten Aufgabentypus. Arbeitsleistung oder Erfolg im Beruf ist jedoch das Ergebnis von eher „typischen“ Verhaltensweisen. Leistungstests erfassen somit nicht das alltägliche Arbeitsverhalten, sondern lediglich Spitzenleistungen. Ackerman und Heggstad (1997) sehen hierin die Ursache für die teilweise niedrigen Korrelationen von Intelligenzmaßen mit schulischen, akademischen oder beruflichen Erfolgs- und Leistungsmaßen.

Ein weiteres Problem bei der Verwendung von Leistungstests ist, dass Angehörige von Minoritäten wie z.B. Kinder von Migranten in diesen Tests tendenziell schlechter abschneiden und somit aufgrund ihrer Herkunft benachteiligt sein könnten (Schuler & Höft, 2006). Dies kann bei der Verwendung von Leistungstests in der Praxis der Personalauswahl in Bezug auf das Antidiskriminierungsgesetz (AGG) problematisch sein.

Zu beachten ist auch, dass neben der prognostischen Validität eines Verfahrens auch die Erklärung der Zusammenhänge im Sinne der Konstruktvalidität nicht vergessen werden darf. „Die Prognosegüte eines Prädiktors gibt ohne weitergehende Untersuchungen wenig Aufschluss über Ursachenzusammenhänge“ (Schuler & Höft, 2006, S. 112). So nehmen beispielsweise Fachkenntnisse eine Mediatorfunktion zwischen Intelligenz und Vorgesetztenbeurteilungen ein. Intelligenz ist somit eher als Basis für den Erwerb von Fachwissen und Kenntnissen anzusehen.

Falls es um die Vorhersage spezifischer Informationen geht, empfehlen Schuler und Höft (2006) darüber hinaus, über allgemeine kognitive Leistungstests weitere Verfahren hinzuzuziehen, um eine detailliertere Analysen der individuellen Stärken und Schwächen eines Bewerbers vorzunehmen zu können. So kann beispielsweise die Ergänzung von Persönlichkeitseigenschaften einen deutlichen Mehrwert zur Vorhersage von Ausbildungs- und Berufserfolg leisten. Die inkrementelle Validität von Gewissenhaftigkeit zu Intelligenz betrug laut Schmidt und Hunter (1998) 18 %. Kein anderes Persönlichkeitsmerkmal reichte an diese Werte heran. Ackerman und Heggstad (1997) fanden in ihrer Metaanalyse, dass Persönlichkeit und Intelligenz sich wechselseitig beeinflussen. Moutafi, Furnham und Paltiel (2004) untersuchten an einer Stichprobe von $N = 211$ Bewerbern, die an einem Assessment Center teilnahmen, den Zusammenhang von Intelligenz und Gewissenhaftigkeit. Die Autoren fanden einen negativen Zusammenhang zwischen dem Leistungskonstrukt und der Persönlichkeitsdimension. Moutafi

und Kollegen (2004) vermuten, dass weniger intelligente Personen durch eine höhere Gewissenhaftigkeit ihren geistigen Nachteil auszugleichen versuchen. Beispielsweise könnten weniger intelligente Schüler durch erhöhtes Lernen und das regelmäßige Erledigen von Hausaufgaben trotzdem gute Noten erreichen, während intelligentere Schüler diesen Mehraufwand nicht betreiben müssen, um gute Ergebnisse zu erzielen (vgl. hierzu auch Horn, 1983). In diesem Sinne ist Gewissenhaftigkeit eher als adaptive Eigenschaft zu verstehen, in dem Sinne, dass man sich bzw. seinen Arbeitsaufwand der Umgebung anpasst, wenn man nicht über eine genügend schnelle Auffassungsgabe verfügt.

Die folgenden Abschnitte (2.3 ff) liefern einen Überblick über Persönlichkeitsmerkmale in der beruflichen Eignungsdiagnostik.

2.3 Persönlichkeitsmerkmale in der Eignungsdiagnostik

Die Persönlichkeitspsychologie ist ein Feld, das eine Person als Ganzes beschreibt. Mayer (2005) definiert *Persönlichkeit* wie folgt: „*Personality is the organized, developing system within the individual that represents the collective action of his or her motivational, emotional, cognitive, social-planning, and other psychological subsystems*“ (S. 296). Persönlichkeitseigenschaften sind Dimensionen individueller Unterschiede in Bezug auf kognitive und emotionale Zustände sowie Verhaltensmuster, die beispielsweise Gesundheit und Lebensalter beeinflussen (Terracciao, Löckenhoff, Zonderman, Ferrucci & Costa, 2008). Handlungen und Reaktionen scheinen zumindest teilweise von der zugrundeliegenden Persönlichkeit gesteuert zu werden. Persönlichkeitsmerkmale sollten demzufolge auch einen Einfluss darauf haben, wie sich eine Person bei der Arbeit normalerweise verhält. Das typische Verhalten einer Person sollte sich somit auch auf die erzielten Ergebnisse auswirken.

Persönlichkeitseigenschaften werden auch als *Dispositionen* oder *Traits* verstanden, d.h. als überdauernde und relativ stabile Merkmale, die das Verhalten eines Menschen in konkreten Situationen beeinflussen (Backhaus, 2004b). Das Ziel der Persönlichkeitstheorien ist es somit, die grundlegenden und bei allen Menschen vorhandenen Dimensionen zu bestimmen und beschreiben, in denen Personen sich individuell voneinander unterscheiden.

Die klassische Methode, Persönlichkeitseigenschaften zu erfassen, ist mittels psychologischer Testverfahren (Schuler & Höft, 2006). Den Autoren zufolge sind psychologische Tests „*standardisierte, routinemäßig anwendbare Verfahren zur Messung individueller Verhaltensmerkmale, aus denen Schlüsse auf Eigenschaften der betreffenden Person oder auf ihr Verhalten in anderen Situationen gezogen werden können*“ (S. 104). Fragebögen bilden einen praktischen und zuverlässigen Rahmen für die Erfassung bzw. Messung einer großen Palette stabiler individueller Unterschiede, die eine große Zeitspanne an Lebensereignissen umfassen, die in einer Laborstudie nur schwer hergestellt werden kann (DeYoung, 2010).

Das Persönlichkeitskonstrukt selbst ist allerdings zu komplex, um als Ganzes erfasst und erforscht zu werden (Mayer, 2005). Philosophen und Psychologen seit Aristoteles' Zeiten postulieren, dass die Persönlichkeit unterteilt werden muss, um sinnvoll erforscht werden zu können. Bereits in der Antike teilte Hippokrates (460 bis 377 v. Chr.) die Menschen in vier Temperamentstypen ein. Auch in der Psychologie hat die Persönlichkeitsforschung eine lange Tradition. Auf die scheinbar einfache Frage nach dem Aufbau der Persönlichkeit gab bzw. gibt es in der Psychologie viele verschiedene Antworten. Psychodynamische Psychologen wie Freud unterteilten die Persönlichkeit in drei Bereiche: *Es*, *Ich* und *Über-Ich*. Humanisten nahmen eine Zweiteilung in das *falsche* und *reale Ich* vor, während die Begründer der sozial-kognitiven Theorie Persönlichkeit mit Hilfe von Enkodierungen, Erwartungen und Überzeugungen sowie Selbstregulierungsplänen zu erfassen versuchten. Trait-Theoretiker erfassen Persönlichkeit mit Hilfe der *Big 5* bzw. des *Fünf-Faktoren-Modells* (vgl. Goldberg, 1993; McCrae & Costa, 1987). Die Unterteilung der Persönlichkeit geschieht im letztgenannten

Ansatz eher strukturell, da man auf der Suche nach möglichst stabilen, überdauernden und festen Persönlichkeitseigenschaften ist.

Zu den Begründern des heute traditionell als „Persönlichkeitspsychologie“ verstandenen Trait-Ansatzes zählen Gordon Allport und H. Odbert (1936) sowie Raymond B. Cattell (1943), die mit ihrem *psycho-lexikalischen Ansatz* den Grundstein für die Untersuchung und Beschreibung der menschlichen Persönlichkeit legten.

Der psycho-lexikalische Ansatz entwickelte sich aus der Sedimentationshypothese, die besagt, dass alle Aspekte bedeutsamer, nützlicher oder interessierender individueller Unterschiede Einfluss in die Sprache gefunden haben und somit als Adjektiv in dieser auftauchen. Je bedeutsamer eine solche individuelle Differenz, desto größer ist die Wahrscheinlichkeit, dass für diese ein gesondertes Wort hervorgebracht wurde (Bühner, 2011). Allport und Odbert (1936) fassten 18.000 Begriffe – überwiegend Adjektive – aus einem Wörterbuch zu vier Kategorien zusammen. Eine gesonderte Kategorie enthielt die sogenannten „persönlichkeitsbeschreibenden“ Begriffe, welche *„konsistente und stabile Formen der Anpassung eines Individuums an seine Umgebung beschreiben und hinsichtlich ihrer wertenden Implikationen möglichst neutral erscheinen“* (Borkenau & Ostendorf, 1993, S. 6). Cattell (1943) führte diese Arbeit fort, indem er die Begriffe mittels faktoren-, korrelations- sowie inhaltsanalytischer statistischer Verfahren auf 35 Variablencluster und später zwölf Faktoren reduzierte. Fiske (1949) versuchte, diese Arbeit zu replizieren und fand jedoch fünf zentrale Eigenschaften, die später auch von einer Vielzahl anderer Forscher unabhängig von der untersuchten Stichprobe, Beobachtern, Instrumenten, Methoden der Faktorenextraktion und -rotation sowie – mit wenigen Einschränkungen – dem Kulturraum gefunden wurden (z.B. Digman, 1990; Goldberg, 1993; McCrae & Costa, 1987) sowie anhand anderer Variablensätze repliziert werden konnten. Auch die faktorenanalytische Untersuchung bestehender Fragebogenverfahren ergab oftmals fünf globale Dimensionen zur Beschreibung der Persönlichkeit (Borkenau & Ostendorf, 2008).

Diese Fünf-Faktoren-Struktur hat als Theorie auf diese Weise als *Fünf-Faktoren-Modell* (FFM) oder auch als *Big 5* im Bereich der Persönlichkeitseigenschaften mittlerweile eine Hegemonialstellung erreicht (Schuler & Höft, 2001). Insbesondere die Forschungsarbeiten von Costa & McCrae (1982; 1992) haben zur Bekanntheit und häufigen Verwendung des FFM sowohl in der Forschung und zunehmend auch in der Praxis beigetragen.

In den 1970er Jahren erlitt die Persönlichkeitspsychologie in der Forschung und Praxis jedoch einen Rückschlag, was auf verschiedene Ursachen zurückgeführt werden kann. Vor allem wurde basierend auf narrativen Zusammenfassungen der bestehenden Forschungsarbeiten in diesem Bereich in der Forschung die geringe prädiktive Validität der Persönlichkeitseigenschaften bemängelt (Guion & Gottier, 1965). Diese Forschungsarbeiten postulierten, dass die Beziehung zwischen Persönlichkeitsmerkmalen und erfolgreicher Leistung am Arbeitsplatz nicht bedeutsam sei. Daraufhin wurde der Nutzen von Persönlichkeitsschätzungen für die Vorhersage effektiven Arbeitsverhaltens infrage gestellt. Der Hauptgrund für diesen frühen Pessimismus lag allerdings in der verwendeten mangelhaften Methodologie (Lord, 2011; Schuler & Höft, 2006). So hatten die Forscher zwar die Beziehung zwischen einer großen Bandbreite von individuellen Persönlichkeitsskalen und verschiedenen Aspekten der Arbeitsleistung untersucht. Allerdings gab es hierfür kein eindeutiges Klassifikationssystem. Einzelne Skalen unterschieden sich trotz gleicher Bezeichnungen inhaltlich deutlich voneinander oder waren trotz verschiedener Benennungen und folglich unterschiedlicher Zuordnungen inhaltlich gleich. Zudem herrschte allgemein Unklarheit darüber, welche Traits überhaupt gemessen wurden. Es wurde beispielsweise auch nicht zwischen Konstrukt- und Fragebogenebene unterschieden (Barrick, Mount & Judge, 2001). Darüber hinaus wurde ähnlich einem „Streuprinzip“ geforscht, d.h. alle verfügbaren Eigenschaften wurden mit allen möglichen gemessenen Leistungskriterien korreliert, ohne diesem Vorgehen eine theoretische Basis zu-

grunde zu legen. Die Metaanalysen zu dieser Forschung waren außerdem eher narrativ denn quantitativ und beinhalteten keine Korrekturen um statistische Artefakte, was zu stark verringerten Vorhersagewerten führte (Barrick, Mount & Judge, 2001; Lord, 2011; Schuler & Höft, 2006). Daher wurde die Erforschung konsistenter Zusammenhänge zwischen Persönlichkeitseigenschaften und Leistungskriterien erschwert und wenig Fortschritt beim Verständnis der Zusammenhänge erzielt.

Diese negative, auf falschen Methoden basierende Zusammenfassung der Arbeiten zur praktischen Nützlichkeit der Persönlichkeitspsychologie wurde das folgende Vierteljahrhundert nicht weiter überprüft. Zudem wuchs in dieser Zeit die Beliebtheit der Sozialpsychologie stark an, die die Persönlichkeitspsychologie als Konkurrenz ansah (Mayer, 2005). Auch aufgrund der andauernden Interaktionismusdebatte (Anlage vs. Umwelt) und der zunehmenden Sensibilisierung für Eingriffe in die Privatsphäre wurde der Einsatz von Persönlichkeitsmaßen in der Vergangenheit kritisch beäugt (Schuler & Höft, 2006). Darüber hinaus könnten demografische Entwicklungen an den Universitäten diesen Trend zusätzlich verstärkt haben, da in dieser Zeit die Persönlichkeitsforscher aus Altersgründen ausschieden, an deren Stelle aber keine neuen Persönlichkeitsforscher eingestellt wurden und die Forschung zu diesem Thema aus diesem Grund ebenfalls weniger wurde. Zudem ist es möglich, dass die nicht mehr zeitgemäße Art, in der über diesen Persönlichkeitsansatz und die zugehörigen Theorien gelehrt wurde, zur schwindenden Popularität beigetragen hat. Dieser Trend scheint in den neueren Lehrbüchern jedoch wieder umgekehrt zu sein.

Barrick, Mount und Judge (2001) sehen die Jahre bis in die 1980er als die erste Phase der Erforschung von Persönlichkeitseigenschaften und Arbeitsleistung an, in der v.a. Primärstudien zum Zusammenhang verschiedener Skalen aus unterschiedlichen Persönlichkeitsinventaren mit verschiedenen Aspekten von Arbeitsleistung durchgeführt wurden. Erst in den 1990er Jahren erwachte das Interesse an Persönlichkeitstests wieder. Es etablierte sich ein Persönlichkeitsmodell mit fünf Dimensionen, das von verschiedenen, unabhängigen Autoren bzw. Autorengruppen vertreten wurde (Borkenau & Ostendorf, 1991). In seiner einflussreichen Arbeit benannte Goldberg (1993) diese fünf Faktoren das erste Mal als *Big 5*, während Costa und McCrae (1987) den Begriff *Fünf-Faktoren-Modell* (FFM) benutzten. Das Aufkommen des FFM oder einiger Varianten des Modells gemeinsam mit einigen bedeutenden Einzelstudien, die beispielsweise die inkrementelle Validität von Persönlichkeitstests belegten, werden allgemein als Ursachen für das Wiederaufleben des wissenschaftlichen Interesses am traittheoretischen Ansatz angesehen. Ab den 1990er Jahren wurden außerdem aussagekräftige Überblicksarbeiten erstellt, die metaanalytische Methoden verwendeten, um die Daten aus den verschiedenen Studien systematisch zusammen zu fassen und z.B. Artefakt-Korrekturen vorzunehmen.

Es gibt nur wenige Themen, die in der Psychologie solch großes Interesse und so viel Forschung nach sich gezogen haben, wie das FFM. Es ist keine Frage, dass die *Big 5* einflussreich sind, in dem Sinne, dass sie viel Varianz der allgemeinen Persönlichkeit aufklären (Jackson, Paunonen, Fraboni & Goffin, 1996). Die Frage ist aber, ob diese fünf Faktoren ausreichend sind, um eine umfassende Repräsentation der Persönlichkeitsstruktur darzustellen.

Trotz der Fülle an unabhängigen und übereinstimmenden Forschungsergebnissen, die für das FFM sprechen, ist die Struktur der Persönlichkeit bis heute umstritten (vgl. Ackerman & Heggestad, 1997; Backhaus, 2004b). Die *Big 5* wurden immer wieder in Persönlichkeitsinventaren sowie in Studien mit lexikalischen Ansätzen anhand von Faktoranalysen nachgewiesen bzw. aufgefunden. Aber das Replizieren von fünf Faktoren ist nicht dasselbe wie keine weiteren Faktoren zu finden. Wenn man nach fünf Faktoren sucht, wird man auch fünf Faktoren finden (Jackson et al., 1996).

Je nach Forschungsrichtung und -tradition existieren in der differentiellen Psychologie verschiedene Faktormodelle der Persönlichkeit, die beispielsweise drei (Larson & Borgen, 2006), sechs (Andresen, 2000; Ashton & Lee, 2007; Becker, 1996; Saucier, 2009), sieben (Saucier, 2003) oder acht Faktoren (Andresen, eingereicht) gefunden haben. Vor allem lassen sich in der Literatur Hinweise darauf finden, dass ein sechster Faktor „Gefühlsbetontheit“ oder „Risikobereitschaft“ nicht mit erfasst wird (Andresen, 1995; Becker, 1996).

Becker (1996) überprüfte in einer Studie die lexikalische Struktur des FFM bzw. dessen Angemessenheit für die deutsche Sprache anhand einer studentischen Stichprobe ($N=127$). Die Probanden beantworteten verschiedene Persönlichkeitsfragebögen und wurden auch von ihren Eltern, Geschwistern und/ oder Bekannten mit Hilfe der verwendeten Fragebögen eingeschätzt. Der Autor interessierte sich dafür, inwiefern das NEO-Fünf Faktoren Inventar (NEO-FFI; Borkenau & Ostendorf, 1993) die Varianz in den Persönlichkeitsvariablen vollständig erfassen kann. Die Ergebnisse der Hauptkomponentenanalyse auf Itemebene sprachen für eine Faktorenlösung mit mehr als fünf Dimensionen für eine optimale Varianzaufklärung. Das FFM sollte dem Autor zufolge somit um einen sechsten Faktor der „Gefühlsbetontheit vs. Verstandesbetontheit“, der auch Aspekte der Risikobereitschaft enthält, erweitert werden.

Der Fünf-Faktoren-Ansatz hat weitere Nachteile, da Eigenschaften wie „Locus of Control“ oder „Sensation Seeking“ nicht beachtet werden (Mayer, 2005). Der lexikalische Ansatz hat diese Konstrukte nicht als wichtig erkannt, da diese in der Sprache nicht explizit vorhanden sind. Zudem haben Becker (1996) zufolge die Forscher des lexikalischen Ansatzes bei der Auswahl der relevanten Eigenschaften eine zum Teil sehr rigorose Auswahlstrategie gehabt. Lediglich 8 % der gefundenen Adjektive wurden zur Analyse beibehalten und die Beurteiler hatten Probleme, die Adjektive in verschiedene Kategorien einzuordnen. Abwandlungen oder andere Faktorenmodelle haben es bisher allerdings verfehlt, ihre Überlegenheit gegenüber dem FFM zu beweisen. Die Tatsache, dass auch andere Faktoren-Modelle gefunden wurden, ist allerdings kein notwendiger Grund, dass die Big 5 keinen guten Rahmen für die Erfassung der Persönlichkeit liefern.

Trotz aller Kritik hat das Fünf-Faktoren-Modell im Bereich der allgemeinen Persönlichkeitskonstrukte „im Laufe des letzten Jahrzehnts eine Hegemonialstellung erreicht“ (Schuler & Höft, 2006, S. 117), was v.a. auch der regen Publikationstätigkeit der beiden Originalautoren Paul T. Costa und John R. McCrae zuzuschreiben ist. Den Autoren zufolge kann der Streit, wie viele Faktoren die Persönlichkeit eines Menschen ausmachen, als ausgetragen angesehen werden. Zumal auch verschiedene Forschungsansätze zu vergleichbaren Ergebnissen kommen. So finden sich vergleichbare Befunde sowohl in den lexikalischen Ansätzen, als auch in deren faktorenanalytischen Replikationen und interkulturellen Vergleichsstudien. Vor allem sich geografisch und historisch nahe stehende Kulturen tendieren dazu, ähnliche Persönlichkeitsausprägungen aufzuzeigen, wie z.B. Deutschland und Österreich (Heine, Buchtel & Norenzayan, 2008).

Darüber hinaus handelt es sich beim FFM um ein sehr kompaktes Modell, mit dessen Aufkommen es möglich wurde, einzelne Skalen in Form eines umfassenden Klassifikationssystems zu kategorisieren und so Ordnung in die zuvor eher ungeordneten Persönlichkeitstheorien zu bringen (DeYoung, 2010; Poropat, 2009; Terracciano et al., 2008). Die Big 5 bieten somit einen hilfreichen Rahmen für die Erforschung der Persönlichkeit (Dudley et al., 2006). Zusätzlich hat die zunehmende Verfeinerung statistischer Techniken es den Forschern ermöglicht, Ergebnisse über verschiedene Studien hinweg genauer zu interpretieren (Lord, 2011). Das FFM wird daher auch als Orientierungsrahmen für die angewandte Forschung in der Eignungsdiagnostik angesehen.

Eine vorsichtige Analyse der englischen Sprache resultierte in fünf voneinander unterscheidbaren Faktoren: *Neurotizismus (Neuroticism)*, *Extraversion (Extraversion)*, *Offenheit für Er-*

fahrungen (*Openess to Experiences*), *Verträglichkeit* (*Agreeableness*) sowie *Gewissenhaftigkeit* (*Conscientiousness*) (Costa & McCrae, 1992; Goldberg, 1993). Eine aktuelle Definition und ausführliche Beschreibung der Konstrukte sowie bisherigen Forschung hierzu kann bei Borkenau und Ostendorf (2008) oder Lord (2011) nachgelesen werden.

Die Persönlichkeitsvariablen sind hierarchisch organisiert und lassen sich jeweils noch weiter in Facetten bzw. Sub-Faktoren unterteilen. Diese Unterteilung ist ebenfalls – je nach Forschungstradition – unterschiedlich und – ähnlich wie die Unterteilung der Persönlichkeit – umstritten (vgl. Andresen, eingereicht; DeYoung, 2010; Roberts, Walton & Bogg, 2005; Saucier & Ostendorf, 1999). Barrick et al. (2001) berichten in ihrer Übersichtsarbeit von verschiedenen Konstrukten zur Substruktur des FFM. Unabhängige Forschergruppen schreiben den Big 5 demzufolge zwischen 18 bis 45 Facetten zu. Costa und McCrae (1992) ordnen jeder ihrer fünf Persönlichkeitsdimensionen jeweils sechs Facetten zu.

DeYoung (2010) nennt drei Ebenen, auf denen Persönlichkeitseigenschaften analysiert werden können. Zum einen gebe es die *Traitebene*, die relativ stabile Verhaltens-, Motivations-, Emotions- und Denkmuster bei einer Person erfasst – unabhängig von der sozialen Herkunft oder kulturellen Einflüssen. Jeder Trait kann dem Autor zufolge in vielen Situationen in jeder Kultur beobachtet werden. Die zweite Ebene sind charakteristische *Adaptionen*, die sich auf die Anpassungsfähigkeit einer Person an sich verändernde Situationen und Umstände beziehen. Die dritte Ebene ist die individuelle *Lebensgeschichte* einer Person. Die beiden letzten Ebenen beschreiben somit eher die Antworten eines Individuums auf die jeweiligen Lebensumstände der Person. Sie sind für das Verständnis des Verhaltens einer Person bedeutend, allerdings erschwert ihre Komplexität ihre Erforschung. Entsprechend hat sich die klassische Persönlichkeitspsychologie vor allem auf die Erforschung grundlegender Persönlichkeitseigenschaften fokussiert.

Zur Substruktur der Dimensionen muss zudem angemerkt werden, dass die Unterteilung der Dimensionen in Subfacetten nicht einfach aufgebaut ist. So sind viele Facetten und Faktoren interkorreliert. Dies wird der Einfachheit in der Regel aber nicht dargestellt. Die Existenz der Facetten deutet darauf hin, dass einige Zusammenhänge nur einzelne Facetten betreffen, andere wiederum mit mehreren Aspekten oder Dimensionen interferieren (DeYoung, 2010).

DeYoung, Quilty und Peterson (2007) analysierten verschiedene Persönlichkeitsinventare, die das FFM erfassen, und berichten von einer zweifaktoriellen Subdomänenstruktur der fünf Dimensionen. Die Forschergruppe um Colin DeYoung schlägt eine hierarchische Ordnung der Persönlichkeitsdimensionen vor, die auf oberster Ebene zwei *Metatraits* (*Stabilität* bzw. *Stability* und *Plastizität* bzw. *Plasticity*) postuliert, auch wenn diese ursprünglich als unabhängige Faktoren und höchste Dimensionsebene konstruiert wurden. Zum Metatrait *Stabilität* (α -Faktor) zählt DeYoung (2010) die drei *Dimensionen* bzw. Faktoren *Neurotizismus*, *Verträglichkeit* und *Gewissenhaftigkeit*. Dem Metatrait *Plastizität* (β -Faktor) gehören die beiden Dimensionen *Extraversion* und *Offenheit für Erfahrungen* an. Alle fünf Faktoren besitzen eine zweifaktorielle Substruktur, deren Eigenschaften DeYoung et al. (2007) als *Aspekte* bezeichnen. Für die Gewissenhaftigkeitsdimension sind dies beispielsweise die Aspekte *Fleiß* (*Industriousness*) und *Ordnung* (*Orderliness*). Den Aspekten ordnen die Forscher wiederum auf der spezifischen Ebene die sog. *Facetten* zu. Es handelt sich somit um ein integratives Modell der Persönlichkeit, das eine Möglichkeit aufzeigt, die verschiedenen Strukturansätze miteinander zu integrieren. Abbildung 2-2 veranschaulicht das hierarchische Modell der Persönlichkeit.

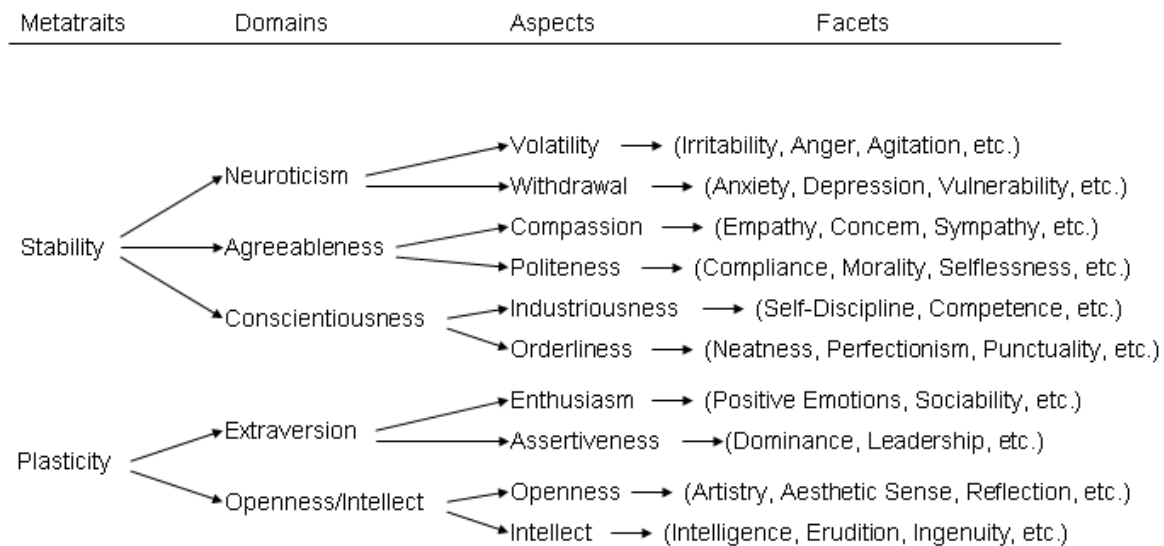


Abbildung 2-2: Hierarchisches Modell der Persönlichkeit (Quelle: DeYoung, 2012)

Neben der Faktorenstruktur wurde seit den 1990er Jahren aber auch wieder vermehrt die Vorhersagekraft von Persönlichkeitsfaktoren in Bezug auf berufliche Leistungs- und Erfolgskriterien erforscht. Bereits Anfang der 1990er Jahre entstanden hierzu auch die ersten Metaanalysen. Diese bestätigten zum ersten Mal eindeutig die günstige, wenn auch moderate Vorhersagekraft von Persönlichkeitseigenschaften für die Vorhersage von Berufserfolg und Arbeitsleistung (Risavy & Hausdorf, 2011). Barrick und Mount (1991) publizierten die erste Metaanalyse zur Kriteriumsvalidität von Big 5-Persönlichkeitseigenschaften. Die Forscher fanden auf diese Weise, dass Gewissenhaftigkeit die einzige Dimension der Big 5 ist, die sowohl berufs- als auch kriteriumsübergreifend konsistente Zusammenhänge mit Arbeitsleistung aufweist. Sie berichteten von einer prädiktiven Validität von $r = .22$ von Gewissenhaftigkeit in Bezug auf Berufserfolg. Nachfolgende Metaanalysen von Salgado (1997) sowie Anderson und Viswesvaran (1998) ergaben, dass sowohl Gewissenhaftigkeit als auch Ausgeglichenheit hoch mit Arbeitsleistung korrelierten. Allgemein haben sich diese beiden Dimensionen als zuverlässige Vorhersagefaktoren erwiesen, die auch inkrementelle Validität über kognitive Fähigkeiten bei der Vorhersage von Berufserfolg besitzen (vgl. McCrae & Costa, 1987; Salgado, 1998).

Bis zur Jahrtausendwende entstanden 15 Metaanalysen zum Zusammenhang von Persönlichkeitseigenschaften und Arbeitsleistung, die Barrick, Mount und Judge (2001) in ihrer Metaanalyse 2. Ordnung wiederum zusammenfassten. Den Autoren zufolge hatte die Verwendung des FFM es ermöglicht, spezifischere Hypothesen über die prädiktive Validität von Persönlichkeitskonstrukten im Arbeitssetting zu formulieren. Die meisten Metaanalysen deuten den Ergebnissen zufolge an, dass die zwei Faktoren Neurotizismus und Gewissenhaftigkeit negativ bzw. positiv mit Arbeitsleistung über fast alle Tätigkeiten hinweg korreliert sind. Wiederrum die meisten dieser Metaanalysen ergaben, dass Gewissenhaftigkeit von allen Persönlichkeitsdimensionen die höchsten Zusammenhänge mit Berufsleistung aufweist. Gewissenhaftigkeit korreliert den Ergebnissen von Barrick et al. (2001) zufolge mit Arbeitsleistung allgemein ($r = .20$), aber auch mit spezifischeren Kriterien wie Vorgesetztenbeurteilungen ($r = .26$), objektiven Leistungskriterien ($r = .19$), Fortbildungsleistung ($r = .23$) und Teamwork ($r = .23$). In Bezug auf verschiedene Berufsgruppen zeigten sich vergleichbare Ergebnisse der prädiktiven Validität von Gewissenhaftigkeit für Verkäufer ($r = .21$), Manager ($r = .21$), Experten ($r = .20$), Polizisten ($r = .22$) sowie gelernte oder angelernte Arbeiter ($r = .19$). Salgado (1997) führte eine Metaanalyse nur mit Studien aus Europa bzw. mit Ergebnissen von europäischen Stichproben durch und fand ähnliche Ergebnisse. Auch hier ergab sich Gewissenhaftigkeit als

stärkster und zuverlässigster Prädiktor von Berufserfolg. Dies beweist, dass die Ergebnisse auch kulturübergreifend gültig sind.

Es bleibt festzuhalten, dass Gewissenhaftigkeit ein valider Prädiktor zur Vorhersage von beruflicher Leistung über Kriteriumstypen, Kulturen und verschiedene Berufe hinweg ist. Gewissenhaftigkeit scheint Barrick et al. (2001) zufolge die traitbezogene Variable zu sein, nach der von Arbeitspsychologen so lange gesucht wurde, die in zukünftigen Theorien zur Vorhersage von Berufserfolg bzw. Erklärung von Arbeitsleistung eine zentrale Rolle spielen sollte.

Poropat (2009) führte eine Metaanalyse zum Zusammenhang von Persönlichkeit und Schulleistung durch. Eine der ersten Anwendungen traitbasierter Persönlichkeitsbewertung war die Vorhersage schulischer Leistung. Bereits Anfang des letzten Jahrhunderts fanden Forscher heraus, dass Persönlichkeitsmessungen mit Schulleistung korrelierten. Allerdings krankten diese ersten Forschungsversuche wie bereits zuvor angesprochen an methodischen Fehler. Die Ergebnisse in diesem Bereich waren daher recht inkonsistent, weshalb auch in diesem Bereich die frühen Ansätze, die Forschungsergebnisse zusammenzufassen, zu dem Schluss kamen, dass keine oder nur tendenziell prädiktive Effekte vorhanden waren. So nutzte Hough (1992) in seiner Übersichtsarbeit eine Vielzahl verschiedener Persönlichkeitsmaße, deren Faktoren sogar interkorreliert waren. Der Autor fand eine mittlere Korrelation von Persönlichkeit mit Schulleistung von $r = .39$. Darüber hinaus vermischte er auch verschiedene Leistungsmaße (z.B. Durchschnittsnoten und Anwesenheitszeiten), die teilweise ebenfalls stark interkorrelieren, um die Zusammenhänge zu errechnen. Zudem wurde bemängelt, dass den Zusammenhängen keine besondere Theorie zugrunde lag.

Die Ergebnisse der existierenden Metaanalysen zum Zusammenhang von Persönlichkeitseigenschaften und Schulleistung sind vergleichbar zu den Ergebnissen der Metaanalysen im beruflichen Kontext. Wie auch bei Arbeitsleistung weist die Gewissenhaftigkeitsdimension die höchsten Zusammenhänge mit Schulleistung auf und stellt (ab weiterführenden Schulen) sogar eine Alternative zur bisher in der Regel verwendeten Vorhersagevariable Intelligenz dar (O'Connor & Paunonen, 2007; Poropat, 2009).

Die Passung des FFM für die berufliche Eignungsdiagnostik mag nicht sofort erkennbar sein, da es ursprünglich als rein deskriptives Modell gedacht war, das keinen Aufschluss darüber gab, warum genau diese fünf Dimensionen die großen Dimensionen der Persönlichkeit bzw. die Kovariation zwischen diesen Traits beschreiben (DeYoung, 2010). Persönlichkeitseigenschaften sind jedoch nachweislich Indikatoren für Verhaltensweisen, die die Kandidaten auch bei der Arbeit zeigen, und daher interessant für die Vorhersage von Arbeitsverhalten bzw. Leistung und somit interessant für die Praxis der Personalauswahl und -entwicklung (Lee, Piccone, Isaacson, Trejo & Griffith, 2010).

Anhand einer quantitativen Analyse der bestehenden Literatur bzw. deren Zitationshäufigkeiten von Landers (2009) ist erkennbar, dass im Bereich der Arbeitspsychologie vor allem Artikel zu Persönlichkeitseigenschaften viel zitiert werden, wie beispielsweise die Metaanalyse von Barrick und Mount (1991). Dieser Befund spiegelt den allgemeinen Trend in diesem Bereich wider, in dem nach wie vor Persönlichkeitseigenschaften im Fokus der Aufmerksamkeit stehen.

Barrick und Mount (2004) führten in ihrer Metaanalyse zur prädiktiven Validität von Persönlichkeitseigenschaften sechs Gründe an, warum Persönlichkeit im Arbeitskontext wichtig ist.

1. Für Manager sind Persönlichkeitseigenschaften im Rahmen von Einstellungsentscheidungen wichtig. Sie gewichten diese genauso hoch wie kognitive Fähigkeiten. Dies scheint intuitiv logisch. Es ist in der Tat schwer, einen Manager zu finden, der sagen würde, dass er jemanden einstellen würde, der unzuverlässig, faul, impulsiv und wenig zielstrebig, also wenig gewissenhaft, ist. Es werden somit grundsätzlich eher Bewerber eingestellt, die

über eine hohe Gewissenhaftigkeit verfügen. Auch wenn dies auf den ersten Blick nicht sehr überraschend ist, wurde diese Tatsache zwischen den Jahren 1965 und 1990 eher verleugnet.

2. Die Vielzahl an Metaanalysen zu prädiktiven Validität von Persönlichkeitseigenschaften im Kontext der Berufseignungsdiagnostik hat signifikant dazu beigetragen, die Zusammenhänge zwischen Persönlichkeitseigenschaften (vor allem die der Big 5) und Arbeitsleistung aufzuklären. Neben Neurotizismus bzw. Ausgeglichenheit hat sich hierbei vor allem Gewissenhaftigkeit als stabiler und zuverlässiger Prädiktor herauskristallisiert.
3. Jede Schätzung eines Korrelationskoeffizienten zur Vorhersage von Arbeitsleistung stellt eine Unterschätzung des Zusammenhangs dar, da dieser nur die Validität eines einzelnen Kriteriums misst. Scheinbar validere Maße, wie z.B. allgemeine kognitive Fähigkeiten, sind hingegen Maße, die aus der Aggregation verschiedener Konstrukte, wie z.B. verbaler, mathematisch-rechnerischer und räumlicher Fähigkeiten sowie Fachkenntnissen, bestehen. Daher sollte man zur Bestimmung der Validität von „Persönlichkeit“ diese auch aus verschiedenen Merkmalen zusammensetzen, um vergleichbare Werte zu erhalten. Untersuchungen, in denen ein solches Vorgehen gewählt wurde, berichten von deutlich höheren Zusammenhängen von Persönlichkeit und Arbeitsleistung. Beispielsweise fanden Judge, Heller und Mount (2002) einen Zusammenhang von $r = .41$ zwischen den Big 5 und Arbeitszufriedenheit. Man sollte somit nicht nur die Traits, sondern auch die Kombination von Traitausprägungen bei den einzelnen Personen beachten.
4. Verschiedene Studien zeigen einen inkrementellen Validitätszuwachs von Persönlichkeitsmaßen über allgemeine kognitive Fähigkeiten und biografische Daten hinaus (Schmidt & Hunter, 1998).
5. Über die Zeit hinweg sind Persönlichkeitseigenschaften sehr gute Prädiktoren für Berufserfolg. So berichten Judge, Higgins, Thoresen und Barrick (1999) von einem Zusammenhang von $r = .49$ zwischen Persönlichkeitseigenschaften (z.B. Gewissenhaftigkeit), die in der Kindheit erfasst wurden, und Berufserfolg, der bis zu 50 Jahre später erhoben wurde. Die gemeinsame Korrelation aller Big 5 Persönlichkeitsmerkmale betrug $r = .60$ bezogen auf beruflichen Status und Einkommen, die 30 bis 50 Jahre später gemessen wurden. Dies weist darauf hin, dass – auch wenn die konkurrente Validität zwischen Gewissenhaftigkeit und Vorgesetztenbeurteilung nur moderat ist – die prädiktive Validität deutlich über diesen Effekten liegt und man somit von einer Akkumulation der Persönlichkeitseffekte ausgehen kann. Dies kann auch auf das reziproke Verhältnis von Gewissenhaftigkeit und Berufserfolg zurückzuführen sein. So ist es möglich, dass der größere berufliche Erfolg von sehr gewissenhaften Menschen dieses Verhalten wiederum verstärkt.
6. Zudem spricht die Vielzahl der Befunde, die Zusammenhänge zwischen relevanten Arbeitsverhaltensweisen und insbesondere der Dimension Gewissenhaftigkeit gefunden haben, die nicht immer auch von GMA vorhergesagt werden können, weiter für die große Bedeutsamkeit von Persönlichkeitseigenschaften in Bezug auf den Erfolg einer Person und letztlich auch des Unternehmens.

Diese sechs Punkte lassen sich auch auf die Vorhersage von Ausbildungserfolg übertragen. Bipp (2010) zufolge können Persönlichkeitseigenschaften auch bestimmen, für welche Art von Job sich eine Person bewirbt bzw. interessiert, da es sich um ein reziprokes Zusammenspiel von Eigenschaften und Situationen handelt. Während Persönlichkeitsmerkmale als stabile Eigenschaften angesehen werden, die das Verhalten einer Person in einer bestimmten Situation beeinflussen, wird ihnen auch zugeschrieben, dass sie mitbestimmen, welche Situationen von einer Person überhaupt aufgesucht werden.

De Raad und Schouwenburg (1996) kritisierten diesbezüglich, dass Persönlichkeitseigenschaften jedoch immer erst am Ende der Liste aller Überlegungen zum Thema Lernen und Erziehung kommen würden und fordern dass „*personality should take a more prominent role*

in future theories of academic performance, and not merely as an adjunct to intelligence“ (S. 39).

Talentmanagement, d.h. die Anwerbung, Auswahl, Entwicklung und Bindung besonders geeigneter Mitarbeiter, beginnt damit, dass man die richtigen Menschen für die jeweiligen Positionen findet. Da die Anstellung passender Mitarbeiter ausschlaggebend für den Erfolg eines Unternehmens ist bzw. die Anstellung nicht geeigneter Personen (= eine falsch-positive Auswahlentscheidung) einen negativen Einfluss auf den Erfolg eines Unternehmens hat, sollte die Personalauswahl sich – auch im Bereich der Auswahl von Auszubildenden – an den wissenschaftlich gesicherten Standards zur prädiktiven Validität der verschiedenen Auswahlmethoden orientieren. Die negativen Konsequenzen, die mit falsch-positiven Entscheidungen verbunden sind, sind substanziell. Unqualifizierte Mitarbeiter können teure Fehler machen, eine enge Überwachung benötigen und einen erhöhten Bedarf an Training haben, um die notwendigen Qualifikationen (nachträglich) zu erwerben. Daher liegen die ökonomischen Vorteile richtiger Auswahlentscheidungen auf der Hand. Persönlichkeitstests sind eine nützliche und hilfreiche Erweiterung in Auswahlprozessen (Risavy & Hausdorf, 2011). Vor allem Gewissenhaftigkeit hat sich vorrangig als diejenige Dimension herausgestellt, die berufs- und hierarchieübergreifend die höchste Vorhersagekraft beruflichen Erfolges besitzt.

Allerdings zeigt sich bei der Integration der wissenschaftlich gesicherten Befunde in die Praxis der Personalauswahl weiterhin eine große Diskrepanz. Unabhängige Studien berichten, dass Organisationen immer mehr Persönlichkeitstests nutzen, um ihre Personalauswahlentscheidungen zu treffen. Im Laufe der Jahre habe das FFM auch in der eignungsdiagnostischen Praxis eine breite Akzeptanz und Beliebtheit erreicht und sei nicht mehr nur in der Persönlichkeitsforschung, sondern auch in vielen Anwendungsfeldern in Europa und Nordamerika stark verbreitet (Anderson & Ones, 2003; DuVernet, Meade, Coughlin & Kantrowitz, 2010; Rammstedt & John, 2005). Nicht nur in höheren Hierarchieebenen und bei Arbeitnehmern mit Berufserfahrung werden Persönlichkeitstests eingesetzt. Auch in niedrigeren Ebenen und bei Berufseinsteigern, wie beispielsweise bei der Auswahl von Auszubildenden, werden diese Verfahren verwendet. Hülshager und Maier (2008) merkten allerdings an, dass im Gegensatz zu Ländern wie den Niederlanden, Großbritannien oder Frankreich in Deutschland den Persönlichkeitsaspekten in der praktischen Personalauswahl noch immer vergleichsweise wenig Beachtung geschenkt wird.

Schuler und Kollegen (2007) befragten die größten deutschen Unternehmen über die Verwendungshäufigkeit von Personalauswahlverfahren zur internen und externen Bewerberauswahl. Neben der Verwendungshäufigkeit fragten sie die Praktiker auch nach ihrer Einschätzung der Vorhersagekraft, Akzeptanz und Praktikabilität der verschiedenen Auswahlverfahren. Insgesamt antworteten 125 der 550 angeschriebenen Unternehmen (= 22,7 % Rücklaufquote). Die Ergebnisse der Studie besagten, dass bei der Auswahl von Auszubildenden im kaufmännischen und technischen Bereich allgemeine Einstellungsinterviews die am häufigsten eingesetzte Auswahlmethode (99,1 % bzw. 97,9 %) und graphologische Gutachten (0 %) die am wenigsten verwendeten Auswahlmethoden waren. Persönlichkeitstests wurden nur von 6,3 % bzw. 6,4 % der Unternehmen eingesetzt.

Die Autoren verglichen die Daten mit Ergebnissen aus einer früheren Studie zur Verwendungshäufigkeit von Auswahlverfahren in der Praxis. Die Verwendung psychologischer Testverfahren hat sich den Ergebnissen zufolge in den vergangenen 10 bis 15 Jahren nur wenig verändert. Leistungs-, Intelligenz- und Persönlichkeitstests wurden im Jahre 2007 mit 20 % im Vergleich zu 1993 (21 %) insgesamt etwas weniger häufig verwendet. Tabelle 2-2 gibt einen Überblick über die Ergebnisse der Studie zur Verwendungshäufigkeit verschiedener eignungsdiagnostischer Verfahren für die Auswahl von kaufmännischen und technischen Auszubildenden in den Jahren 2007 und 1993. Welche Auswahlmethoden im Rahmen der Assessment Center kombiniert wurden, wurde nicht berichtet.

Tabelle 2-2: Einsatzhäufigkeit von Auswahlverfahren für Auszubildende (Auszug aus Schuler et al., 2007)

	kaufm. Auszubildende		techn. Auszubildende	
	2007	1993	2007	1993
Analyse der Bewerbungsunterlagen	99,1 %	99,0 %	97,9 %	93,0 %
Überprüfung zusätzlicher Referenzen	0,9 %	4,0 %	1,1 %	6,0 %
Strukturiertes Interview durch Personalabteilung	67,0 %	62,0 %	67,0 %	48,0 %
Unstrukturiertes Interview durch Personalabt.	15,2 %	28,0 %	13,8 %	28,0 %
Strukturiertes Interview durch Fachabteilung	26,8 %	20,0 %	28,7 %	15,0 %
Unstrukturiertes Interview durch Fachabteilung	17,0 %	19,0 %	21,3 %	25,0 %
Assessment Center	31,3 %	11,0 %	21,3 %	0,0 %
Persönlichkeitstest	6,3 %	10,0 %	6,4 %	10,0 %
Leistungstest (z.B. Bürotest)	36,6 %	40,0 %	33,0 %	40,0 %
Intelligenztest	31,3 %	35,0 %	23,4 %	28,0 %
Graphologisches Gutachten	0,0 %	0,0 %	0,0 %	0,0 %
Medizinische Begutachtung	35,7 %	60,0 %	38,3 %	70,0 %

Auffällig war die Abnahme der Verwendungshäufigkeit von Leistungs-, Intelligenz- und Persönlichkeitstests. Diese Befunde widersprechen den Aussagen von DuVernet et al. (2010) sowie Rammstedt und John (2005), dass Persönlichkeitskonstrukte in der Personalauswahl immer beliebter und mehr verwendet würden. Ursache hierfür könnte die weiterhin geringe Akzeptanz dieser Verfahren bei den Bewerbern sein, auch wenn die Praktikabilität und Validität von den Personalverantwortlichen auf einem mittleren Niveau eingeschätzt werden (Schuler et al., 2007). Verglichen mit anderen Ländern werden den Autoren zufolge in Deutschland vergleichsweise selten Intelligenz- und Persönlichkeitstests eingesetzt. Vor allem in kleinen und mittelständischen Unternehmen gehören freie und unstandardisierte Interviews immer noch zu den Hauptauswahlmethoden. Hierzu muss allerdings angemerkt werden, dass diese Unternehmen in der Regel auch nicht über entsprechend qualifiziertes Fachpersonal (z.B. Psychologen) verfügen, um beispielsweise Persönlichkeitstests angemessen anwenden, auswerten und interpretieren zu können. In Großbritannien (69 %) und Spanien (74 %) gehören Persönlichkeitstests demgegenüber mittlerweile zu den Standardauswahlverfahren (Nachtwei & Schermuly, 2009).

Allerdings wurden vermehrt standardisierte Interviews durchgeführt. Insgesamt gibt es also eine Tendenz hin zu einer Verwendung valider Auswahlverfahren, verbunden mit einer rückläufigen Verwendung invalider Verfahren wie bspw. graphologischer Gutachten. Dies deutet auf einen zumindest teilweise erfolgreichen Transfer der Wissenschaft in die Praxis hin.

Nachtwei und Schermuly (2009) verglichen die Personalauswahlverfahren in großen Konzernen mit denen kleinerer und mittelständischer Unternehmen (KMU). Zudem fragten sie nach Vorbehalten der Personaler ($N=113$) gegenüber psychologischen Eignungstests, die eine konstruktive und objektive Auseinandersetzung mit der Problematik überlagern.

Die Ergebnisse zeigten auf, dass in Deutschland gerade die Kombination von Auswahlverfahren mit der höchsten Vorhersagekraft am wenigsten in der Praxis verwendet wird. Lediglich 13 % der Konzerne und 8 % der kleineren und mittelständischen Unternehmen orientierten sich bei der Zusammenstellung ihrer Auswahlprozeduren an wissenschaftlichen Standards. Stattdessen führten 70 % der KMU und 58 % der Konzerne v.a. unstrukturierte Interviews (s. Tab. 2-3).

Tabelle 2-3: Verwendungshäufigkeit von Auswahlverfahren in deutschen KMU bzw. Konzernen
(aus Nachtwei & Schermuly, 2009)

	KMU	Konzerne	Präd. Validität
Einstellungstests	8 %	13 %	.36
Arbeitsprobe	26 %	42 %	.29
Strukturiertes Interview	30 %	42 %	.25
Fachwissentest	9 %	13 %	.23
Assessment-Center	14 %	21 %	.20
Unstrukturiertes Interview	70 %	58 %	.04

Diese Ergebnisse widersprechen nicht nur der empirischen Befundlage zur Validität von allgemeiner kognitiver Fähigkeit und Persönlichkeitseigenschaften, sondern auch der in vielen Unternehmensauftritten getroffenen Aussage, dass neben Ausbildung und Berufserfahrung auf die „Persönlichkeit“ der Mitarbeiter besonderer Wert gelegt werde. Zudem muss kritisch angemerkt werden, dass Nachtwei und Schermuly (2009) Intelligenz- und Persönlichkeitstests als „Einstellungstests“ zusammenfassten. Dies verschleiert den Einfluss bzw. die Verwendungshäufigkeit der einzelnen Testverfahren. Die wahren Werte könnten bei den Persönlichkeitstests somit noch viel niedriger liegen.

Eine weitere Fragestellung ermittelte, welche Vorbehalte dazu führen, dass Eignungstests in der Praxis der Personalauswahl in Deutschland immer noch eine eher untergeordnete Rolle spielen. Die Autoren sammelten acht typische Vorbehalte von Personalern gegenüber Eignungstests, die im Folgenden kurz dargestellt werden.

1. *Eignungstests sind nur in großen Unternehmen sinnvoll.* Dass Leistungs- und Persönlichkeitstests über eine gute Vorhersagekraft bzw. inkrementelle Validität besitzen und somit auch für die ergänzende Feindiagnostik taugen, ist den meisten nicht bewusst.
2. *Eignungstests sind zu unpersönlich.* Die Frage, ob ein Bewerber ins Team oder zur jeweiligen Kultur eines Unternehmens passt, ist ein wichtiges Passungskriterium. Dieses wollen die Unternehmen bzw. zuständigen Personalern jedoch nicht durch Persönlichkeitstests beantworten, sondern lieber im Interview erfragen bzw. „erspüren“.
3. *Eignungstests schrecken Bewerber ab.* Die Personalern haben Angst, dass durch die Verwendung von Intelligenz- oder Persönlichkeitstests im Rahmen der Auswahlverfahren potenzielle Bewerber von einer Bewerbung bei dem Unternehmen Abstand nehmen könnten.
4. *Eignungstests setzen Bewerber unnötig unter Druck und messen am ehesten Angst.* Viele Personalern fürchten, Eignungstests könnten Angst auslösen, was beispielsweise die Leistung in einem Intelligenztest verringern könne.
5. *Eignungstests erfassen die Komplexität einer Person nicht.* Dieser Vorbehalt, Tests seien nicht aussagekräftig genug, entsteht meist dann, wenn ein Test nicht als Ergänzung, sondern als Ersatz für bestehende Verfahren betrachtet wird.
6. *Eignungstests sind durchschaubar und lassen sich manipulieren.* Viele Personalern besitzen Grundkenntnisse über Fragebogenformate und glauben, dass auch die Bewerber – wie sie selbst – die Absichten der Tests durchschauen könnten. Abschnitt 2.4.3 dieser Arbeit beschäftigt sich ausführlicher mit der Problematik der Verfälschbarkeit von Selbsteinschätzungen.
7. *Eignungstests liefern Ergebnisse, die mit der Realität nichts zu tun haben.* Dieser Vorbehalt zeigt, dass die Ergebnisse aus der Wissenschaft nicht in die Praxis der Personalarbeit einfließen bzw. wie wenig die Personalern über Theorien und aktuelle Forschungsergebnisse zur Personalauswahl wissen. Sofern Personalern überhaupt etwas über wissenschaftliche Ergebnisse im Bereich der beruflichen Eignungsdiagnostik wissen, unterschätzen

sie zumeist den Zusammenhang zwischen Intelligenz und Berufserfolg. Stattdessen denken sie, dass die Testergebnisse nicht generalisierbar seien.

8. *Eignungstests sind für Führungskräfte sinnlos – an der Spitze ist man geeignet.* Die Personalern denken, dass eine hohe Position eines Mitarbeiters aussagekräftig genug ist, um die Eignung der jeweiligen Person zu bestätigen. Streng genommen, bestätigen sie mit dieser Aussage jedoch nur ihre eigene Personalpolitik. Es ist somit möglich, dass dieses Argument eher dazu dient, die eigene Arbeit zu legitimieren bzw. rechtfertigen, und weniger explizit den Nutzen von Persönlichkeitstests mindern soll.

Tabelle 2-4 gibt einen Überblick darüber, wie groß der Grad der Zustimmung der Personalern in den befragten KMU und Konzernen hinsichtlich der zuvor erläuterten acht Vorbehalte gegenüber Eignungstests ist.

Tabelle 2-4: Vorbehalte von Personalern gegenüber Eignungstests (aus Nachtwei & Schermuly, 2009)

Eignungstests...	Personalern, die Eignungstests verwenden	Personalern, die Eignungstests nicht verwenden
... sind nur in großen Unternehmen sinnvoll	51 %	44 %
... sind zu unpersönlich	50 %	49 %
... schrecken Bewerber ab	44 %	65 %
... messen vor allem Testangst	40 %	43 %
... erfassen Komplexität nicht	56 %	61 %
... sind durchschaubar	51 %	56 %
... haben nur geringen Realitätsbezug	35 %	52 %
... sind sinnlos bei Führungskräften	28 %	29 %

Die Ergebnisse deuten an, dass bis auf den letzten Vorbehalt („Eignungstests sind bei Führungskräften sinnlos“), alle Vorbehalte bei ca. der Hälfte der Befragten bekannt waren. Im Schnitt gaben 48 % der Befragten an, dass sie Vorbehalte gegenüber Eignungstests haben. Daraus lässt sich schließen, dass die Vorbehalte gegenüber Eignungstests in der praktischen Personalarbeit sehr weit verbreitet sind. Vor allem Vorbehalte bezüglich der Durchschaubarkeit, hinsichtlich der Angst der Bewerber vor solchen Verfahren und einem angeblich mangelnden Realitätsbezug sind sehr verbreitet. Allerdings wurde nicht erhoben, welche Personalern diese Vorbehalte äußerten. Es ist denkbar, dass vor allem Praktiker ohne psychologischen Hintergrund diese Bedenken äußerten. Allerdings ließ sich zwischen den Aussagen von Nutzern und Nicht-Nutzern der Tests ein bedeutsamer Unterschied erkennen. So gaben jene Unternehmensvertreter, die keine Persönlichkeitstests einsetzten, um fünf Prozentpunkte höhere Vorbehalte an als Personalverantwortliche mit Erfahrung im Umgang mit Persönlichkeitstests. Dies legt den Schluss nahe, dass der praktische Umgang mit Eignungsdiagnostik die Skepsis und Vorurteile diesen gegenüber verringert. Den Autoren zufolge war auch beobachtbar, dass je mehr die Befragten einem Vorbehalt zustimmten, desto eher dieser sie vom Einsatz von Persönlichkeitstests abhielt. V.a. der angeblich mangelnde Realitätsbezug, die geringe Bewerberakzeptanz sowie das Komplexitätsargument entpuppten sich den Autoren zufolge als „Killerkriterien“.

Warum sich diese Vorbehalte in Deutschland so hartnäckig halten, während Unternehmen in anderen Ländern vorurteilsfreier an Tests herangehen, bleibt weiter unklar. Möglicherweise spielt ein mangelnder Zugang zu psychologischer Fachliteratur eine Rolle, eventuell wurden Forschungsbefunde aber bisher auch nicht praxisgerecht genug aufbereitet.

Zukünftig scheint daher ein pragmatischer und praxisorientierter Ansatz in der Eignungsdiagnostik zielführender. Ausgehend von den aktuellen Problemen der angewandten psychologisch-diagnostischen Praxis sind solche Tests zu konstruieren, für die ein Bedarf entsteht.

Erforderlich ist hierzu nicht nur ein stärkeres Hinwenden zu Anforderungsanalysen, etwa in Bezug auf die Beschreibung von Verhaltensweisen in (berufs-) kritischen Situationen (Ortner et al., 2007).

So sollte man die Dimensionen Extraversion, Verträglichkeit und Offenheit, die keine Zusammenhänge über unterschiedliche berufliche Leistungs- und Erfolgskriterien hinweg aufweisen und lediglich einige Beziehungen zu Performanzkriterien bestimmter Berufsgruppen aufzeigen, auch nicht standardmäßig im Rahmen eignungsdiagnostischer Prozesse mit erheben (Hülshager & Maier, 2008). In den Fokus könnten in diesem Zuge demgegenüber aber Konstrukte rücken, die bisher (noch) nicht in der klassischen Persönlichkeitstheorie verankert sind, aber deutliche Zusammenhänge mit Ausbildungs- und Berufserfolg aufweisen, wie beispielsweise das Selbstwirksamkeitskonzept.

Auch sollte man weniger über den Sinn und Zweck von Persönlichkeitsmaßen in der Eignungsdiagnostik diskutieren, sondern sich fragen, wie die Vorhersagekraft der verwendeten eignungsdiagnostischen Verfahren, insbesondere die der Persönlichkeitstests weiter verbessert werden kann. Vor allem sollte der Fokus auf solche Faktoren gelegt werden, die hohe prädiktive Validitäten in Bezug auf die Vorhersage von Berufs- und Schulerfolg besitzen. In den folgenden Abschnitten werden mit der Gewissenhaftigkeitsdimension (Abschnitte 2.3.1) und dem Selbstwirksamkeitskonzept (Abschnitt 2.3.2) zwei solche Faktoren vorgestellt.

2.3.1 Gewissenhaftigkeit

Der am meisten in Bezug auf seine Vorhersagekraft von Arbeitsleistung beforschte Persönlichkeitsfaktor ist die *Gewissenhaftigkeit* (Barrick et al., 2001; Sutherland et al., 2007). Die Persönlichkeitsdimension unterscheidet ordentliche, zuverlässige, anspruchsvolle, pünktliche, disziplinierte sowie systematisch vorgehende Menschen von nachlässigen, gleichgültigen, undisziplinierten und inkompetenten Personen (Borkenau & Ostendorf, 1991, 1993, 2008). In anderen Persönlichkeitstheorien spielen neben der Planungsfähigkeit, Strukturiertheit und Zielstrebigkeit von Individuen weitere Konzepte wie beispielsweise die Impuls- oder Selbstkontrolle sowie die Fähigkeit, Belohnungen aufzuschieben, eine wichtige Rolle im Gewissenhaftigkeitskonzept (Corker, Oswald & Donnellan, 2012; DeYoung, Hirsh, Shane, Papademetris, Rajeevan und Gray, 2010; Lievens et al., 2009). Gewissenhaftigkeit kann somit definiert werden als individuelle Unterschiede in der Neigung, sozial vorgeschriebenen Normen zur Impulskontrolle zu folgen, zielorientiert und planvoll zu sein, Belohnungen aufzuschieben sowie soziale Normen und Regeln zu befolgen (Jackson et al., 2009; Jackson, Wood, Bogg, Walton, Harms & Roberts, 2010). Lord (2011) beschreibt Gewissenhaftigkeit als „Stärke der Zielsetzung und den Antrieb zur Zielerreichung“ (S. 25), d.h. die Art und Weise, wie Personen Aufgaben in Angriff nehmen, wie Menschen Dinge vollbringen und das Ausmaß, in dem sie ihre Energie auf die Erreichung konkreter Ergebnisse ausrichten. Die Autorin bezeichnet Gewissenhaftigkeit als Schlüsselmerkmal, wenn die Aufgabenerfüllung ein wichtiges Erfordernis ist. Gewissenhaftigkeit sei das Merkmal, das den Unterschied zwischen dem Leistungspotenzial und der tatsächlich gezeigten Leistung einer Person ausmache. Der Gewissenhaftigkeitsfaktor bezieht sich somit nicht nur darauf, Zeit zu investieren oder Aufgaben systematisch anzugehen und zu bearbeiten, sondern auch auf die strategische Ausrichtung auf bestimmte Ziele, den Willen zum Erfolg und den Glauben an die eigenen Fähigkeiten bzw. erfolgreich zu sein. Lord (2011) schließt also auch Selbstvertrauen als Komponente von Gewissenhaftigkeit – im Sinne des Wunsches oder Antriebs, erfolgreich zu sein und in Bezug auf die Art und Weise, wie eine Person mit den Faktoren umgeht, die erforderlich sind, um Ergebnisse hervorzubringen – mit in das Gewissenhaftigkeitskonzept ein. Gewissenhaftigkeit kann also auch als Wahrscheinlichkeit, die Ergebnisse zu erreichen, die mit den angestrebten Zielen übereinstimmen, bezeichnet werden. Die Autorin definiert Gewissenhaftigkeit in Bezug auf den Arbeitskontext als den effektiven Umgang mit Detailfragen und

das Befolgen von Regeln. Sehr gewissenhafte Menschen „besitzen einen Drang zur Vollendung von Aufgaben, obwohl sie sich möglicherweise zu lange mit Projekten aufhalten“ (S. 25). Darüber hinaus können sie eine Tendenz aufweisen, in so hohem Maße strukturiert zu sein, dass sie bei unerwarteten Ereignissen keine adäquaten Reaktionsmuster aufweisen und sie aus der Fassung geraten. Sehr gewissenhafte Menschen „sind eher loyale Angestellte, können aber unzufrieden werden, wenn Gelegenheiten für persönliche Leistung und Beförderung fehlen“ (S. 25).

DeYoung und seine Kollegen (2010) konnten in einer Studie Zusammenhängen der Big 5 Persönlichkeitseigenschaften mit Gehirnaktivitäten finden. Sie ließen 116 Personen den NEO-PI-R (Costa & McCrae, 1992) bearbeiten und untersuchten mittels Magnetresonanztomographie (MRT), welche Gehirnareale besonders aktiv waren. Die Ergebnisse der Forscher zeigten, dass Gewissenhaftigkeit positiv mit einem größeren Volumen des lateralen präfrontalen Cortex‘ zusammenhängt, was sich auch über den mittleren frontalen Gyrus ausweitet und vom frontalen bis zum hinteren Bereich des lateralen präfrontalen Cortex reichte. Diese Region ist unter anderem in die Informationsverarbeitung im Arbeitsgedächtnis sowie das Befolgen von Regeln involviert und für die Ausführung geplanter Handlungen zuständig. Die Ergebnisse sprechen somit dafür, dass eine hoch ausgeprägte Gewissenhaftigkeit mit effektiver Selbstregulation auf verschiedenen Komplexitätsebenen verbunden ist, was sich insbesondere auf die Einhaltung von Regeln und die Auswahl von Verhaltensweisen bezieht. Dies unterstreicht die Wichtigkeit der Gewissenhaftigkeitsdimension als Prädiktor für schulischen und beruflichen Erfolg sowie Gesundheit und Lebensdauer. Auch für weitere Persönlichkeitsmerkmale und Gehirnaktivitäten bzw. aktive Gehirnregionen konnten die Forscher Zusammenhänge finden. Die Ergebnisse unterstützen somit Annahme, dass den Big 5 biologische Systeme und Mechanismen zugrunde liegen.

Terracciano et al. (2008) führten eine längsschnittliche Untersuchung in einer repräsentativen US-Bevölkerungsstichprobe ($N=2.359$) durch. Über mehrere Jahre hinweg führten sie wiederholte Messungen u.a. auch mit Persönlichkeitsmaßen durch und erhoben beispielsweise Informationen über den Gesundheitszustand der befragten Personen. Sie fanden, dass eine hoch ausgeprägte Gewissenhaftigkeit mit einer geringeren Sterblichkeitsrate verbunden ist. Sie führten die höhere Lebenserwartung gewissenhafter Personen darauf zurück, dass sehr gewissenhafte Personen gesundheitsschädliche Verhaltensweisen (z.B. Rauchen, Einnahme von Drogen) eher vermeiden und eher gesundheitsförderliche Verhaltensweisen an den Tag legen (z.B. Sport treiben, Gewicht kontrollieren). Auch sind sie gesundheitsförderlichen Praktiken gegenüber aufmerksamer und halten eher medizinische Ratschläge ein (z.B. regelmäßige Medikamenteneinnahme).

Obwohl die Traits, zu denen ja auch Gewissenhaftigkeit zählt, als stabile, d.h. ein Leben lang konstante Eigenschaften einer Person, definiert sind, konnten sowohl querschnittliche als auch längsschnittliche Untersuchungen zeigen, dass Menschen mit zunehmendem Alter gewissenhafter werden, dass dieser Anstieg aber oftmals eher spät im Leben stattfindet (Jackson et al., 2009).

Wie bereits oben angeführt zeigen Metaanalysen übereinstimmend, dass Gewissenhaftigkeit und Arbeitsleistung positiv miteinander korrelieren und dass dieser Zusammenhang über Berufe und Hierarchieebenen hinweg generalisierbar ist (Barrick et al., 2001). Le, Oh, Robbins, Ilies, Holland und Westrick (2011) fanden zudem, dass Menschen mit höheren Ausprägungen in Gewissenhaftigkeit über ihre definierten Arbeitsaufgaben hinaus allgemein ein höheres Engagement in ihrer Firma zeigen, was sie als „Extra-Rollenverhalten“ definierten. Demzufolge tendieren gewissenhafte Menschen dazu, solches Verhalten an den Tag zu legen, das die sozialen bzw. psychologischen Rahmenbedingungen der Arbeit verbessert und weniger abweichendes Verhalten zeigen, das dem Erfolg des Unternehmens abträglich ist.

Doch nicht alle Untersuchungen konnten Zusammenhänge zwischen Gewissenhaftigkeit und Leistungskriterien finden. So konnten Sutherland et al. (2007) in einer Feldstudie mit 101 süd-afrikanischen Kundentechnikern im IT-Bereich verschiedenen Alters, Bildung und Arbeits-erfahrung keine Zusammenhänge mit Arbeitsleistung finden. Als Erklärungsmöglichkeit für diese Ergebnisse führten die Forscher die ohnehin vorherrschende starke Strukturierung an. In einer solchen Umgebung gebe es sehr viele starre Regeln und Normen, an die man sich bei der Arbeit halten müsse, sodass nur wenig Raum dafür bliebe, dass sich individuelle Unterschiede in der Gewissenhaftigkeit zeigen könnten bzw. im Arbeitsergebnis bemerkbar seien. Der IT-Kontext ist den Autoren zufolge also eher hinderlich für die Erforschung von Gewissenhaftigkeit. Es ist denkbar, dass sich hohe Ausprägungen in dieser Dimension vor allem in Situationen zeigen, in denen Menschen größere Handlungsspielräume und Wahlmöglichkeiten haben, was ihr Arbeitsverhalten betrifft. Beispielsweise wird eine sehr gewissenhafte Person ohne Vorgesetzten sich eher von allein gemäß ihrer eigenen Maßstäbe und dem individuellen Naturell an Regeln und Normen halten. Demgegenüber werden Mitarbeiter in einer Situation mit strikter Überwachung und wenig Wahlmöglichkeiten vergleichbare Arbeitsleistungen und insgesamt weniger Verhaltensvariationen zeigen – unabhängig davon, wie unterschiedlich stark ihre Ausprägungen im Merkmal Gewissenhaftigkeit sind. Sutherland et al. (2007) nahmen weiter an, dass sehr gewissenhafte Menschen hoch strukturierte Situationen und Menschen mit geringer Ausprägung in Gewissenhaftigkeit eher lockere Strukturen vorziehen. Sie vermuteten, dass hoch gewissenhafte Menschen auch in unstrukturierten Umgebungen erfolgreicher sind als wenig gewissenhafte, da diese aufgrund ihrer zuverlässigeren und ausdauernderen Arbeitsweise bessere Ergebnisse erzielen werden.

Ones und Viswesvaran (1997) führten verschiedene Gründe an, warum Gewissenhaftigkeit, schulischer Erfolg und Arbeitsleistung allgemein zusammenhängen bzw. warum die Persönlichkeitseigenschaft Berufs- und Schulerfolg vorhersagt:

1. Gewissenhafte Personen arbeiten länger an ihren Aufgaben. Dies betrifft sowohl den Lernstoff in der Schule bzw. Ausbildung als auch arbeitsrelevante Themen. Es führt zu einer höheren Produktivität bzw. einer verbesserten Lernleistung, was letztlich in besseren Produktivitätskennzahlen bzw. höheren Schulnoten resultiert.
2. Gewissenhafte Personen tun mehr als sie müssten. Zu dieser These passen die o.g. Ergebnisse von Le et al. (2011), dass gewissenhaftere Menschen mehr Extra-Rollenverhalten zeigen und sich über ihre Arbeitsaufgaben hinweg im und für ein Unternehmen engagieren. Im Schulkontext wären hierfür die Übernahme einer Sprecherfunktion oder die Leitung einer AG beispielhaft.
3. Gewissenhafte Personen setzen sich selbst eigenständig mehr Ziele und verfolgen diese ausdauernder. Dies kann man darauf zurückführen, dass diese Personen eher in der Lage sind, Belohnungen aufzuschieben und über eine höhere Zielorientierung und intrinsische Motivation verfügen.
4. Gewissenhafte Personen vermeiden kontraproduktives Verhalten. Sie sind eher in der Lage, sich zu strukturieren und fokussiert zu arbeiten. Dies hängt vermutlich mit ihrer erhöhten Zielorientierung zusammen. So weisen gewissenhaftere Menschen weniger Fehlzeiten auf und gehen trotz widriger Umstände zur Schule bzw. Arbeit (Lievens et al., 2009).

Insgesamt sind somit vor allem die leistungsorientierten Aspekte der Gewissenhaftigkeitsdimension aufgrund ihrer motivationalen Effekte in der Lage, Schulleistung vorherzusagen, da Schüler mit hohen Werten ausdauernder sowie fleißiger sind und allgemein mehr leisten wollen. Diese Personen tendieren dazu, Regeln und Normen einzuhalten und Belohnungen hinauszuzögern. Adjektive wie „kompetent“, „sorgfältig“, „gründlich“, „getrieben sein“ und „arbeitsorientiert“ beschreiben Individuen mit hohen Ausprägungen auf dieser Dimension (Lievens et al., 2009; Nofle & Robins, 2007).

Schon früh wurde der starke positive Einfluss von Gewissenhaftigkeit auf schulische und akademische Leistungsmerkmale nachgewiesen (z.B. Wiggins, Blackburn & Hackman, 1969; Wolfe & Johnson, 1995). Schmit, Ryan Stierwalt und Powell (1995) führten eine Studie zur Vorhersage von Schulnoten durch Gewissenhaftigkeit durch. Sie untersuchten 400 College-Studenten und erfassten neben der Gewissenhaftigkeitsdimension auch die zugehörigen Facetten der Dimension mit. Als Kriterium diente die Durchschnittsnote der Studenten (Grade Point Average; GPA). Die Untersuchungsgruppe wurde von den Forschern in zwei Subgruppen geteilt. Eine Gruppenhälfte bearbeitete den Selbstbeschreibungsbogen unter normalen Bedingungen (= „Studenten“), während der anderen Gruppenhälfte ein „Preisgeld“ für die „besten“ Selbstbeschreibungen geboten wurde (= „Bewerber“). Schmit et al. (1995) wollten auf diese Weise eine Bewerbungssituation simulieren und die Motivation der Studenten kontrollieren. Zudem erhielt wiederum eine Hälfte der Studenten einen schulbezogen formulierten Fragenbogen, während die andere Hälfte der Studenten eine neutrale, d.h. nicht kontextbezogene Version des Selbstbeschreibungsinventars bearbeitete. Es handelte sich also um ein klassisches 2x2-Design. Die Ergebnisse der Studie sind in Tabelle 2-5 aufgeführt.

Tabelle 2-5: Kriteriumsvaliditäten und interne Konsistenzen von Gewissenhaftigkeit bzw. deren Facetten (aus Schmit et al., 1995)

	allgemeine Items				schulbezogen formulierte Items			
	Studenten		Bewerber		Studenten		Bewerber	
	<i>r</i>	α	<i>r</i>	α	<i>r</i>	α	<i>r</i>	α
Gewissenhaftigkeit	.25*	.91	-.02	.90	.41**	.89	.46**	.94
C1 Kompetenz	.31**	.71	-.02	.68	.38**	.69	.53**	.84
C2 Ordnung	.02	.64	-.08	.55	.20	.51	.12	.56
C3 Pflichtbewusstsein	.19	.34	-.08	.42	.20	.24	.46**	.56
C4 Leistungsstreben	.25*	.81	-.10	.68	.38**	.82	.44**	.87
C5 Selbstdisziplin	.17	.79	.06	.79	.36**	.74	.38**	.86
C6 Besonnenheit	.23*	.77	.10	.72	.31**	.64	.35**	.76

Anmerkungen: * $p < .05$; ** $p < .01$. r = Korrelationskoeffizient; α = interne Konsistenz.

Es zeigte sich, dass die Gewissenhaftigkeitsdimension ein guter Prädiktor von Schulleistung bzw. der Durchschnittsnote der Studenten war ($r = .25$; $p < .05$). Die Facetten der allgemeinen Fragebogenversion korrelierten nur teilweise mit der Schulleistung (z.B. C1 Kompetenz $r = .31$, $p < .01$). Für die Probanden in der „Bewerber“-Bedingung ergab sich in Kombination mit den allgemein formulierten Items sowohl für die Gewissenhaftigkeitsdimension als auch die zugehörigen Facetten kein signifikanter Zusammenhang. Waren die Items hingegen schulbezogen formuliert, zeigten sich sowohl in der „Studenten-“ als auch in der „Bewerbergruppe“ deutliche Zusammenhänge mit der Gewissenhaftigkeitsdimension ($r = .41$ bzw. $r = .46$, $p < .01$) und den Facetten ($r = .31$ bis $r = .53$, $p < .01$). Lediglich die Facette Ordnung (C2) korrelierte unter keiner Bedingung mit dem Kriterium. Die höheren Interkorrelationen in der Bewerber-Bedingung sind mit Ergebnissen von Paulhus, Bruce und Trapnell (1995) sowie Dreier (2008) konsistent. Dieser Effekt könnte damit erklärt werden, dass Bewerber bzw. Probanden in Bewerber-Situationen klarere Konzepte (Schemata) über die sozialen Normen parat haben, auf die sie zur Beantwortung der Items zurückgreifen und sich somit konsistenter auf bestimmte Verhaltensweisen beziehen als in einer unspezifischen Situation. In Abschnitt 2.7 dieser Arbeit wird der Punkt des berufsbezogenen Testens noch einmal ausführlicher behandelt.

Schmit et al. (1995) folgerten aus den Ergebnissen weiter, dass mehr auch mit engeren Facetten der Dimensionen geforscht werden sollte. Vor allem die Facetten Ordnung (C2) und Pflichtbewusstsein (C3) waren weniger kriteriumsrelevant in Bezug auf das Kriterium GPA

im Vergleich zu den anderen Gewissenhaftigkeitsfacetten. Die Autoren interpretierten diesen Effekt allerdings nicht inhaltlich, sondern eher als Ergebnis der geringeren Reliabilitätskennwerte der Subskalen.

MacCann, Duckworth und Roberts (2009) identifizierten acht Subfaktoren von Gewissenhaftigkeit, die signifikant Schulerfolg bzw. Schulerfolgskriterien vorhersagten: *Industriousness* (Fleiß), *Perfectionism* (Perfektionismus), *Tidiness* (Ordnung), *Procrastination* (Aufschubverhalten), *Control* (Kontrolle), *Cautiousness* (Vorsicht), *Task Planning* (Aufgabenplanung) und *Perserverance* (Ausdauer). Den Ergebnissen zufolge war Fleiß die Facette, die die höchsten prädiktiven Validitäten aufwies. So konnte sie sowohl Lehrerbeurteilungen ($r = .29, p < .01$) als auch GPA ($r = .20, p < .01$) signifikant vorhersagen. Auch in dieser Untersuchung war Ordnung mit den meisten akademischen Kriterien nicht korreliert, während Fleiß mit diesen recht stark zusammenhing. Im Vergleich mit der allgemeinen Gewissenhaftigkeitsdimension war Fleiß ein besserer Prädiktor von Abwesenheitszeiten. Auch MacCann et al. (2009) kamen zu dem Fazit, dass die Verwendung von Facetten zur Vorhersage von z.B. Schulleistung besser geeignet ist als Gewissenhaftigkeit per se. Vor allem, wenn man konkrete (Schul-) Leistungs- und Erfolgsvariablen wie z.B. Fehlzeiten vorhersagen möchte, scheinen Facetten bzw. Subfaktoren bessere Prädiktoren zu sein.

Bereits Rothstein, Paunonen, Rush und King (1994) untersuchten, ob die Big 5-Faktoren und ihre entsprechenden Facetten bessere Prädiktoren schulischer Leistung sind. Keiner der Faktoren der Big 5 korrelierte mit dem Kriterium (GPA). Nur zwei Facetten zeigten Zusammenhänge mit GPA. Die Gewissenhaftigkeitsfacette *Achievement* (Leistungsorientierung) und eine Verträglichkeitsfacette zeigten Zusammenhänge mit Schulleistung ($r = .21$). Chamorro-Premuzic und Furnham (2003) nutzten einen ähnlichen Ansatz und verglichen die allgemeinen NEO-PI-R-Dimensionen mit den zugehörigen Facetten in Bezug auf ihre Vorhersagekraft von Schulleistung (GPA). Sie fanden, dass GPA signifikant mit der Gewissenhaftigkeitsdimension ($r = .36$) sowie den zugehörigen Facetten *Dutifulness* ($r = .38$), *Achievement-Striving* ($r = .35$) sowie *Self-discipline* ($r = .22$) korrelierte.

Paunonen und Ashton (2001) untersuchten die relative prädiktive Nützlichkeit von Facetten gegenüber Faktoren. Sie verglichen jeweils nur einen Faktor mit jeweils einer Facette, z.B. den Faktor Gewissenhaftigkeit mit der Facette *Achievement*. In Bezug auf die Vorhersagekraft des Kriteriums (Abschlussnote) erwies sich die Facette ($r = .26$) als besserer Prädiktor als die Dimension ($r = .21$).

In einer Untersuchung von Ziegler, Danay, Schölmerich und Bühner (2010) zur Vorhersagekraft der Gewissenhaftigkeitsfacetten zeigte *Ordnung* einen negativen Zusammenhang mit Leistung. Dies kann dadurch erklärt werden, dass sehr ordentliche Personen sich eher an bestehenden Strukturen orientieren und weniger an Veränderungen ihrer Arbeitsweisen interessiert sind, was jedoch gerade in Bezug auf das Erreichen guter Noten im Abschlusszeugnis evtl. notwendig ist. Wenn man seine Lerngewohnheiten somit nicht verändert, könnte dies zu minderen Leistungen führen, z.B. auch zu einer verminderten Ausbildungsleistung im Vergleich zur Schulleistung vorher. Wenn man im Gegensatz dazu eher aktiv ist und beispielsweise Dinge tut, die zur Zielfindung beitragen, kann dies zu positiven Leistungen führen. Somit hängt die Facette *Fleiß* positiv mit Leistung zusammen. Diese gegensätzlichen Effekte der Gewissenhaftigkeitsaspekte könnten in der Vergangenheit dazu geführt haben, dass die prädiktive Validität der Gewissenhaftigkeitsdimension nicht korrekt erfasst werden konnte. Zukünftige Studien sollten sich daher eher auf die Spezifität von Kriterium und Prädiktoren konzentrieren.

O'Connor und Paunonen (2007) führten eine Metaanalyse zur prädiktiven Validität der Persönlichkeitsdimensionen der Big 5 im Vergleich mit den jeweils zugehörigen Facetten in Bezug auf schulische Leistungskriterien durch. Sie kritisieren, dass allgemein bisher v.a. nur

korrelative Zusammenhänge berichtet wurden, die keine Aussage über Ursache und Wirkung erlauben. Die Vorhersagekraft und auch inkrementelle Validität der Facetten wurde den Autoren zufolge bisher kaum bzw. nur unzureichend beforscht. Lediglich Paunonen (1998) untersuchte die inkrementelle Validität von Facetten über Dimensionen hinaus mit Bezug auf schulische Leistung (GPA) als Kriterium. Den Ergebnissen zufolge erklärten die Dimensionen 6 % der Varianz der Schulleistung. Die Facette *Achievement* (Leistung) erklärte zusätzliche 7,2 % der Varianz auf. In einer zweiten Studie von Paunonen (1998) waren die Dimensionen nicht in der Lage GPA vorherzusagen, aber die Facette *Responsibility* (Verantwortung) klärte 5,8 % der Varianz auf. Paunonen und Ashton (2001) fanden in einer anderen Untersuchung heraus, dass fünf ausgewählte Facetten des NEO-PI-R zusätzliche 6,3 % an Varianz hinsichtlich des Kriteriums GPA aufklären konnten. Laut O'Connor und Paunonen (2007) sind Facetten somit in der Lage, ca. 5-7 % zusätzliche Varianz über allgemeine Dimensionen der Persönlichkeit hinaus aufzuklären. Die Autoren identifizierten Gewissenhaftigkeit ($r = .24$) als stärksten Prädiktor von Schulleistung, neben Extraversion ($r = -.05$) und Offenheit ($r = .06$). Die Ergebnisse der Metaanalyse demonstrieren weiter, dass – zumindest in einigen Situationen – engere Persönlichkeitsfacetten Vorteile gegenüber den allgemeinen Dimensionen haben. Die Autoren kritisierten an der bisherigen Forschung allerdings, dass in der Regel immer wenige Dimensionen mit vielen Facetten verglichen würden, sodass allein auf Basis der Anzahl der untersuchten Facetten eine zumindest statistisch größere Wahrscheinlichkeit vorliegt, dass eine der Facetten auch mit dem Kriterium korreliert.

Seit Veröffentlichung der Metaanalyse liegen einige neuere Untersuchungen zur inkrementellen Validität der Facetten über die allgemeinen Dimensionen hinaus vor. So verglichen O'Neill und Allen (2007) in einer Stichprobe von 69 studentischen Teams, ob engere und spezifischere Facetten genauso viel oder sogar mehr Varianz menschlichen Verhaltens in Bezug auf Teamleistung aufklären können wie die allgemeinen, groben Dimensionen. Das Ziel der Studie war es, den relativen Anteil spezifischer Traits im Vergleich zu den Dimensionen der Big 5 zu erforschen. Das Kriterium Teamleistung wurde durch die Abschlussnote am Ende eines Ingenieurkurses operationalisiert. Mit Hilfe von Regressionen wurde die inkrementelle Validität der spezifischen Facetten über die Big 5-Persönlichkeitsdimensionen hinaus berechnet. Die Ergebnisse der Untersuchung zeigten, dass spezifischere Facetten über die Big 5 hinaus Varianz aufklärten und somit inkrementelle Validität besaßen. Das bedeutet, dass die Facetten im Vergleich mit den Dimensionen der Big 5 für die Vorhersage von Leistung genauso gut oder sogar besser geeignet sind. Zudem sind sie den Autoren zufolge einfacher zu interpretieren, da ihre Zusammenhänge mit beispielsweise Leistungsvariablen stärker sind. Die allgemeine Gewissenhaftigkeitsdimension verschleierte den Ergebnissen zufolge sogar bedeutungsvolle Zusammenhänge zwischen den Facetten und Teamleistung.

O'Neill und Allen (2007) empfehlen für die Praxis darüber hinaus, die Facetten konstruktorientiert auszuwählen. Das heißt, dass die Prädiktorvariablen theoretisch und konzeptuell mit den Kriteriumsvariablen verbunden sein sollten. Nur so könne man bessere Validitäten erhalten.

Zu vergleichbaren Ergebnissen, dass Facetten bessere Hinweise gaben, welche Aspekte von Gewissenhaftigkeit mit Jobcharakteristika korrelieren, kamen auch Sutin und Costa (2010). Ihren Ergebnissen zufolge korrelierten *Kompetenz* (C1) und *Leistungsstreben* (C4) hoch positiv mit Entscheidungsfreiheit. Nicht alle Korrelationen auf Facettenebene wiesen die gleichen Vorzeichen auf wie die Korrelationen der Traits, sodass man auch hier von einer differenzierteren Vorhersagekraft seitens der Facetten sprechen kann. Die Facetten C1, C3, C4, C5 und C6 korrelierten hoch mit Entscheidungsverhalten im Job, was darauf hinweist, dass diese Eigenschaften höhere Eigeninitiative mit sich bringen. Die Autoren folgerten aus den Ergebnissen, dass gewissenhaftere Menschen dazu tendieren, eher in sehr anspruchsvollen Jobs tätig zu sein, die ihnen dennoch die Freiheit lassen, eigene Entscheidungen zu treffen und sich

die Arbeit selbst einzuteilen. Persönlichkeit spielte in dieser Studie eine wichtige Rolle, um das Umfeld einer Person zu bestimmen. Eine hohe Ausprägung in der Gewissenhaftigkeitsdimension war assoziiert mit einem Anstieg in der Entscheidungsfreiheit, d.h. sehr gewissenhafte Personen gestalteten ihren Arbeitsplatz über die Zeit hinweg aktiv, um mehr Entscheidungsspielräume zu erschaffen und um ihrer Kreativität Ausdruck verleihen zu können. Diese Zusammenhänge waren berufsübergreifend zu beobachten und waren auch unabhängig vom beruflichen Prestige sowie Berufswechseln.

Es bleibt somit festzuhalten, dass im Hinblick auf manche Kriterien einzelne Persönlichkeitsfacetten eine größere Vorhersagekraft haben können als umfassende Bereichswerte oder Kombinationen umfassender Bereichswerte. Entsprechend kann man die Ergebnisse von Mount und Barrick (1995) interpretieren, die zeigten, dass Leistungsstreben als Facette von Gewissenhaftigkeit die investierte Anstrengung besser vorhersagte ($r = .58$) als der Gesamtwert der Dimension Gewissenhaftigkeit ($r = .51$). Paunonen und Ashton (2001) verglichen die Faktoren der Big 5 mit den Facetten des FFM im Hinblick auf ihre relative Fähigkeit, spezifische Verhaltensweisen vorherzusagen. Sie stellten fest, dass die Berücksichtigung der enger umschriebenen Facetten zu einer wesentlichen Erhöhung der Vorhersagegenauigkeit führte.

Schon lange gibt es in der Wissenschaft eine Diskussion über die Nützlichkeit allgemeiner vs. differenzierter Konstrukte in Bezug auf die (verbesserte) Vorhersage von Leistungs- und Erfolgskriterien (Ones & Viswesvaran, 1996). Ein Hauptargument für die Verwendung der Subfaktoren ist, dass der Fokus auf engeren Aspekten eine bessere Einsicht in die Zusammenhänge der allgemeinen Konzepte wie Gewissenhaftigkeit mit Leistung geben kann (Corker et al., 2012; Hülshager & Maier, 2008).

Lord (2011) widmet sich dieser Problematik für den praktischen Anwender und erläutert diese anschaulich anhand des NEO-PI-R für die Praxis der Personalauswahl und -entwicklung. Ihr zufolge spricht ein hoher Gesamtwert auf einer Dimension nicht gleichzeitig dafür, dass auch alle ihr zugeordneten Facetten in derselben Richtung oder ebenfalls hoch ausgeprägt sein müssen. Wenn der Hauptskalenwert außerhalb des mittleren Bereichs liegt – egal ob hoch oder niedrig – kann es immer ein oder zwei Facetten geben, die innerhalb des durchschnittlichen Bereichs liegen oder sich – relativ zur Ausprägung der jeweiligen Dimension – am entgegengesetzten Pol der Skala befinden. Ebenso ist eine durchschnittliche Ausprägung auf einem Faktor nicht immer auch das Ergebnis durchschnittlicher Ausprägungen auf allen zugehörigen Facetten. Eine solche kann auch erzielt werden, wenn eine Person hohe Merkmalsausprägungen auf einigen Facetten und niedrige Ausprägungen auf anderen hat.

Es ist somit wichtig zu beachten, wie die Zusammenhänge der Kriteriumskorrelationen einer Persönlichkeitsdimension von den einzelnen zugehörigen Facetten gesteuert werden. Die Ausprägung der allgemeinen Dimension gibt demzufolge Aufschluss über allgemeine Verhaltensweisen und -tendenzen einer Person. Wenn das Merkmal Gewissenhaftigkeit hoch ausgeprägt ist, kann Lord (2011) zufolge entsprechend angenommen werden, dass „eine Person einen Sinn für Zielsetzungen hat und bestimmte Ergebnisse erreichen wird“ (S. 26). Wenn allerdings bestimmte Facetten der Richtung dieses Gesamtwertes entgegen gesetzt sind, sollte dies ebenfalls beachtet werden. Wenn man beispielsweise eine Person mit einer eher niedrigen *Ordnungsliebe* (C2) und ansonsten hohen Werten und eine Person mit einem geringen *Leistungsstreben* (C4) und einer sonst hohen Ausprägung im Merkmal Gewissenhaftigkeit vergleicht, so ist die erste Person zwar schlecht organisiert, wird aber durch persönliche Leistungen motiviert. Die zweite Person ist zwar gut organisiert und methodisch, wird aber nicht von Ehrgeiz oder einem hohen Niveau persönlicher Leistungen getrieben. Innerhalb jedes Persönlichkeitsbereichs variieren die Facetten zudem in der Stärke ihres Einflusses auf den Bereich.

Die Facetten von Gewissenhaftigkeit fasst Lord (2011) zu drei Eigenschaften zusammen. Die Facette *Leistungsstreben* (C4) als Ausdruck einer Person Leistung zu erbringen sowie die Facette *Kompetenz* (C1) als Überzeugung, dass die eigene Zielerreichung von persönlicher Anstrengung und vom eigenen Können abhängt anstatt von Zufall oder äußeren Faktoren, sollten der Autorin zufolge eigenständig bleiben. Die Facetten *Ordnungsliebe* (C2), *Pflichtbewusstsein* (C3), *Selbstdisziplin* (C5) und *Besonnenheit* (C6) fasst sie zu einer dritten Eigenschaft zusammen, die sie als die „Fähigkeit sich auf gegenwärtige Ziele zu konzentrieren ohne abgelenkt zu werden“ beschreibt (S. 45). Wenn auch *Pflichtbewusstsein* (C3) als Gefühl moralischer Verpflichtungen gegenüber denen, für die man arbeitet, nur bedingt in diese Kategorisierung passt. Die Autorin vergleicht auch die Ladungen der einzelnen Facetten des NEO-PI-R auf die Gewissenhaftigkeitsdimension. Tabelle 2-6 gibt einen Überblick über die aus verschiedenen Stichproben zusammengetragenen Ergebnisse, mit welchem Gewicht die Gewissenhaftigkeitsfacetten zur Bedeutung der Gewissenhaftigkeitsdimension beitragen.

Tabelle 2-6: Ladungen der Gewissenhaftigkeitsfacetten auf die Gewissenhaftigkeitsdimension (nach Lord, 2011)

Facette	Ladung auf Gewissenhaftigkeit	Art des Einflusses
C1 Kompetenz	0,63 (0,69) [0,64]	Selbstvertrauen und Gefühle des Könnens
C2 Ordnungsliebe	0,75 (0,75) [0,70]	Persönliche Organisiertheit
C3 Pflichtbewusstsein	0,75 (0,75) [0,68]	Festhalten an ethischen Prinzipien und einem Moral-kodex
C4 Leistungsstreben	0,76 (0,71) [0,74]	Bedürfnis nach persönlicher Leistung
C5 Selbstdisziplin	0,78 (0,78) [0,75]	Fähigkeit, mit Aufgaben anzufangen und sie trotz Ablenkungen zum Abschluss zu bringen
C6 Besonnenheit	0,54 (0,59) [0,57]	Tendenz, die Dinge vor dem Handeln vollständig zu durchdenken

Anmerkungen: Werte ohne Klammern entsprechen den Ergebnissen aus verschiedenen deutschsprachigen Ländern; Werte in runden Klammern entsprechen Ergebnissen einer Studie mit Erwerbstätigen in Großbritannien; Werte in eckigen Klammern entsprechen den Ladungen auf Gewissenhaftigkeit in der Allgemeinbevölkerung der USA.

Den Ergebnissen der verschiedenen Studien zufolge zeigen alle Facetten sprach- und kulturübergreifend ähnlich hohe Ladungen auf die Gewissenhaftigkeitsdimension. Nur *Besonnenheit* (C6) zeigt einen etwas geringeren Einfluss. Diese Facette ist aber dennoch wichtig, da sie das Ausmaß beschreibt, wie sehr Menschen die Folgen ihrer Handlungen in Bezug auf das Erreichen der von ihnen angestrebten Ziele durchdenken. *Leistungsstreben* (C4) ist diesen Befunden zufolge beispielsweise eng mit Gewissenhaftigkeit verbunden. Menschen mit hohen Ausprägungen auf dieser Facette zeigen eine Tendenz, bei der Ausführung von Aufgaben zielstrebig und effizient zu sein. Im Schul- und Arbeitskontext bedeutet dies, dass gewissenhafte Schüler und Angestellte ihre Aufgaben effektiver erledigen und letztlich eine größere Menge an Arbeit erledigen. Gleiches lässt sich auf die anderen Facetten übertragen.

Allgemein zeigen die Gewissenhaftigkeitsfacetten des NEO-PI-R moderate bis hohe Zusammenhänge. Tabelle 2-7 gibt entsprechend einen Überblick über die Interkorrelationen der Gewissenhaftigkeitsfacetten (Rust, 2005, zitiert aus Lord, 2011).

Tabelle 2-7: Interkorrelation der Gewissenhaftigkeitsfacetten (nach Lord, 2011)

	C1	C2	C3	C4	C5	C6
C1 Kompetenz	1.00	.36	.54	.55	.68	.40
C2 Ordnungsliebe		1.00	.43	.42	.51	.40
C3 Pflichtbewusstsein			1.00	.49	.60	.45
C4 Leistungsstreben				1.00	.65	.30
C5 Selbstdisziplin					1.00	.40
C6 Besonnenheit						1.00

Barrick und Mount (2004) versuchten, in der Diskussion um die Faktorenstruktur und Vorhersagekraft von Gewissenhaftigkeit zu einem Kompromiss zu gelangen. Sie postulierten, dass breite, allgemeine Dimensionen, wie z.B. die Big 5, gut für die Verwendung in der Entwicklung von Theorien und für allgemeine Erklärungsansätze geeignet seien, während engere Facetten von Gewissenhaftigkeit, wie z.B. Fleiß und Ordnung, eher für die Vorhersage spezifischer Verhaltensweisen bei der Arbeit verwendet werden sollten, denn die Facetten beinhalten konkretere Beschreibungen von Situationen, Zeiten und Rollen. Durch diese würde sich somit eher die prädiktive Validität für bestimmte Verhaltensweisen erhöhen. Demgegenüber fokussierten allgemeine Dimensionen eher auf allgemeine, übergreifende Verhaltenstendenzen, was die Vorteile höherer Effektivität (in der Messung), größerer Bandbreite und höherer situationsübergreifender Replizierbarkeit mit sich bringt. Dies erklärt, warum diese Faktoren besser allgemeine Kriterien wie z.B. subjektiv empfundenen Erfolg, aber weniger gut spezifischere Kriterien vorhersagen (Ones & Viswesvaran, 1996; Schneider, Hough & Dunnette, 1996). Die Autoren forderten eine hierarchische Organisation der verschiedenen Traitebenen, um die allgemeinen Dimensionen und spezifischen Subskalen zu organisieren. Dieses haben DeYoung und Kollegen (2010) für die Big 5 vorgelegt (s. Abschnitt 2.3 Abbildung 2-2).

Dudley et al. (2006) führten eine Metaanalyse zur prädiktiven Validität der Gewissenhaftigkeitsfacetten durch. Ihnen zufolge repräsentiert eine Lösung mit vier Facetten am besten den aktuellen Forschungsstand. Die vier Facetten von Gewissenhaftigkeit sind ihnen zufolge *Achievement* (Leistung), was eine Tendenz reflektiert, Standards anzuerkennen bzw. auszuführen, um Ziele zu erreichen. Die zweite Facette benennen Dudley et al. (2007) mit *Orderliness* (Ordnung). Sie beschreibt gut organisierte, planend handelnde, detailorientierte und vorsichtig bzw. methodisch vorgehende Personen. Die dritte Facette ist mit *Cautiousness* (Vorsicht) benannt und reflektiert eine Tendenz, Risiken vorab zu bedenken und ist als Gegenspieler zu Impulsivität anzusehen. Die vierte Facette nennen die Autoren *Dependability* (Abhängigkeit). Diese Facette beschreibt vertrauenswürdige, berechenbare, disziplinierte Personen, die Regeln und Gesetze beachten. Die Aspekte sind gering bis moderat miteinander interkorreliert ($r = .14$ bis $.60$), weshalb es den Autoren zufolge Sinn macht, inhaltlich zwischen diesen zu unterscheiden. Die Facetten sind moderat bis hoch mit Gewissenhaftigkeit korreliert ($r = .38$ bis $.73$). Dependability hat die höchsten Korrelationen mit der Dimension, was darauf hinweist, dass dies einer der Kernaspekte von Gewissenhaftigkeit ist.

Zu ähnlichen Ergebnissen kamen Egberink, Meijer und Veldkamp (2010). Sie fanden hohe Korrelationen der Gewissenhaftigkeitsdimension mit den NEO-PI-R-Facetten *Performance*, *Organization*, *Concentration* sowie *Methodicalness* ($r = .75$ bis $.85$), während die Facetten untereinander eher moderate Interkorrelationen aufwiesen ($r = .41$ bis $.66$).

Diese Befunde deuten auch an, dass das Konzept der Gewissenhaftigkeit breiter bzw. unterschiedlicher als das Aggregat der Aspekte ist. Die Facetten bzw. deren prädiktive Validität waren Dudley et al. (2006) zufolge nicht geringer als die berichteten Werte, die mit globalen Gewissenhaftigkeitsmaßen gefunden wurden. V.a. Dependability und Achievement waren

aussagekräftige Prädiktoren für aufgabenbezogenes oder kontextuelles Verhalten. Für allgemeines Leistungsverhalten waren die Facetten jedoch nicht besser zur Vorhersage geeignet als Gewissenhaftigkeit allgemein, was der Annahme entspricht, dass Item- und Kriteriumsspezifität übereinstimmen sollten. Bei der Vorhersage spezifischer Leistungskriterien hingegen erwiesen sich die Facetten als aussagekräftigere und präzisere Prädiktoren, die über Gewissenhaftigkeit hinaus auch inkrementelle Validität in Bezug auf die Vorhersage von Extrarollenverhalten (wie z.B. Hingabe zum Job, kontraproduktives Verhalten oder interpersonelle Hilfe) besitzen. Inkrementelle Validität besitzen die Facetten auch bei der Vorhersage von Leistungsverhalten in verschiedenen Berufsgruppen. Hierzu sei angemerkt, dass die Facetten allerdings nur dann inkrementelle Validität besaßen, wenn zwischen den untersuchten Berufsgruppen unterschieden wurde. Bei einer gemeinsamen Betrachtung der verschiedenen Stichproben traten in der Metaanalyse keine Effekte auf. Daher sollte man Dudley et al. (2006) zufolge die Facetten insbesondere dann einsetzen, wenn es darum geht, spezifisches Verhalten vorherzusagen.

Allerdings herrscht nach wie vor Uneinigkeit darüber, wie der Faktor Gewissenhaftigkeit nun aufzusplitten ist. MacCann et al. (2009) identifizierten wie bereits oben beschrieben acht Subfaktoren von Gewissenhaftigkeit. Costa und McCrae (1992) ordneten jeder Dimension des FFM jeweils sechs Facetten zu, so auch der Gewissenhaftigkeitsdimension (s. Abschnitt 2.3). Jackson et al. (2010) identifizierten fünf der Gewissenhaftigkeitsdimension untergeordnete Faktoren (*Industriousness*, *Orderliness*, *Impulse control*, *Reliability* und *Conventionality*). Ihren Ergebnissen zufolge repräsentieren verhaltensbezogene Faktoren Mischungen aus verschiedenen Facetten allerdings zwei allgemeine Tendenzen. So konnten Facetten wie *Impulse control* und *Reliability* zu einem „antisozialen Verhaltensfaktor“ zusammengefasst werden sowie Facetten wie *Orderliness* und *Conventionality* zu einem Aspekt „Sauberkeit“ aggregiert werden.

Wiederum andere Forschungsarbeiten fanden eine Vier-Faktoren-Lösung für die Substruktur von Gewissenhaftigkeit, deren Facetten mit *Responsibility* (Verantwortung), *Order* (Ordnung), *Impulse control* (Impulskontrolle) und *Laziness* (Faulheit) benannt wurden (Peabody & De Raad, 2002; Perugini & Gallucci, 1997; Roberts et al., 2005; Saucier & Ostendorf, 1999).

Unabhängige Forschungsarbeiten berichten zunehmend und übereinstimmend von einer zweifaktoriellen Substruktur bzw. sinnvollen Zweiteilung der Dimension (Andresen, eingereicht; DeYoung et al., 2007; DeYoung, 2010; Hough & Schneider, 1996; Jackson et al., 2009). Um herauszufinden, wie Gewissenhaftigkeit strukturiert ist, kann man entweder die lexikalische Struktur der Sprache oder die Faktorenstruktur bestehender Persönlichkeitsfragebögen untersuchen. Dies haben beispielsweise DeYoung et al. (2007) und Andresen (eingereicht) getan. Andresen analysierte 13 bestehende Persönlichkeitsinventare bezüglich ihrer Faktorenstruktur. Er fand eine 8-Faktoren-Lösung der Persönlichkeit, in der unter anderem der eher dominant-aggressive *W-Faktor* enthalten ist, der für den Willen, etwas erreichen zu wollen und dafür hart zu arbeiten, steht. Als Gegenspieler hierzu fand Andresen den eher submissiven, passiv-protektiven *C-Faktor*, der für „Gewissenhaftigkeit, Ordnungsliebe und Impulskontrolle“ steht.

DeYoung et al. (2007) untersuchten ebenfalls faktorenanalytisch die 75 Facetten von zwei Selbstbeschreibungsinventaren, die auf dem FFM beruhten. Die Probanden kamen aus einer bevölkerungsrepräsentativen Feldstichprobe ($N = 481$), die die beiden Persönlichkeitsmaße bearbeiteten. Die Autoren konnten aus den Daten für jede Dimension eine zweifaktorielle Substruktur extrahieren. Da sich diese hierarchisch zwischen den Facetten und Dimensionen befindet, benannten die Autoren die von ihnen gefundenen Subfaktoren als *Aspekte*. Die Aspekte erwiesen sich als unterscheidbar, aber dennoch korrelierte Konstrukte. Gewissenhaftigkeit wird den Ergebnissen zufolge durch die beiden Aspekte *Industriousness* (Fleiß) sowie *Orderliness* (Ordnung) gebildet. Anschließend entwickelten DeYoung und Kollegen (2007)

einen Fragebogen, der die Aspekte der fünf Persönlichkeitsdimensionen des FFM erfasst – die Big Five Aspect Scales (BFAS) (s. Abschnitt 2.6.3). Die BFAS erfassen jeden Big 5-Aspekt mit jeweils zehn Items, sodass sich ein 100-Item-Fragebogen ergab. Dieser wurde wiederum anhand einer bevölkerungsrepräsentativen Stichprobe ($N=480$) evaluiert. Die Ergebnisse deuteten auf gute Testkennwerte hin. Auch Hinweise hinsichtlich einer biologischen Verankerung der Aspekte deuteten sich an.

Auch in älteren Forschungsarbeiten lassen sich Hinweise auf eine Zweiteilung des Faktors Gewissenhaftigkeit finden (z.B. Edwards, Abbott & Klockars, 1972; Huba, Segal & Singer, 1977; Nesselroade & Baltes, 1975). Tellegen (1982) postulierte unter seinen drei Hauptfaktoren der Persönlichkeit ein 11-Faktoren-Modell, in das er unter anderem die beiden Persönlichkeitsfaktoren *Control* und *Achievement* aufnahm, die nachweislich mit Gewissenhaftigkeit zusammenhängen (Ackerman & Heggstad, 1997). Jackson, Ashton und Tomes (1996) bzw. Jackson et al. (1996) definierten ein Sechs-Faktoren-Modell der Persönlichkeit, indem sie den Gewissenhaftigkeitsfaktor in einen Ordnungs- und einen Fleißfaktor aufspalteten und diese als *Methodicalness* und *Industriousness* in das Fünf-Faktoren-Modell als jeweils eigenständige Dimensionen aufnahmen.

Hough und Schneider (1996) unterteilten den Faktor Gewissenhaftigkeit in zwei Subskalen: *Achievement* (Leistung) und *Dependability* (Abhängigkeit). Der Subskala *Achievement* ordneten sie die NEO-PI-R-Facetten *Leistungsstreben* (C4) und *Kompetenz* (C1) zu. Der Subskala *Dependability* ordneten sie die anderen vier Facetten *Ordnung* (C2), *Pflichtbewusstsein* (C3), *Selbstdisziplin* (C5) und *Besonnenheit* (C6) zu. Sie fanden, dass sich bei einer Kontextualisierung der Items die Validität sich sowohl bei *Achievement* als auch bei *Dependability* erhöhte. Diese Befunde sind inhaltlich nachvollziehbar, da beide Verhaltenstendenzen allgemeine Voraussetzung für (Schul-) Erfolg sind.

Neben Hough und Schneider (1996) unterteilten auch Mount und Barrick (1995), Gewissenhaftigkeit in die zwei Subkomponenten *Achievement* und *Dependability*. *Achievement* beschreibt den Autoren zufolge die Fähigkeit, hart zu arbeiten und Ziele zu erreichen, während *Dependability* eher eine Tendenz zu Verantwortungs- und Pflichtbewusstsein beschreibt. Allgemein ist eine Tendenz dahingehend zu erkennen, dass die beiden Aspekte von Gewissenhaftigkeit zum einen immer eher aktive Eigenschaften wie Ehrgeiz, Leistungsorientierung und Ausdauer vereinen und zum anderen eher passives Verhalten wie planvolles Vorgehen, Organisation und ein allgemeines Ordnungsbedürfnis beschreiben (Jackson et al., 1996). Zunehmend haben sich für diese beiden Subkomponenten von Gewissenhaftigkeit die Bezeichnungen *Industriousness* bzw. *Fleiß* sowie *Orderliness* und *Order* bzw. *Ordnung* durchgesetzt.

Unabhängig davon, wie viele Subfaktoren von Gewissenhaftigkeit identifiziert werden, bleibt festzuhalten, dass die Persönlichkeitseigenschaft Gewissenhaftigkeit hierarchisch organisiert ist. Da die bisher identifizierten Gewissenhaftigkeitsfacetten alle auf Fragebogenstudien und Einschätzungen von Adjektivlisten beruhen, sind die Spannweite und die Anzahl der gefundenen Facetten möglicherweise begrenzt (Jackson et al., 2010). Die zwei-Faktoren-Lösung bestehend aus einem proaktiven und einem inhibitorischen Faktor, die die verschiedenen Verhaltensweisen von Gewissenhaftigkeit repräsentieren, scheint derzeit eine praktikable und effektive Lösung zu sein, die aktuellen Ergebnisse zu integrieren. Differenziertere Facetten-Lösungen splitten die Aspekte mehr und mehr auf. So differenziert eine Drei-Faktoren-Lösung beispielsweise den Impulsivitäts-Faktor in *Zuverlässigkeit* und *Impulskontrolle*.

Lievens, De Corte und Schoellaert (2008) untersuchten den Zusammenhang der Aspekte von Gewissenhaftigkeit – *Fleiß* und *Ordnung* – mit der Gewissenhaftigkeitsdimension. Gewissenhaftigkeit und Fleiß zeigten einen mittleren Zusammenhang ($r=.58$), während Ordnung und Gewissenhaftigkeit eine hohe Korrelation aufwiesen ($r=.70$). Die beiden Aspekte zeigten

eine hohe Interkorrelation ($r=.73$). Diese Ergebnisse spiegeln auch die bereits oben beschriebenen Befunde der Metaanalyse von Dudley et al. (2006) wider.

Zusammenfassend bleibt festzuhalten, dass präzisere Facetten bzw. Aspekte von Gewissenhaftigkeit besser geeignet sind, Vorhersagen von Verhalten und Leistung in der Schule und bei der Arbeit zu treffen als allgemeine Gewissenhaftigkeitsmaße (Dudley et al., 2006; O'Connor & Paunonen, 2007). Daher kann eine Ergänzung der bestehenden berufseignungsdiagnostischen Testverfahren und -prozesse um präzisere Gewissenhaftigkeitsmaße im Sinne einer verbesserten prädiktiven Validität und des inkrementellen Validitätszuwachses gewinnbringend sein. Aber auch weitere Persönlichkeitseigenschaften, die nicht aus der traittheoretischen Forschungstradition stammen, haben sich als Prädiktoren von Ausbildungs- und Berufserfolg erwiesen. Diese sollen in den folgenden Abschnitten (2.3.2 ff) näher erläutert werden.

2.3.2 Selbstwirksamkeit

Bandura (1977) führte das Konzept der Selbstwirksamkeit als erster ein und definierte Selbstwirksamkeit als „den Glauben an die eigenen Fähigkeiten, eine Aufgabe oder ein bestimmtes Verhalten erfolgreich zu bewältigen oder ausführen zu können“ (S. 79). Das Konstrukt der Selbstwirksamkeit umschreibt Erwartungen eigener Effektivität und macht Vorhersagen darüber, ob, auf welche Art und mit wie viel Mühe sich jemand der Bewältigung einer Aufgabe widmet, auch wenn er bei der Bearbeitung dieser mit Widerständen rechnen muss. Es umfasst auch die Ausdauer, mit der jemand eine Tätigkeit, die subjektiv als unangenehm erlebt wird, weiter bzw. zu Ende führt (Bandura & Wood, 1989).

Das Selbstwirksamkeitskonstrukt kann auch als erlernte Kompetenzerwartung bezeichnet werden (Abele et al., 2000) und ist Bestandteil der sozial-kognitiven Lerntheorie (SKL; Bandura, 1977), die annimmt, dass wir aktive Gestalter unserer Umwelt sind. Selbstwirksamkeit beginnt Bandura (1977) zufolge, sich in der Kindheit zu entwickeln. Die frühzeitige Entwicklung von Selbstwirksamkeitserwartungen wird vor allem durch zwei interagierende Faktoren beeinflusst. Zum einen durch die Entwicklung der Fähigkeit zum symbolischen Denken, v.a. der Fähigkeit, Ursache-Wirkungs-Beziehungen zu erkennen sowie die Fähigkeit zur Selbstbeobachtung und -reflexion. Zum anderen wird die Selbstwirksamkeit auch durch die Responsivität der Umwelt beeinflusst, insbesondere durch die Reaktionen der unmittelbaren sozialen Umgebung auf die Versuche des Kindes, diese zu manipulieren bzw. kontrollieren. Ein gewisses Gefühl der Kontrolle über unser Verhalten, unsere Umwelt, unsere Gedanken und Gefühle ist nämlich essenziell für Zufriedenheit und ein gewisses Wohlbefühl. Wenn die Welt vorherseh- und vorhersagbar erscheint, sind wir eher in der Lage, den Herausforderungen des Lebens zu begegnen und persönliche Zufriedenheit zu erreichen (Maddux, 2001).

Die Selbstwirksamkeit einer Person verändert sich jedoch nicht nur in der Kindheit, sondern über das ganze Leben hinweg, da wir kontinuierlich Informationen aus fünf verschiedenen Quellen integrieren (Maddux, 2001):

1. *Performance experiences* (Leistungserfahrungen). Die eigenen Versuche, die persönliche Umwelt zu kontrollieren, sind die einflussreichste Quelle von Selbstwirksamkeitsüberzeugungen.
2. *Vicarious experiences* (Stellvertretererfahrungen). Selbstwirksamkeit wird auch durch die Beobachtung des Verhaltens anderer und den Konsequenzen aus deren Verhalten beeinflusst, was Bandura (1977) als „Lernen am Modell“ bezeichnete. Dies ist dem Forscher zufolge die einflussreichste Komponente bei der Ausbildung der Selbstwirksamkeitsüberzeugung. Vor allem durch die Beobachtung von Verhalten und der darauf folgenden Reaktionen erkennen Individuen, welches Verhalten in welcher Situation angemessen ist bzw. zum Erfolg führt.

3. *Imaginal experiences* (mental vorgestellte Erfahrungen). Man kann die eigenen Selbstwirksamkeitsüberzeugungen beeinflussen, indem man sich vorstellt, wie man in-/ effektiv in hypothetischen Situationen handeln.
4. *Verbal persuasion* (verbale Überzeugung). Selbstwirksamkeitsüberzeugungen werden auch durch Feedback beeinflusst bzw. dadurch, dass andere einer Person mitteilen, ob sie der Person etwas zutrauen oder nicht. Je nach Expertentum, Glaubhaftigkeit und Attraktivität der Quelle können verbale Überzeugungen mehr oder weniger auf die Selbstwirksamkeit einer Person wirken.
5. *Physiological and emotional states* (physiologische und emotionale Zustände). Wenn schlechte Leistungen oder wahrgenommene Fehler mit aversiven physiologischen Reizen, wie z.B. physiologischer Erregung, und Erfolg mit angenehmen Gefühlen einhergehen, beeinflussen diese intrapersonellen Wahrnehmungen ebenfalls das Selbstwirksamkeitsempfinden einer Person. Im Falle der Wahrnehmung negativer emotionaler oder physiologischer Zustände rechnet man allgemein eher mit schlechten Leistungen.

Durch die verschiedenen o.g. Lernerfahrungen entstehen kognitive Repräsentationen über die Folgen zukünftigen Verhaltens, aus welchen Individuen die Motivation für ihre Handlungen ziehen. Wahrgenommene (negative) Diskrepanzen zwischen Verhalten und Zielerreichung rufen Unzufriedenheit hervor und sind die Ursache für Verhaltensänderungen und letztlich auch für Veränderungen bzw. Anpassungen des Selbstwirksamkeitsempfindens einer Person.

Menschen verändern ihre Selbstwirksamkeit beständig, indem sie Informationen kognitiv verarbeiten, gewichten und in bestehende Netzwerke integrieren. Ihr Verhalten und Bemühen passen sie entsprechend an diese Schemata an. Allerdings ist das individuelle Selbstwirksamkeitsempfinden einer Person resistent gegenüber kurzfristigen Änderungen, da Rückschläge bei der Erreichung eines Zieles in der Regel als Herausforderungen interpretiert und nicht als persönliche Unzulänglichkeit angesehen werden. Dies ist darauf zurück zu führen, dass zur Erreichung großer Leistungen in der Regel Persistenz und Ausdauer vonnöten sind (Bandura & Wood, 1989). Selbstwirksamkeit ist folglich das kumulierte Ergebnis der individuellen Erfolgs- bzw. Misserfolgsbilanz einer Person.

Die Selbstwirksamkeit einer Person ist Bandura (1977) zufolge darüber hinaus abhängig von drei situativen bzw. aufgabenspezifischen Bedingungen. Die erste Bedingung ist der Schwierigkeitsgrad einer Anforderung. Je nach Schwierigkeitsgrad (gering vs. hoch) ist das Selbstwirksamkeitsempfinden einer Person entweder hoch, wenn diese sich das Lösen der Aufgabe zutraut, oder niedrig, wenn ein Misserfolg wahrscheinlicher erscheint. Als zweite Bedingung führt Bandura (1977) die Allgemeingültigkeit der bereits gemachten Erfahrungen einer Person an. Je umfassender und genauer abgestimmt auf die Situation diese Erfahrungen sind, desto höher ist das Selbstwirksamkeitsempfinden der Person. Die letzte Bedingung betrifft die Stärke der Erfolgszuversicht. Je eher eine Person vom Erfolg einer Handlung überzeugt ist, desto eher wird sie eine bestimmte zielführende Verhaltensweise zeigen. Alle drei Bedingungen müssen genau abgewogen werden, um die Selbstwirksamkeit einer Person und folglich das Verhalten abschätzen zu können.

Darüber hinaus postulierte Bandura (1977), dass auch Kontextfaktoren wie beispielsweise soziale, situationsspezifische oder zeitliche Umstände, denen Ereignisse unterliegen, Selbstwirksamkeitserwartungen beeinflussen würden. Das Selbstwirksamkeitsempfinden kann also je nach Situation variieren (Abele et al., 2000). Bandura (1977) betont auch, dass Selbstwirksamkeit keine Disposition ist, wie es beispielsweise die Traittheoretiker über die Big 5 sagen. Selbstwirksamkeit muss zudem von Ergebniserwartungen abgegrenzt werden, die sich auf potenzielle Resultate von Handlungen beziehen. Selbstwirksamkeit betrifft nur die individuelle Handlungskontrolle und -ausführung, nicht das Ergebnis der Handlung (Abele et al., 2000; Bandura, 1977; Schwarzer, Bäßler, Kwiatek, Schröder & Zhang, 1997). Ergebnis- und Konsequenzerwartungen unterliegen demgegenüber eher dem Einfluss von Umweltgegebenheiten.

Das Denken betreffend verbessert ein hoch ausgeprägtes Selbstwirksamkeitsempfinden die kognitiven Denkprozesse und Leistungen in den verschiedensten Situationen. Menschen mit hohem Selbstwirksamkeitsempfinden fokussieren vor allem die Problemlösung. Sie denken organisiert und erarbeiten konstruktive Lösungsstrategien. Personen mit niedriger Selbstwirksamkeit hingegen zweifeln an den eigenen Fähigkeiten und beschäftigen sich vor allem mit Bewertungen potenzieller Handlungsergebnisse. Sie richten ihren Aufmerksamkeitsfokus nach innen und erleben folglich auch mehr Stress (Bandura & Wood, 1989).

Bandura (1977) postulierte, dass Menschen, die sich mehr anstrengen, auch mehr Erfolg haben. Diese Annahme besagt, dass Menschen, die davon überzeugt sind, eine bestimmte Handlung erfolgreich ausführen zu können oder ein Problem lösen zu können, an den betreffenden Aufgaben ausdauernder und intensiver arbeiten, als Menschen mit geringeren Selbstwirksamkeitsüberzeugungen (Schwarzer et al., 1997). Selbstwirksamkeit beeinflusst den Autoren zufolge das Fühlen, Denken und Handeln einer Person. Beispielsweise fühlen Menschen mit niedrigem Selbstwirksamkeitsempfinden sich eher depressiv, hilflos und ängstlich sowie haben ein eher pessimistisches Zukunfts- und Weltbild. Ursprünglich wurde das Konzept zur Erforschung von Phobien entwickelt.

Anhand klinischer Stichproben untersuchte Bandura (1977), wie sich das Selbstwirksamkeitsempfinden auf die Bewältigung von potenziell angstausslösenden Situationen auswirkt. Die Ergebnisse zeigten, dass Selbstwirksamkeit für 85 % aller Aufgaben und 83 % der Aufgaben, die zur Prätestung von den Probanden selbst noch als „nicht schaffbar“ eingeschätzt wurden, als Prädiktor fungierte. Bandura (1977) führte an, dass Selbstwirksamkeit demzufolge ein besserer Verhaltensprädiktor für die Bewältigung unbekannter Aufgaben sei als vergangenes Verhalten.

Das Konstrukt der Selbstwirksamkeit wird darüber hinaus als ein wesentlicher Bestandteil in Motivationsprozessen angesehen. So wählen Menschen mit hoch ausgeprägter Selbstwirksamkeit schwierigere Aufgaben aus, setzen sich höhere Ziele und verfolgen diese auch länger (Bandura & Wood, 1989; Schwarzer et al., 1997). Dies ist auf die gedankliche Antizipation einer Handlung, die sich am jeweiligen positiven oder negativen Zukunfts- und Weltbild orientiert, zurückzuführen. Bandura und Wood (1989) zufolge erhöht das Selbstwirksamkeitsempfinden direkt die Motivation einer Person, wirkt aber auch indirekt über die erhöhte Zielsetzung motivierend. Dabei sollte man Selbstwirksamkeit jedoch nicht mit positiven Illusionen oder unrealistischem Optimismus verwechseln. Bandura (1977) merkte hierzu an, dass Erwartungen allein keinen Erfolg nach sich ziehen, wenn wichtige Komponenten, die Erfolg begründen, wie beispielsweise notwendige Übung, fehlen. Darüber hinaus ist Leistung auch von äußeren Bedingungsfaktoren abhängig. Bandura und Wood (1989) untersuchten an einer Stichprobe von Wirtschaftsstudenten ($N=60$), inwiefern das Selbstwirksamkeitsempfinden abhängig ist von der subjektiv empfundenen Kontrollierbarkeit einer Situation. Die Annahme der Studie war, dass Personen, denen suggeriert wird, sie könnten in einer Situation nicht viel ausrichten, eine geringere Selbstwirksamkeitserwartung und folglich weniger Erfolg haben, als Personen, denen suggeriert wird, sie könnten viel bewirken. In einer Computersimulation konnten die Autoren erwartungsgetreu zeigen, dass das subjektive Selbstwirksamkeitsempfinden der Versuchsteilnehmer durch unkontrollierbare situative Bedingungen verringert werden konnte und diese folglich auch unter eigentlich guten Voraussetzungen für Erfolg geringere Leistungen zeigten als Personen, die die gleiche Aufgabe unter schwierigeren, aber wissentlich kontrollierbaren Außenbedingungen bearbeiteten.

Schwarzer und Mitarbeiter (1997) postulieren, dass ein ausgeprägtes Selbstwirksamkeitsempfinden mit besserer Gesundheit, erhöhter Leistung und mehr sozialer Integration einhergeht.

Einige Theorien zur Selbstwirksamkeit gehen davon aus, dass es eine „globale“ Selbstwirksamkeit gibt, die sich auf die allgemeine Zuversicht eines Individuums bezieht, verschiedene neuartige Situationen erfolgreich bewältigen zu können (Schwarzer et al., 1997). Dieser dispositionelle Optimismus bezieht sich jedoch nur auf allgemeine individuelle Kompetenz- und Ressourcenüberzeugungen und lässt andere Quellen für erhöhtes Selbstwirksamkeitsempfinden außer Acht. Bandura (1977) lehnt eine globale Erfassung des Selbstwirksamkeitskonzepts ab und betont den Vorteil situativer bzw. bereichsspezifischer Erfassung von Selbstwirksamkeit. Ihm zufolge ist es bei der Erhebung von Selbstwirksamkeit wichtig, die Aufgabe, auf die sich die Einschätzung bezieht, zu spezifizieren. Dies basiert auf der Idee, dass Prädiktor und Kriterium immer den gleichen Spezifitätsgrad haben sollten. Nur dann ist eine gute Vorhersage möglich.

Schyns und von Collani (2002) unterscheiden zwischen globaler, arbeitsbezogener und aufgabenbezogener Selbstwirksamkeit. Während globale Selbstwirksamkeit die allgemeine Überzeugung einer Person widerspiegelt, unbekannte Situationen erfolgreich bewältigen zu können, erfasst arbeitsbezogene Selbstwirksamkeit den Autoren zufolge eher Erfolgserwartungen über verschiedene berufliche Tätigkeiten, Organisationen und Hierarchieebenen hinweg. Aufgabenbezogene Selbstwirksamkeit erfasst demgegenüber sehr spezifisch, inwiefern ein Arbeitnehmer kognitiv und emotional davon überzeugt ist, eine bestimmte Aufgabe meistern zu können. Larson und Borgen (2006) definieren aufgabenspezifische Selbstwirksamkeit als *“the confidence to successfully execute a desired action in the near future”* (S. 296). Die aufgabenbezogene Selbstwirksamkeit muss für jede Tätigkeit neu abgefragt werden und kann je nach Aufgabe und Erfahrungshintergrund auch intraindividuell erheblich variieren.

Das Selbstwirksamkeitskonzept wird viel in der organisationalen Forschung verwendet. Laut Judge, Jackson, Chaw und Rich (2007) wurde Selbstwirksamkeit bis dahin in mehr als 10.000 Untersuchungen erforscht. Vor allem in der Organisationspsychologie in der Erforschung von beispielsweise Lernerfolg oder Leistungsbewertungen ist das Konzept sehr populär. So wird u.a. der Einfluss der Selbstwirksamkeit von Mitarbeitern in einem Unternehmen in Bezug auf dessen Auswirkungen auf die Unternehmensergebnisse wie z.B. Leistungsvariablen (Sadri & Robertson, 1993; Stajkovic & Luthans, 1998) oder Arbeitszufriedenheit (Judge & Bono, 2001) untersucht. In ihrer Metaanalyse fanden Judge und Bono (2001) eine Korrelation von $r = .45$ zwischen Arbeitszufriedenheit und (allgemeiner) Selbstwirksamkeit. Arbeitsleistung und Selbstwirksamkeit hängen auch positiv miteinander zusammen ($r = .23$).

Zu Beginn der Erforschung des Selbstwirksamkeitskonstrukts im beruflichen Kontext wurde vor allem zum Zusammenhang mit gesundheitsrelevantem Verhalten geforscht. So sind geringe Selbstwirksamkeitserwartungen ein Indikator für Depression (Bandura, 1997; Maddux & Meier, 1995). Dysfunktionale Angst und Vermeidungsverhalten sind oftmals ein direktes Ergebnis von geringen Selbstwirksamkeitserwartungen. Menschen, die über ein starkes Selbstbewusstsein verfügen, was ihre Fähigkeiten, potenziell schwierige Situationen erfolgreich zu meistern betrifft, werden solchen Situationen eher ruhig begegnen. Personen, denen es an dieser Zuversicht mangelt, werden diesen Situationen eher mit Besorgnis begegnen und somit die Wahrscheinlichkeit, dass sie effektiv handeln, verringern (Maddux, 2001). Selbstwirksamkeit beeinflusst die Durchführung von Gesundheitsverhalten (Bandura, 1997). Die Verstärkung von Selbstwirksamkeitsüberzeugungen ist ein wesentliches Element in der Therapie, um eine erfolgreiche Veränderung und Aufrechterhaltung von gesundheitsrelevantem Verhalten zu erreichen. Zudem beeinflussen Selbstwirksamkeitsüberzeugungen die physiologischen Antworten auf Stress, was auch das Immunsystem betrifft. Ein subjektiver Mangel an wahrgenommener Kontrolle über die Anforderungen der Umwelt kann die Anfälligkeit für Infektionen erhöhen und die Heilungschancen von Krankheiten verringern. Selbstwirksamkeitsüberzeugungen können auch nachweislich die Aktivierung von Katecholaminen (Neuro-

transmitter, die für die Stressregulierung zuständig sind) und Endorphinen (= endogenen Schmerzmitteln) beeinflussen. Selbstwirksamkeit hängt daher nachweislich mit verschiedenen gesundheitsrelevanten Konstrukten zusammen, wie z.B. Self-Esteem, Locus-of-Control oder Neurotizismus (Judge, Bono & Locke, 2000; Sherer & Adams, 1983). In ihrer Metaanalyse fanden Judge, Erez und Bono (1998) hohe Zusammenhänge zwischen diesen vier Konstrukten und konnten auch einen Einfluss auf Arbeitsleistung nachweisen (s. auch Judge & Bono, 2001).

Betz und Hackett (1981) übertrugen das Konzept der Selbstwirksamkeit zum ersten Mal in den beruflichen Kontext, da sie eine Ursache für die Unterrepräsentanz von Frauen in Führungspositionen suchten. Innerhalb der letzten dreißig Jahre wuchs daher das Interesse an beruflicher Selbstwirksamkeit immer mehr an, da das Konstrukt mit beruflichem Erleben und karriererelevantem Verhalten sowohl korreliert als auch in der Lage ist, diese Variablen vorherzusagen (Abele et al., 2000).

Lent, Brown und Hackett (1994) formulierten auf Basis ihrer Erkenntnisse die „Sozialkognitive Theorie der Karriereentwicklung“, der zufolge Selbstwirksamkeits- und Ergebniserwartungen sowie Ziele relevante Prädiktoren für eine erfolgreiche Berufsausübung darstellen.

In der organisationalen Forschung konnte beispielsweise nachgewiesen werden, dass Selbstwirksamkeit mit beruflicher Zufriedenheit (Judge & Bono, 2001) und Leistung zusammenhängt (Judge & Bono, 2001; Stajkovic & Luthans, 1998). Rigotti, Schyns und Mohr (2008) zufolge ist dies damit erklärbar, dass Selbstwirksamkeit als persönliche Ressource angesehen werden kann, die als „Puffer“ zwischen Stressor und Beanspruchung agiert. Es ist auch möglich, dass Selbstwirksamkeit als Mediator des Zusammenhangs zwischen Persönlichkeitseigenschaften und Karriereinteressen fungiert (Dreier, 2008). In einigen Feldstudien konnte zudem nachgewiesen werden, dass Vorgesetztenbeurteilungen einen Einfluss auf die Selbstwirksamkeit ihrer Mitarbeiter hatten (Eden & Kinnar, 1991; Natanovich & Eden, 2001). Murphy und Ensher (1999) konnten in einer Längsschnittstudie zeigen, dass Vorgesetztenbeurteilungen die Selbstwirksamkeit der Mitarbeiter erhöhen können.

Berufliche Selbstwirksamkeit und Aufgabenanforderungen korrelieren positiv, aber etwas geringer als erwartet (Schyns & von Collani, 2002). Dies könnte ein Hinweis darauf sein, dass diese Eigenschaft vergleichsweise stabil ist und sich nur marginal durch neue Aufgaben, eine neue Arbeit o.ä. verändert, sondern v.a. auch durch Erfahrungen in vorherigen Arbeitsverhältnissen herausgebildet wird.

Auch konnten Zusammenhänge zwischen beruflicher Selbstwirksamkeit und Berufen (Donnay & Borgen, 1999) sowie akademischen Zielen (Rottinghaus, Lindley, Green & Borgen, 2002) nachgewiesen werden. Allerdings sind dies lediglich einfache Zusammenhänge, die nur wenig über die prädiktive Validität des Konzeptes aussagen.

Bell und Kozlowski (2002) untersuchten an einer Stichprobe von College-Studenten ($N=125$), inwiefern Personen, die in der Lage sind, auch in schwierigen Situationen erfolgreich zu sein und Aufgaben erfolgreich zu bewältigen, auch eine hohe Selbstwirksamkeitsüberzeugung aufweisen. Personen mit geringen Fähigkeiten sollten demgegenüber eher geringe Selbstwirksamkeitserwartungen haben. Die Ergebnisse der Studie zeigten signifikante Zusammenhänge von Selbstwirksamkeit mit allgemeinen Fähigkeiten ($r=.28$), Leistung ($r=.53$) und Wissen ($r=.46$).

Die prädiktive Validität von beruflicher Selbstwirksamkeit wurde auch viel bei Berufsanfängern untersucht. Diese Personengruppe ist für die Selbstwirksamkeitsforschung besonders interessant, da Absolventen in der Regel noch über keine oder nur wenig Berufserfahrung verfügen, aus der sie Fähigkeitseinschätzungen bzw. -erwartungen ableiten könnten. Übereinstimmend mit den Annahmen des sozial-kognitiven Modells der Karriereentwicklung (Lent et

al., 1994) fanden unabhängige Untersuchungen signifikante Zusammenhänge von Selbstwirksamkeitsüberzeugungen mit schulischen und beruflichen Leistungsindikatoren bei Berufsanfängern, wie z.B. Studiendauer, Examensnote oder einen erfolgreichen Berufsstart (Abele et al., 2000).

Abele et al. (2000) untersuchten anhand von zwei Stichproben von Hochschulabsolventen ($N_1 = 1500$ und $N_2 = 430$ Absolventen) die prädiktive Validität eines von ihnen selbst entwickelten berufsbezogen formulierten Fragebogens zur Erfassung von Selbstwirksamkeit. Sie fanden signifikante Zusammenhänge zwischen beruflicher Selbstwirksamkeit und Studiendauer sowie Examensnote. Allerdings waren die Zusammenhänge eher niedrig ausgeprägt. Mit einem erfolgreichen Berufsstart korrelierte die berufsbezogene Selbstwirksamkeitsskala jedoch hoch. Die gefundenen Effekte sprechen den Autoren zufolge dafür, dass berufliche Selbstwirksamkeit Aspekte eines erfolgreichen Berufsstarts anspricht, die durch beispielsweise Leistungsmaße im engeren Sinne nicht erfasst werden. Auch für Trainings- und Lernprozesse sei berufliche Selbstwirksamkeit den Autoren zufolge interessant.

Lee et al. (2010) versuchten, eine Auswahlssituation herzustellen, indem sie ihren Versuchspersonen ($N = 694$ College-Studenten) suggerierten, dass diese sich bei einem Unternehmen bewerben würden. Die „Bewerber“ bearbeiteten u.a. einen Persönlichkeitsfragebogen (NEO-FFI) und eine allgemeine Selbstwirksamkeitsskala. Teilnehmer, die einen signifikant erhöhten Gewissenhaftigkeitswert aufwiesen, wurden als „Faker“ klassifiziert. Die Ergebnisse deuteten an, dass das Ausmaß, in dem Faker ihren Gewissenhaftigkeitswert veränderten mit Selbstwirksamkeit korrelierte ($r = -.10$, $p < .05$). Das heißt, Personen, die ihre Testwerte verfälschten, wiesen eine geringere Selbstwirksamkeit auf als ehrliche Kandidaten. Eine mögliche Erklärung für diesen Effekt ist den Autoren zufolge, dass Menschen mit nur gering ausgeprägter Selbstwirksamkeitserwartung nicht daran glauben, dass sie eine gewünschte Position allein aufgrund ihrer (ehrlichen) Eigenschaften erhalten (können) und daher ihre Werte in eine sozial erwünschte Richtung verzerren. Lee et al. (2010) warnten, dass dieser Effekt die Kriteriumsvalidität eines Selbstwirksamkeitsmaßes empfindlich stören könne, da die später gezeigte Leistung nicht den wirklichen Annahmen entspreche.

Rigotti und Kollegen (2008) untersuchten in einer kulturübergreifenden Studie 1.535 Personen aus fünf verschiedenen europäischen Ländern hinsichtlich ihrer beruflichen Selbstwirksamkeitserwartung. Die Ergebnisse ihrer Studie ergaben zum einen, dass sich die Studienteilnehmer hinsichtlich ihrer demografischen Variablen bzw. deren Zusammenhang mit Selbstwirksamkeit nicht deutlich voneinander unterschieden. Lediglich in der deutschen Subgruppe gab es einen leichten Alterseffekt ($r = .28$, $p < .001$) dahingehend, dass die Selbstwirksamkeitserwartung mit dem Alter anstieg. Bezüglich der Kriteriumsvalidität fanden Rigotti et al. (2008), dass Selbstwirksamkeit mit Zufriedenheit im Job ($r = .17$, $p < .05$), Commitment ($r = .19$, $p < .01$), Arbeitsleistung ($r = .32$, $p < .01$) und Jobsicherheit ($r = -.19$, $p < .01$) korrelierte.

Zwei Metaanalysen haben gezeigt, dass Selbstwirksamkeit positiv mit beruflicher Leistung zusammenhängt (Sadri & Robertson, 1993; Stajkovic & Luthans, 1998). Einhergehend mit theoretischen Annahmen, dass Menschen, die eine höhere Selbstwirksamkeit aufweisen, auch länger an Aufgaben arbeiten, allgemein ausdauernder sind und sich höhere Ziele setzen, konnten höherer Leistungen bei Personen mit hohen Ausprägungen der Selbstwirksamkeit gefunden werden. Hoch selbstwirksame Individuen fühlen sich darüber hinaus dem Unternehmen gegenüber verbunden, das Aufgaben für sie bereithält, bei deren Erfüllung sie gute Leistungen zeigen können.

Selbstwirksamkeitsforscher haben sowohl allgemeine (z.B. Schwarzer & Jerusalem, 1995) als auch berufsbezogene Selbstwirksamkeitsinventare (z.B. Abele et al., 2000; Schyns & von Collani, 2002) entwickelt, um das Konstrukt standardisiert erfassen zu können. Allgemeine

Selbstwirksamkeitsmaße haben sich in der Forschung als nicht so nützlich erwiesen wie spezifische Selbstwirksamkeitsmaße (Abele et al., 2000; Maddux, 2001; Schyns & von Collani, 2002). Aber auch eine aufgabenspezifische Erfassung der Selbstwirksamkeit macht (vor allem im organisationalen Kontext) keinen Sinn, wenn man die verschiedenen Mitarbeiter miteinander vergleichen will. In Summe spricht die Befundlage für eine tätigkeitsübergreifende, aber arbeitsbezogene Erfassung des Selbstwirksamkeitskonstrukts im Rahmen von Personalauswahl- und -entwicklungsverfahren. Bereichs- bzw. berufsspezifische Selbstwirksamkeit bezieht sich auf die Überzeugung einer Person, dass sie die notwendigen Fähigkeiten und Fertigkeiten besitzt, um eine berufliche Aufgabe oder Tätigkeit erfolgreich bewältigen und erfüllen zu können (Rigotti et al., 2008).

Bandura (1997) deutete eine Verbindung zwischen Gewissenhaftigkeit und Selbstwirksamkeit an, da *“people see the extraordinary feats of others, but not the unweavering commitment and countless hours of perseverant effort that produce them”* (S. 119). D.h. man würde in Betracht hervorragender Leistungen anderer diese eher auf Talent als auf Bemühungen attribuieren. Selbstwirksamkeitsüberzeugungen, Einsatz und Ausdauer seien aber wichtigere Prädiktoren als lediglich Talent. Menschen, die über eine allgemein hoch ausgeprägte Gewissenhaftigkeit verfügen, werden mit vielen positiven Attributen beschrieben, die im Arbeitsbereich vorteilhaft sind. Hierzu zählt oftmals auch eine höhere Selbstwirksamkeit (Minbashian, Wood & Beckmann, 2010).

Es ist möglich, dass die beiden Konstrukte über die gemeinsamen Aspekte der Motivation miteinander verbunden sind (Judge et al., 2007). Hierzu liefert die Metaanalyse von Barrick und Mount (2004) Hinweise. Personen, die über eine hohe Gewissenhaftigkeit verfügen, werden sich auch eher (anspruchsvolle) Ziele setzen und höhere Erwartungen haben, dass ihre Bemühungen erfolgreich sein werden. Dies wiederum führt zu einer erhöhten Leistungsmotivation. Den Autoren zufolge zeigte Selbstwirksamkeit mit Persönlichkeitsvariablen und Leistungsmotivation einen mittleren Zusammenhang ($r = .49$). Sie sahen die Persönlichkeitseigenschaften als distale Motivation an, die wiederum die proximale Motivation bzw. Selbstwirksamkeit beeinflusse. Die Erforschung dieser Zusammenhänge sollte signifikant zum Verständnis des Zusammenhangs von Gewissenhaftigkeit mit Erfolgs- bzw. Leistungskriterien beitragen.

Nauta (2004) ließ $N = 147$ College-Studenten Testverfahren zur Erfassung der Big 5, arbeitsbezogene Selbstwirksamkeit sowie Karriereinteressen bearbeiten. Die Autorin identifizierte Selbstwirksamkeit als Mediator des Zusammenhangs zwischen den Persönlichkeitseigenschaften und Karriereinteressen, was sie als Bestätigung der sozial-kognitiven Lerntheorie im akademisch-beruflichen Kontext wertete. Dieses Ergebnis wurde von Dreier (2008) teilweise bestätigt.

Zusammenfassend lässt sich sagen, dass Selbstwirksamkeit als Prädiktor für Berufserfolg, Zufriedenheit, Bindung an das Unternehmen sowie Arbeitsleistung angesehen werden kann (Rigotti et al., 2008), was auch in verschiedenen Metaanalysen bestätigt werden konnte (Judge & Bono, 2001; Stajkovic & Luthans, 1998). Auch schulischer Erfolg kann durch Selbstwirksamkeit vorhergesagt werden (Gore, 2006; Okech & Harrington, 2002). Untersuchungen haben weiterhin die Verbindung von Selbstwirksamkeit mit Gewissenhaftigkeit aufzeigen können. Positive Zusammenhänge zwischen den beiden Konstrukten wurden vor allem in Bezug auf Training und Lernerfahrungen gefunden (Barrick & Mount, 1991; Chen et al., 2001; Salgado, 1997). Vor allem für Berufsanfänger ist berufliche Selbstwirksamkeit ein wichtiger Prädiktor von beispielsweise Ausbildungserfolg (Abele et al., 2000), weshalb das Konstrukt bei der Vorhersage von Schulabschluss- und IHK-Noten einen wichtigen Beitrag leisten kann.

Neben Selbstwirksamkeit gibt es ein weiteres Konstrukt, die *wahrgenommenen eigenen Fähigkeiten*, das ebenfalls einen Einfluss auf den Zusammenhang zwischen den Gewissenhaftigkeitsaspekten und Ausbildungserfolg haben könnte. Dieses wird im folgenden Abschnitt (2.3.3) näher erläutert.

2.3.3 Wahrgenommene eigene Fähigkeiten

Als vielversprechende beeinflussende Variable bei der Vorhersage von akademischer Leistung haben sich die wahrgenommenen eigenen Fähigkeiten gezeigt (Dreier, 2008; Nofle & Robins, 2007). Brown, Lent und Gore (2000) bezeichnen mit *wahrgenommenen eigenen Fähigkeiten* bzw. *perceived abilities* (PA) Selbsteinschätzungen der beruflichen oder schulischen Fähigkeiten einer Person. Messungen dieser Einschätzungen beinhalten in der Regel normative Urteile, d.h. Vergleiche der eigenen Leistung mit der von Kollegen, Mitschülern oder anderen relevanten Vergleichsgruppen.

Das Konzept wurde bisher vor allem in der Sportpädagogik verwendet. Hier fungiert es als Indikator der Motivation einer Person, besser zu sein als andere. Erfolg wird durch den Vergleich der eigenen Leistung mit der anderer während des Trainings oder im Wettkampf definiert. Die Einschätzung der eigenen Fähigkeiten erwies sich in diesem Kontext als wichtiger Mediator bei der Erreichung von Zielen und der Attribuierung von Erfolg (Cury, Sarrazin & Famose, 1997). Schätzt man die eigenen Fähigkeiten als hoch ein, haben die Personen mit als hoch wahrgenommenen eigenen Fähigkeiten eher Erfolg, als Personen, die die eigenen Fähigkeiten als vergleichsweise schlechter wahrnehmen.

Aber auch im Schulkontext wird das Konzept der wahrgenommenen eigenen Fähigkeiten zunehmend angewendet und erforscht. Staff, Schulenberg und Bachmann (2010) fanden heraus, dass der akademische Einsatz im Sinne von „das Beste geben“ positiv mit Schulleistung korrelierte ($r = .36$). Keith, Diamond-Hallam und Fine (2004) konnten zeigen, dass die Zeit, die jemand mit Hausaufgaben verbrachte, positiv mit Schulnoten korrelierte ($r = .20$ bis $.31$). In einer neueren Metaanalyse erhielten Cooper, Robinson und Patall (2006) ähnliche Ergebnisse zum Zusammenhang zwischen Hausaufgabenzeit und Schulnoten ($r = .14$). Nauta, Epperson und Wagoner (1999) zeigten, dass ausdauernde Studenten dazu tendieren, den eigenen Erfolg eher ihren Bemühungen als ihren Fähigkeiten zuzuschreiben. Dieser Effekt zeigte sich auch noch, nachdem die Intelligenz der Studenten kontrolliert wurde.

Nofle und Robins (2007) verglichen in einer Studie mit College-Studenten ($N = 10.497$) die Vorhersagekraft verschiedener Gewissenhaftigkeitsmaße in Bezug auf akademische Leistung. Die Autoren nahmen die wahrgenommenen eigenen Fähigkeiten als beeinflussende Variable in der Studie mit auf und ermittelten hohe Zusammenhänge zwischen GPA und akademischen Fähigkeiten ($r = .60$ bis $.61$) sowie dem selbsteingeschätzten Aufwand für die Ausbildung ($r = .17$ bis $.25$). Auch mit Gewissenhaftigkeit zeigten die wahrgenommenen eigenen Fähigkeiten deutliche Zusammenhänge ($r = .35$ bis $.37$). Konsistent mit sozial-kognitiven Ansichten über Persönlichkeit und Leistung nahmen die Autoren an, dass Gewissenhaftigkeit sich in der realen Welt durch die Gedanken und Gefühle einer Person ausdrücken würde (d.h. die Selbstwahrnehmung) und so das eigentliche Verhalten beeinflusst. Sie nahmen an, dass Gewissenhaftigkeit nur deshalb mit höheren Noten assoziiert ist, da diese Eigenschaft in hohen Ausprägungen auch einen allgemein höheren Einsatz und als höher wahrgenommene akademische Fähigkeiten mit sich bringe. Diese Hypothese wurde von den Autoren mittels Mediationsanalyse überprüft. Die Ergebnisse der Studie berichten von einer signifikanten Reduktion des Zusammenhangs zwischen Gewissenhaftigkeit und GPA von $r = .20$ auf $r = .06$ bzw. $r = .19$ auf $r = .05$, wenn die wahrgenommenen eigenen Fähigkeiten in die Regression hinzugenommen wurden. Die wahrgenommenen eigenen Fähigkeiten erwiesen sich somit als signifikanter Mediator des Zusammenhangs zwischen Gewissenhaftigkeit und Schulleistung.

Furnham, Chamorro-Premuzic und McDougall (2003) verglichen in einer ähnlichen, aber kleineren Studie mit 93 Studenten die Zusammenhänge zwischen selbsteingeschätzter Intelligenz mit Persönlichkeitseigenschaften und Intelligenz. Sie untersuchten ein den wahrgenommenen Fähigkeiten ähnliches Konstrukt, das sie *Beliefs About Intelligence* (= BAI) nannten. BAI beschreiben den Autoren zufolge die Ausprägung der subjektiven Überzeugung einer Person, ob und inwiefern (die eigene) Intelligenz über das Leben hinweg zunehmen kann oder nicht. Den Autoren zufolge sind realistische und zuversichtliche Erwartungen an die eigene Leistung bzw. Leistungsfähigkeit mit besseren Schulleistungen verbunden. Negative Annahmen über die eigenen kognitiven Eigenschaften oder Annahmen über die Überlegenheit der eigenen Fähigkeiten können demgegenüber auch in geringeren Leistungen resultieren. Wenn eine Person beispielsweise von der Überlegenheit der eigenen Fähigkeiten überzeugt ist oder denkt, dass Lernen sich sowieso nicht lohnt, wird dies zu geringerem Einsatz, beispielsweise hinsichtlich des Lernens für Prüfungen, und letztlich in geringeren Leistungen resultieren.

Die Ergebnisse von Furnham et al. (2003) zeigten eine signifikante Korrelation der Gewissenhaftigkeitsdimension mit selbsteingeschätzter Intelligenz ($r = .29$). Wenn Geschlecht und selbsteingeschätzte Intelligenz auspartialisiert wurden, korrelierten Gewissenhaftigkeit und Schulleistung signifikant miteinander in dem Sinne, dass Personen, die glaubten, dass Intelligenz unveränderbar sei und es keinen Mehrwert bringe, wenn man sich um höhere Leistungen bemühe, eine geringere Gewissenhaftigkeit aufwiesen. Weder BAI noch Intelligenz waren allerdings signifikante Prädiktoren von Schulleistung. Ob jemand Erfolg in der Schule hat, basiert diesen Ergebnissen zufolge somit eher auf der Persönlichkeit und weniger auf der Überzeugung, ob jemand glaubt, dass Intelligenz durch harte Arbeit verändert werden kann. Allerdings zeigte sich dieser Effekt bei den gewissenhafteren Schülern. Dies könnte somit auf eine Art „selbsterfüllende Prophezeiung“ hindeuten, in dem Sinne, dass gewissenhafte Studenten gelernt haben, dass sich Einsatz lohnt, während weniger gewissenhafte Studenten diese Lernerfahrung (noch) nicht machen konnten.

Hierzu passen auch die Ergebnisse von Andresen (eingereicht), dass Selbsteingeschätzte eigene Intelligenz mittlere bis hohe Korrelationen mit den von ihm postulierten W- und C-Faktoren der Gewissenhaftigkeit zeigt. Auch Corker et al. (2012) setzten in ihrer Untersuchung zwei Items zur Erfassung des Einsatzes ein („*I put a lot of effort into preparing for this exam.*“ und „*I worked very hard to prepare for this exam.*“). Die Ergebnisse ihrer Studie deuten an, dass der damit gemessene Einsatz (*Effort*) als Mediator des Zusammenhangs zwischen Gewissenhaftigkeit und Leistung agiert.

Duckworth und Seligman (2005) fanden heraus, dass Selbstdisziplin (eine Facette von Gewissenhaftigkeit) positiv mit der Anzahl der Stunden, die mit Hausaufgaben verbracht wurden, und negativ mit Abwesenheitszeiten von der Schule korrelierte. Beides kann als Indikator für die Bemühungen angesehen werden, die wiederum schulischen Erfolg vorhersagen. Trautwein, Lüdtke, Roberts, Schnyder und Niggli (2009) zeigten in einer Studie an deutschen Achtklässlern, dass Gewissenhaftigkeit den akademischen Einsatz in Mathematik und Englisch vorhersagen konnte. Sie fanden darüber hinaus, dass der Einsatz, den ein Schüler zeigte, den Zusammenhang zwischen Gewissenhaftigkeit und Mathematiknoten medierte. Daraus kann man schließen, dass Einsatz und Durchhaltevermögen die Schlüsselmediatoren des Zusammenhangs zwischen Gewissenhaftigkeit und Schulleistung sind.

Aufgrund der theoretischen Nähe von Zielorientierung und individuellen Einschätzungen der eigenen Fähigkeiten, haben einige Studien auch den Einfluss von wahrgenommenen Fähigkeiten auf die Effekte der Zielorientierung untersucht (Kaplan & Midgley, 1997; Miller, Behrens, Greene & Newman, 1993). Diese Arbeiten haben i.d.R. die moderierenden Effekte wahrgenommener Fähigkeiten auf das Verhältnis zwischen Zielorientierung und Verhalten

untersucht. Allerdings gab es keine eindeutigen Ergebnisse und auch aufgrund der angewendeten Methoden ist die Aussagekraft dieser Studien begrenzt (Bell & Kozlowski, 2002).

Zusammenfassend gibt es deutliche Hinweise, dass wahrgenommene eigene Fähigkeiten den Zusammenhang zwischen Gewissenhaftigkeit und (schulischer) Leistung medieren. Rigotti et al. (2008) berichten zudem von Hinweisen, dass wahrgenommene (eigene) Leistungen starke Zusammenhänge mit Selbstwirksamkeit aufweisen. Allerdings liegen bis dato bezüglich dieser Variablen hauptsächlich Ergebnisse aus US-Studien mit College-Studenten vor. Das Konstrukt wurde somit zwar in vergleichbaren Stichproben erforscht, was die Altersstruktur der Probanden betrifft. Allerdings gibt es bis dato keine Hinweise darauf, inwiefern die wahrgenommenen eigenen Fähigkeiten auch das Verhältnis zwischen Gewissenhaftigkeit bzw. deren Facetten in Bezug auf berufliche Ausbildungsleistung mediert. Aufgrund der bisherigen Forschungsergebnisse zu möglichen mediierenden Effekten des Konstrukts, sollen die wahrgenommenen eigenen Fähigkeiten in der vorliegenden Studie in die Modellbildung zur Vorhersage von Ausbildungsleistung mit aufgenommen werden und hinsichtlich ihrer Nützlichkeit überprüft werden.

Der folgende Abschnitt (2.3.4) beschäftigt sich kurz mit der Abgrenzungsfähigkeit der wahrgenommenen eigenen Fähigkeiten vom Selbstwirksamkeitskonstrukt, bevor in Abschnitt 2.4 eine kritische Auseinandersetzung mit Persönlichkeitsmaßen in der beruflichen Eignungsdiagnostik allgemein erfolgt.

2.3.4 Zur Abgrenzung der Selbsteinschätzungskonstrukte

Unterschiedliche Forscher bzw. Forschergruppen haben die Frage aufgeworfen, ob die verschiedenen Selbsteinschätzungskonstrukte eventuell auch zu einer gemeinsamen Dimension zusammengefasst werden könnten. Beispielsweise wird diesbezüglich das Konstrukt der *Core Self-Evaluations* vorgeschlagen, das verschiedene Selbsteinschätzungskonstrukte zusammenfasst (z.B. Selbstwirksamkeit, Self-Esteem, Neurotizismus u.a.). Einhergehend mit Brunswiks *Linsenmodell* der optimalen Vorhersage (1943), haben sich zusammengesetzte Persönlichkeitsvariablen als vorteilhaft für die Vorhersage von allgemeinen Kriterien wie z.B. Arbeitsleistung herausgestellt (Bipp, 2010). Judge, Bono, Erez und Locke (2005) zeigten beispielsweise, dass Personen mit positiver Selbstbeurteilung eigenständig nach Zielen suchen bzw. diese anstreben und weniger von äußerlichen Einflüssen im Sinne einer extrinsischen Motivation dahin gedrängt werden müssen. Bierhoff, Wegge, Bipp, Kleinbeck, Attig-Grabosch und Schulz (2005) warfen die Frage auf, ob die genannten Konstrukte eventuell auch zu einer Dimension *Eigenverantwortung* zusammengefasst werden könnten, die verschiedene kognitive und motivationale Komponenten umfasst, z.B. Gewissenhaftigkeit und Selbstwirksamkeit. Diese Dimension sollte sich von diesen Konstrukten selbst allerdings psychometrisch abgrenzen lassen.

Insbesondere mit Blick auf die verschiedenen Spezifitätsgrade der Selbstwirksamkeit stellt sich daher die Frage, inwieweit sich die Konstrukte wahrgenommene eigene Fähigkeiten und Selbstwirksamkeit voneinander unterscheiden.

Die beiden Konstrukte wahrgenommene eigene Fähigkeiten und Selbstwirksamkeit stammen aus verschiedenen theoretischen Herangehensweisen, weisen aber einige Gemeinsamkeiten auf. So beinhalten beide subjektive Einschätzungen der eigenen Fähigkeiten und in beiden Theorien, die den Konstrukten zugrunde liegen, werden ähnliche Annahmen über die berufliche Bedeutung der Konstrukte gemacht (Dreier, 2008).

Neben diesen Gemeinsamkeiten gibt es aber auch deutliche konzeptuelle Unterschiede. So handelt es sich bei wahrgenommenen eigenen Fähigkeiten lediglich um ein Vergleichsmaß, das nach normativen Urteilen fragt. Selbstwirksamkeit hingegen erfasst Erwartungen über

zukünftige Leistungen (Lent et al, 1994). Auch unterscheiden sich die beiden Konstrukte in Inhalt, Abstraktionslevel, Skalierung und temporären Dimensionen (Brown et al., 2000).

In einer Vergleichsstudie der beiden Konstrukte erhoben Brown und Mitarbeiter (2000) verschiedene Maße zu wahrgenommenen eigenen Fähigkeiten, beruflicher Selbstwirksamkeit, Karriereinteressen und wahrgenommenen Karrieremöglichkeiten an einer studentischen Stichprobe ($N=229$). Ihre Ergebnisse belegen, dass wahrgenommene eigene Fähigkeiten und Selbstwirksamkeitsüberzeugungen zwar korreliert sind, es sich aber dennoch um empirisch voneinander unterscheidbare Konstrukte handelt, die auch unterschiedlich mit den wahrgenommenen Karriereoptionen und Karriereinteressen zusammenhängen.

Diese Ergebnisse konnten mittels konfirmatorischer Faktorenanalysen in einer vergleichbaren Stichprobe ($N=217$ College-Studenten) von Dreier (2008) bestätigt werden. Hierzu beantworteten die Studenten eine Skala zur Einschätzung allgemeiner Selbstwirksamkeit (Schwarzer & Jerusalem, 1995), acht Items zu ihren wahrgenommenen eigenen Fähigkeiten (Nofle & Robins, 2007) sowie eine englische Vorversion der Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS). In dieser Untersuchung zeigt sich außerdem ein starker Zusammenhang der wahrgenommenen eigenen Fähigkeiten in Bezug auf die Vorhersage der Collegewerten (GPA).

Von einem mediiierenden Effekt wahrgenommener eigener Fähigkeiten auf den Zusammenhang von Gewissenhaftigkeit und Schulleistung kann ausgegangen werden (Nofle & Robins, 2007).

Aufgrund dieser Befunde sollen die wahrgenommenen eigenen Fähigkeiten in dieser Forschungsarbeit als eigener Einflussfaktor in das Vorhersagemodell (s. Abschnitt 2.8.2) mit aufgenommen werden, da anzunehmen ist, dass es zusätzlich zur Varianzaufklärung beitragen und so einen wichtigen Beitrag zur Vorhersage von Schul- und Ausbildungserfolg leisten kann. Nofle und Robins (2007) nahmen an, dass Gewissenhaftigkeit mit höheren Noten assoziiert ist, da diese Eigenschaft höheren Einsatz und höhere wahrgenommene akademische Fähigkeiten mit sich bringt.

Diese Ergebnisse sprechen insgesamt dafür, dass es sich bei den wahrgenommenen eigenen Fähigkeiten, Selbstwirksamkeit und Gewissenhaftigkeit zwar um korrelierte, aber voneinander konzeptuell und empirisch unterscheidbare Konstrukte handelt, die alle entweder direkt oder indirekt Schul-, Ausbildungs- und Berufsleistung vorhersagen. Ein Ziel dieser Studie ist es, die unterschiedlichen Zusammenhänge der Konstrukte in Bezug auf deren Vorhersagekraft hinsichtlich Ausbildungs- und Schulerfolg zu untersuchen.

2.4 Kritische Bewertung von Persönlichkeitsmaßen in der Eignungsdiagnostik

Im Vergleich mit anderen Methoden sind Persönlichkeitstests ein viel diskutiertes Instrument in der beruflichen Eignungsdiagnostik. Je nach Standpunkt ergeben sich zahlreiche Vor- und Nachteile, die für oder gegen eine Verwendung dieser im Rahmen von Personalauswahl- und -entwicklungsmaßnahmen sprechen. Im Folgenden werden daher zuerst die Vorteile (2.4.1) und Nachteile (2.4.2) von Persönlichkeitseigenschaften erläutert. Da im Rahmen dieser Diskussion der Aspekt der sozialen Erwünschtheit besonders stark umstritten und beforscht ist, wird dieser nachfolgend in einem gesonderten Abschnitt (2.4.3) erläutert.

2.4.1 Vorteile

Die Forschung hat, wie bereits in den vorherigen Abschnitten (2.1 und 2.3 ff.) ausgeführt, hinreichend viele Studien und Belege hervorgebracht, die die Nützlichkeit von Persönlich-

keitstests in der beruflichen Eignungsdiagnostik belegen. Verschiedene Gründe sprechen somit für die Verwendung von Persönlichkeitstests in der Eignungsdiagnostik. Diese werden im Folgenden aufgezählt und erläutert.

1. *Gute prädiktive und inkrementelle Validität.* Der Zusammenhang zwischen Persönlichkeitseigenschaften, allen voran Gewissenhaftigkeit, und Arbeitsleistung ist wahrscheinlich am besten erforscht. Wiederholt haben Untersuchungen zeigen können, dass Persönlichkeitsmaße inkrementelle Validität zu kognitiven Tests besitzen, um beispielsweise Arbeitsleistung vorherzusagen (Schmidt & Hunter, 1998), was diese Instrumente für die Praxis der Personalauswahl und -entwicklung interessant macht (Lee et al., 2010).
2. *Getestete Kandidaten haben kein Misserfolgserleben.* Persönlichkeitsmaße haben einen weniger negativen Einfluss auf die Kandidaten als andere Auswahlinstrumente, wie beispielsweise kognitive Tests (Hogan, Hogan & Roberts, 1996).
3. *Sozial gerechte Auswahlmethode.* Auch wenn einige Forscher in der Vergangenheit behauptet haben, dass Persönlichkeitseigenschaften nur ein schwacher Prädiktor beruflicher Erfolgskriterien seien, weisen Persönlichkeitstests weniger soziale Ungerechtigkeit auf als andere Testverfahren bzw. Auswahlmethoden (Risavy & Hausdorf, 2011).
4. *Kostengünstige Verfahren.* Persönlichkeitstests sind günstig in der Durchführung. Sowohl in der Durchführung als auch bezüglich ihrer Auswertung sind diese Verfahren wenig zeitintensiv und mit ein wenig Übung ist letzteres auch schnell erledigt. Der Erwerb der entsprechenden Testlizenzen und Materialien ist ebenfalls vergleichsweise günstig, da diese teilweise mehrfach verwendet werden können, was auch aus ökologischen Gesichtspunkten positiv zu bewerten ist.
5. *Leichte Anwendbarkeit.* Persönlichkeitstests sind leicht anwendbar, sodass auch sorgfältig instruierte Laien diese teilweise administrieren können.
6. *Vielseitige Einsetzbarkeit.* Die meisten Persönlichkeitstests sind darüber hinaus vielseitig einsetzbar. Die gemessenen Merkmale können, insbesondere in Bezug auf die Gewissenhaftigkeitsdimension und ihre Substruktur, für die verschiedensten Stellen und Bewerber angewendet werden. Da es hier um Passungsprozesse geht und die Merkmale der Kandidaten quantitativ und zuverlässig gemessen werden, benötigt man hierfür lediglich eine Analyse, welches Merkmal in welchem Ausmaß für die zu besetzende Stelle von Interesse ist.
7. *Effektive und effiziente Methode der Eignungsdiagnostik.* Es handelt sich bei Persönlichkeitsinventaren um effektive und effiziente Instrumente, die schnell, umfassend und ökonomisch die Eigenschaften von Bewerbern erfassen.
8. *Einfache und schnelle Vergleichbarkeit der Ergebnisse.* Das Vorhandensein von Normwerten stellt außerdem die Vergleichbarkeit der Testwerte sowohl in Bezug auf die Gruppe der Bewerber als auch mit relevanten Vergleichsgruppen her.
9. *Hohe Augenscheinvalidität.* Auch auf Seiten der Wirtschaft werden Persönlichkeitsmerkmale, vor allem Gewissenhaftigkeit, als relevantes und wichtiges Kriterium eingeschätzt, wie eine Praxisanalyse von Moldzio & Partner (2009) zeigen konnte. Dies wurde im Rahmen einer Online-Studie in den Personalabteilungen von 500 führenden deutschen Unternehmen erfragt. Die teilnehmenden Personalr (N= 108) erachteten Gewissenhaftigkeit für alle Bereiche der Personalarbeit als wichtig und relevant. Eine Unterteilung des Merkmals in Aspekte wie *Fleiß* und *Ordnung* macht ihrer Meinung nach auch aus pragmatischen Gesichtspunkten Sinn.

Insgesamt sind Persönlichkeitstests somit relativ einfach in bestehende Auswahlprozesse integrierbar, ohne dass dem auswählenden Unternehmen hierdurch viel Aufwand entstehen würde. Vor allem im Hinblick auf die Mehrerkenntnis, die diese Instrumente mit sich bringen, sind die damit verbundenen Kosten eher zu vernachlässigen.

Auch rechtlich ist es erlaubt und wird es sogar als sinnvoll erachtet, auf Persönlichkeitstests in der beruflichen Eignungsdiagnostik zurückzugreifen. Dies wird in der Praxis der Personalauswahl allerdings teilweise anders wahrgenommen.

Maties und Wottawa (2011) widmeten sich in einem Aufsatz der aktuellen gesetzlichen Regelung zur Erfassung und Weitergabe von Daten aus eignungsdiagnostischen Untersuchungen. Den Autoren zufolge sollte die Grundlage des fachlich und rechtlich korrekten Einsatzes – nicht nur von Persönlichkeitstests, sondern sämtlichen eignungsdiagnostischen Methoden – eine fundierte Anforderungsanalyse sein, die einer Begutachtung vorausgehen sollte. Das Vorgehen hierzu wird in der DIN 33430 beschrieben. Der Norm zufolge stellen Persönlichkeitstests im Prinzip nur eine Art technisch optimierte Prüfung oder Arbeitsprobe dar. Ein Bewerber sollte mit dieser Methode hinsichtlich der optimalen Passung auf das oder die Anforderungskriterien hin überprüft werden.

Das Gesetz zur Regelung des Beschäftigtendatenschutzes (BDSG) beinhaltet auch einen Paragraphen (§ 32) zum Umgang mit Daten aus eignungsdiagnostischen, psychologischen Untersuchungen. Der Hintergrund dieses Gesetzes ist, dass Bewerber bzw. Arbeitnehmer allgemein so wenig wie möglich von sich preisgeben wollen, um eventuellen Nachteilen, die sich hieraus für sie ergeben könnten, vorzubeugen. Dagegen möchte der Arbeitgeber so viel wie möglich über einen Bewerber bzw. Mitarbeiter erfahren. Beide Positionen finden im neu überarbeiteten § 32 BDSG rechtlichen Schutz. Während man sich dem Gesetz zufolge als Bewerber auf sein Recht auf informationelle Selbstbestimmung berufen kann, kann der Arbeitgeber die eigene Berufsfreiheit ins Feld führen. Der Arbeitgeber darf über allgemeine Personendaten hinaus auch weitere personenbezogene Daten erheben, sofern diese notwendig sind, um die Eignung des Bewerbers bzw. Mitarbeiters für eine bestimmte Stelle festzustellen.

Vor einem Eignungstest muss der jeweilige Kandidat dem § 32 BDSG zufolge über Art und Umfang der verwendeten Instrumente aufgeklärt werden und in die Weitergabe der Ergebnisse an den Arbeitgeber einwilligen. Der Eignungstest selbst ist nach den gültigen wissenschaftlich anerkannten Standards durchzuführen. Dem Bewerber sind die Ergebnisse der Eignungsuntersuchung hinterher (ebenfalls) mitzuteilen (Maties & Wottawa, 2011).

In § 32 Abs. 1 BDSG ist darüber hinaus auch ausdrücklich gestattet, dass zu dem Zweck einer beruflichen Eignungsbeurteilung auch persönliche Fähigkeiten erfasst werden dürfen. Dies gilt somit also auch für Persönlichkeitstests. Denn psychologische Eignungstests tun der rechtlichen Auffassung zufolge nicht anderes, als die Fähigkeiten schriftlich zu erfassen, die sonst in einem mündlichen Interview abgefragt würden. Es sollte daher keinen Unterschied machen, ob ein Arbeitgeber seinen Bewerbern geschickte Fragen stellt und daraus auf bestimmte Eigenschaften schließt oder er dies mittels eines Fragebogens macht. Wichtigstes Kriterium ist in jedem Fall die Erforderlichkeit des jeweiligen eignungsdiagnostischen Verfahrens. Der Einsatz von Instrumenten, die auch Aspekte ohne Anforderungsbezug erfassen, ist § 32 BDSG zufolge rechtlich unzulässig.

Der Zweck des § 32 BDSG-E besteht primär darin, Bewerber vor irrationalen und obskuren Tests zu schützen und die Erhebung von zur Beurteilung irrelevanten Daten zu schützen. Werden vom Arbeitgeber jedoch fachlich fundierte Instrumente eingesetzt, die ausschließlich der Erfassung relevanter Eignungsmerkmale dienen, die der auszuübenden Tätigkeit zuzuordnen sind, so ist dies auch nach der neuen Gesetzeslage zulässig (Maties & Wottawa, 2011).

2.4.2 Nachteile

Neben den soeben angeführten Vorteilen der Persönlichkeitstestung im Rahmen eignungsdiagnostischer Untersuchungen gibt es jedoch auch zahlreiche Nachteile, die sowohl in der Wissenschaft als auch in der Praxis viel diskutiert werden und vor allem Praktiker davon ab-

halten, diese Verfahren im Rahmen eignungsdiagnostischer Untersuchungen einzusetzen. Die folgende Auflistung nennt und erläutert die (potenziellen) Nachteile von Persönlichkeitstests in der Berufseignungsdiagnostik.

1. *Geringe Akzeptanz der Verfahren.* Zuerst fällt die geringe Akzeptanz von Persönlichkeitstests sowohl auf Seiten der Bewerber als auch auf Seiten der Personaler ins Auge. Vielfach bestehen auf beiden Seiten große Zweifel hinsichtlich der Aussagekraft und Relevanz von Persönlichkeitstests für den Auswahlprozess bzw. die spätere Tätigkeit. Viele Bewerber haben Angst, dass Persönlichkeitstests ihnen unbekannte und/ oder negative Eigenschaften zu Tage fördern, die womöglich zu einer Ablehnung führen könnten. Vorurteile von „Seelenstrips“, „Durchleuchten“ oder der Offenbarung womöglich intimer Details lassen Bewerber vor der Beantwortung von Persönlichkeitsfragebögen zurückschrecken (von Polier, 2010). Der Praktiker muss – sofern er selbst von der Aussagekraft der Testverfahren überzeugt ist – gegen Voreingenommenheiten sowohl gegenüber der Psychologie im Allgemeinen als auch der Nützlichkeit von Persönlichkeitstests im Kontext der Personalauswahl kämpfen.
2. *Ablehnung der Ergebnisse.* Passt ein Testergebnis nicht zum Selbstbild eines Kandidaten, werden die Testergebnisse häufig als nicht zutreffend abgetan. Eine tiefergehende Beschäftigung mit den Verfahren und den zugrundeliegenden Theorien wird nicht vorgenommen, sondern der Test oder oftmals sogar der gesamte diagnostische Prozess pauschal in Frage gestellt.
3. *Bandwidth-fidelity-Dilemma.* Das Bandwidth-fidelity-Dilemma geht auf Cronbach (1990) zurück und beschreibt den Konflikt zwischen einer möglichst breiten Informationsausbeute und der Detailtiefe der erhobenen Informationen. Die Diskussion ist im Rahmen der Diskussion um das FFM neu entflammt. Einige Autoren kritisieren die Big 5 als zu breit definiert (z.B. Hough & Schneider, 1996). Sie befürworten einen Rückgriff auf engere Eigenschaften. Hilfestellung gibt es hier durch die sog. *Symmetriehypothese*, der zufolge das Abstraktionsniveau von Prädiktor und Kriterium vergleichbar sein muss (Schuler & Höft, 2006). Neuere Forschungsarbeiten zeigen, dass zur Vorhersage spezifischer Leistungskriterien auch spezifische Prädiktoren eingesetzt werden sollten. Für die Praxis bedeutet dies, dass zuerst das zu erfassende bzw. erwünschte Kriterium definiert werden sollte, bevor ein geeignetes Maß zur Erhebung der Prädiktorvariablen ausgewählt wird.
4. *Geringe prädiktive Validität.* Auch wenn viele Untersuchungen die Vorhersagekraft von Persönlichkeitsmerkmalen sowohl in Arbeits- als auch Schulsettings belegen, werden Persönlichkeitsvariablen immer noch wegen ihrer moderaten Validitätskoeffizienten kritisiert (Lievens et al., 2009). Hierfür führen die Autoren verschiedene Ursachen an. So kann eine schlechte Passung von Prädiktor und Kriterium für die eher geringen Zusammenhänge sein. Auch das zuvor angeführte Bandwidth-Fidelity-Dilemma könnte hierfür ursächlich sein. Einschränkungen der Bandbreite der erhobenen Merkmale, wie es die Range Restriction beschreibt, reduzieren ebenfalls die Vorhersagekraft eines Verfahrens – unabhängig von der eigentlichen prädiktiven Validität des Verfahrens (Schmidt et al., 2008).
5. *Falsche Vorinformationen und/ oder Vorurteile gegenüber Persönlichkeitstests.* Oft vermuten Bewerber projektive Verfahren hinter Persönlichkeitstests, die auf Basis der Antwortmuster des Kandidaten auf unbewusste Motive oder Konflikte schließen würden. Die auf Freud basierende Methode wird in der Personalentwicklung allerdings nur noch selten genutzt und ist für den Bereich der Eignungsdiagnostik auch ungeeignet (Backhaus, 2004a). Weniger sachlich formulieren Hesse und Schrader ihre Kritik an den gängigen Persönlichkeitsverfahren. Sie bezeichnen psychologische Testverfahren als „inhumane, gefährliche Testwaffen“ und „sadistische Rituale, die Arbeitnehmer zu Psychosklaven herabzuwürdigen“ (Hesse & Schrader, zitiert nach Backhaus, 2004a). Die beiden Psychologen veröffentlichen regelmäßig Ratgeber für Bewerber, in denen sie diesen Tipps zum „Knacken“ eignungsdiagnostischer Verfahren geben und die rationalen Begründungen für

die Verwendung dieser Verfahren leugnen. Auch differenzieren sie nicht, was einen geeigneten von einem ungeeigneten Verfahren unterscheidet, und tragen so noch mehr zur Skepsis gegenüber psychologischen Testverfahren bei. Die selbsternannten „Testknacker“ raten ihren Klienten, sich möglichst so darzustellen „wie es das Unternehmen wünscht“ und stiften Bewerber so zu sozial erwünschten Antworten (s. Abschnitt 2.3.4) an.

6. *Einsatz falscher diagnostischer Instrumente.* Klinische Tests, wie beispielsweise der MMPI (Minnesota Multiphasic Personality Inventory; Butcher, 2011) werden teilweise auch in der Personalauswahl und -entwicklung eingesetzt, sind aber ebenfalls wie projektive Verfahren nicht für diesen Bereich geeignet, da sie ausschließlich für die Diagnose klinischer Störungen entwickelt wurden. Auch typologisierende Verfahren, die zwar ursprünglich für den Einsatz im Kontext von Beratung und Training entwickelt wurden, sind für die berufliche Eignungsdiagnostik ungeeignet. Graduelle Abstufungen oder mittlere Ausprägungen von Eigenschaften werden durch diese Verfahren nicht berücksichtigt (Backhaus, 2004a).
7. *Einsatz selbstentwickelter, nicht-validierter Testverfahren.* Angeregt durch die augenscheinliche Einfachheit von Persönlichkeits- und auch Leistungstests entwickeln viele Unternehmen eigene Testbatterien, die jeglicher theoretischer Grundlagen entbehren sowie wissenschaftlichen und testtheoretischen Ansprüchen nach Objektivität, Reliabilität und Validität nicht genügen. Die Vorhersagekraft der Ergebnisse solcher „selbstgestrickter“ Verfahren geht gegen null. Wenn man keine wissenschaftlich überprüften Instrumente verwendet, sind auch alle damit erhobenen Ergebnisse fraglich, also wertlos. Außerdem bergen diese Verfahren zudem die Gefahr, dass Bewerber aufgrund negativer Erlebnisse mit diesen dazu tendieren, sämtliche eignungsdiagnostischen Verfahren – auch die wissenschaftlich fundierten – generell abzulehnen. Darüber hinaus werden auch oft veraltete Testverfahren eingesetzt, die jeglicher Konstruktvalidität entbehren und auch nicht neu normiert werden (Anderson & Ones, 2003; Schimmack, 2010). Es ist jedoch notwendig, dass auch kommerziell genutzte Persönlichkeitstests den gängigen Teststandards entsprechen und unabhängigen Überprüfungen ihrer Validität unterworfen werden. Vor allem zur Erforschung berufsbezogener Persönlichkeitstests liegen nur wenige Studien vor (Anderson & Ones, 2003). Demgegenüber fehlt vielen wissenschaftlich entwickelten Verfahren die Datengrundlage, um beispielsweise für die Vorhersage von berufsrelevanten Kriterien als validiert zu gelten (Hülshager & Maier, 2008). Hier herrscht Nachholbedarf, was die Aufklärung der Praktiker sowie die Kooperation von Forschung und Praxis betrifft. So werden Tests bislang noch zu häufig an Gelegenheits- und Zufallsstichproben validiert, wie beispielsweise Schülern und Studenten. Benötigt werden jedoch systematische Daten aus dem Feld, insbesondere auch von älteren Mitarbeitern und Bewerbern (Kersting, 2011).
8. *Verfälschbarkeit von Testergebnissen.* Vor allem Selbstauskünfte im Rahmen von Potenzialanalysen werden als diagnostische Methode aufgrund ihrer potenziellen Verfälschbarkeit stark kritisiert bzw. ihre Aussagefähigkeit sowohl von praktischer als auch von wissenschaftlicher Seite angezweifelt (Schuler & Höft, 2006). Es ist allgemein nicht nachprüfbar, ob ein Kandidat einen Fragebogen wahrheitsgemäß beantwortete oder versuchte, sich in einem anderen, meist positiven Licht darzustellen. Es wird häufig befürchtet, dass Menschen sich v.a. Auswahl-situationen möglichst gut darstellen wollen (Hülshager & Maier, 2008). Dieses Phänomen wird als „sozial erwünschtes Antwortverhalten“ bezeichnet. Der folgende Abschnitt (2.4.3) widmet sich diesem Phänomen ausführlicher.
9. *Keine hinreichenden Diagnosemethoden.* Bei der Auswahl bzw. Anwendung von Persönlichkeitstests sollte man bedenken, dass diese Verfahren allein kein hinreichendes Auswahlkriterium darstellen, sondern lediglich dazu dienen, Hinweise für Verhaltensdispositionen zu liefern. Die Erkenntnisse aus diesen Verfahren sollten in jedem Fall in einen strukturierten Auswahlprozess, im Sinne eines inkrementellen Validitätszuwachses, ein-

gebettet werden (Backhaus, 2004a). Informationen zur Einzelvalidierung sagen darüber hinaus wenig über die mögliche Validitätsverbesserung durch Verfahrenskombinationen aus (Schuler & Höft, 2006). Vor einem kombinierten Verfahrenseinsatz müssen daher beispielsweise mittels multipler Regressionen Aussagen über die genauen Zusammenhänge der Prädiktoren mit einem Kriterium gemacht werden. Vorher signifikante Korrelationen können sich bei diesen Verfahren dann als redundant oder nicht signifikant herausstellen. Dies kann allerdings erst überprüft werden, wenn für ein Verfahren ausreichend Kriteriumsdaten vorliegen und eine Evaluation der Diagnoseinstrumente somit möglich ist. Schmidt und Hunter (1998) schlagen diesbezüglich z.B. vor, zuerst ein Verfahren zur Überprüfung der allgemeinen kognitiven Fähigkeiten einzusetzen und dann die inkrementelle Validität der weiteren Verfahren zu berechnen.

10. *Fehlende Kenntnisse über Persönlichkeitsmerkmale.* Bei der Online-Studie von Moldzio & Partner (2009) wurde auch deutlich, dass mehr als einem Drittel der befragten Personalverantwortlichen das Gewissenhaftigkeitskonstrukt bisher noch nicht bekannt ist oder manche Personalverantwortliche das Konstrukt gar nutzen, obwohl sie über keine differenzierten Kenntnisse darüber verfügen. Auch diesbezüglich scheint es noch Aufklärungsbedarf zu geben.
11. *Vielzahl angebotener Verfahren kann verwirren.* Darüber hinaus kann die Vielzahl der angebotenen Verfahren auf mögliche Anwender verwirrend und abschreckend wirken. Der Laie kann nicht zwischen „guten“ und „schlechten“ Verfahren unterscheiden. Die Gefahr, eines der theoretisch wenig fundierten Verfahren auszuwählen, ist groß. Auch hier besteht weiterhin Aufklärungsbedarf.
12. *Situationsabhängigkeit von Antworten.* Auch wenn Persönlichkeitseigenschaften stabil sind, sind die Antworten auf Items dennoch immer auch situationsabhängig. Je nach persönlicher Verfassung, aktueller Lebenslage sowie den Begleitumständen einer Testung variieren die Antworten einer Person auf Fragebogenitems (Bühner, 2011). Je stärker der Einfluss dieser Begleitumstände, desto mehr können auch Antworten verzerrt sein. Als besonders stark beeinflussende Situationen gelten beispielsweise Auswahlverfahren, in denen das Ergebnis eines Testverfahrens einen erheblichen Einfluss auf die Auswahlentscheidung eines Unternehmens und somit ggf. den weiteren Lebensverlauf einer Person hat. Aus diesem Grund ist es notwendig, Persönlichkeitstests für die Anwendung in eignungsdiagnostischen Untersuchungen anhand von Feldstichproben zu normieren, um vergleichbare und aussagekräftige Ergebnisse zu erhalten.
13. *Widerstand von Betriebsräten.* Der Einsatz von Persönlichkeitstests wird von den Betriebsräten in Unternehmen oft als Eingriff in die Privatsphäre angesehen und es wird befürchtet, dass diese Verfahren die Persönlichkeitsrechte von Bewerbern verletzen (Hülshager & Maier, 2008).
14. *Testschutz.* Die lange Zeit offiziell vertretene Forderung, dass psychologische Tests nur an Psychologen verkauft und von diesen angewendet werden dürfen, führte bei anderen Berufsgruppen zu der Vorstellung, dass beispielsweise auch Persönlichkeitstests etwas sehr „Schwieriges“ und „Geheimnisvolles“ sind (Maties & Wottawa, 2011). Dieser sogenannte „Testschutz“ wurde erst mit den DIN 33430 offiziell beendet. Dennoch halten sich die Vorurteile gegenüber psychologischen Testverfahren weiterhin hartnäckig.

Auch unter Experten wird die Zweckmäßigkeit von Persönlichkeitstests kontrovers diskutiert. Testautoren selbst weisen darauf hin, dass allgemeine Testverfahren eher als grobes Diagnoseinstrument oder „Breitbandverfahren“ zu verstehen sind (Borkenau & Ostendorf, 1993, 2008). Die Testautoren geben an, dass die Verwendung ihres Verfahrens nicht angezeigt ist, wenn Interesse an einer spezifischen und vollständigen Erfassung oder genauen Untersuchung bestimmter Persönlichkeitskonstrukte besteht. Zudem widerspricht es der DIN 33430, Fähigkeiten oder Eigenschaften von Bewerbern zu erfassen, die nichts mit den konkreten Anforderungen einer bestimmten Position zu tun haben. Trotzdem werden solche Tests pauschal in der

Berufseignungsdiagnostik eingesetzt. Ein Hauptgrund hierfür sind vermutlich testkonstruktivistische Beschränkungen, da einzelne Subskalen nicht ausgegliedert werden können bzw. sollen und entsprechende Vergleichsdaten fehlen (Maties & Wottawa, 2011).

Andere Forscher sagen, dass breite, allgemeine Dimensionen, wie z.B. die Big 5, gut zur Theorieentwicklung und als allgemeine Erklärungsansätze geeignet sind, während engere Facetten oder Aspekte (z.B. Fleiß und Ordnung) eher für die Vorhersage spezifischer Verhaltensweisen bei der Arbeit verwendet werden sollten (Bühner, 2011). Enge Traits beinhalten genaue Beschreibungen von beispielsweise Situationen, Zeiten oder Rollen. Durch das konkrete Abfragen von Verhaltensweisen wird somit die prädiktive Validität der Verfahren erhöht. Demgegenüber fokussieren allgemeine Dimensionen eher auf die Erfassung übergreifender Verhaltenstendenzen, was die Vorteile höherer Effektivität (in der Messung), größerer Bandbreite und höherer situationsübergreifender Replizierbarkeit mit sich bringt. Dies erklärt, warum diese Faktoren besser allgemeine Kriterien wie z.B. subjektives Erfolgserleben, aber weniger gut spezifischere Kriterien vorhersagen.

Zudem kommt hier erschwerend hinzu, dass Persönlichkeitseigenschaften in eignungsdiagnostischen Kontexten oftmals nur als Zusatzvariable mit erfasst werden und somit die Zeit, die zur Erfassung der Persönlichkeitseigenschaften aufgewendet wird, häufig stark beschränkt ist. Die herkömmlichen Verfahren zur Erfassung der Big 5 bzw. von Persönlichkeitseigenschaften sind hierfür jedoch oft zu umfangreich (Rammstedt & John, 2005).

Bevor in Abschnitt 2.5 auf berufsbezogenes Testen in der Eignungsdiagnostik eingegangen wird, gibt der folgende Abschnitt (2.3.4) einen Überblick über den aktuellen Status der Diskussion um sozial erwünschtes Antwortverhalten bei der Beantwortung von Selbsteinschätzungsinventaren.

2.4.3 Sozial erwünschtes Antwortverhalten

Die Verfälschbarkeit von Fragebogenwerten durch die Probanden selbst ist einer der häufigsten Kritikpunkte an Maßen, die auf Selbstauskünften basieren. Insbesondere in Situationen, in denen man einen möglichst positiven Eindruck hinterlassen möchte, wie beispielsweise in Situationen der beruflichen Eignungsdiagnostik, besteht die Gefahr, dass Kandidaten sich in einem besseren Licht darzustellen versuchen. Für diese Effekte werden in der Literatur verschiedene Begriffe verwendet, wie beispielsweise *sozial erwünschtes Antwortverhalten* oder *Impression Management* (O'Brien & LaHuis, 2011; Ziegeler, 2011). Im englischen Sprachgebrauch wird sozial erwünschtes Antwortverhalten mit *Social Desirability* oder *Faking* bezeichnet. In dieser Arbeit soll die Tendenz, die eigenen Fähigkeiten in einem positiven Licht darzustellen, als sozial erwünschtes Antwortverhalten oder Faking bezeichnet werden.

Soziale Erwünschtheit kann definiert werden als die absichtliche Verbesserung von erwünschten Verhaltensweisen und Eigenschaften sowie die Unterdrückung unerwünschter Charakteristika, um sich einem potenziellen Arbeitgeber in einem möglichst positiven Licht zu zeigen. „Motiviert durch die Furcht vor sozialer Verurteilung neigt man zu konformen Verhaltensweisen und orientiert sich in seinen Verhaltensäußerungen strikt an verbreiteten Normen und Erwartungen“ (Bortz & Döring, 1995, S. 212). Das Ausmaß sozial erwünschten Antwortverhaltens ist dabei abhängig von der Quelle der Erwartung. So gibt es einerseits internalisierte Normen, die zu angepasstem Antwortverhalten führen. Andererseits haben auch situative Faktoren einen Einfluss auf den Grad sozial erwünschten Antwortverhaltens. Im letzten Fall kommt es durch die in der Befragungssituation aktiven Stimuli zu einer Generierung von Erwartungen, an die das Antwortverhalten angepasst wird (Ziegeler, 2011). Starke Hinweisreize werden grundsätzlich durch Auswahl-situationen erzeugt.

Renner (2011) besagt, dass jeder Mensch mehr oder weniger stark dazu tendiere, sich selbst bzw. die eigenen Fähigkeiten und Verdienste positiv zu inszenieren. Erfolgreiche Selbstdarstellung fördere den eigenen Selbstwert, indem sie positive Reaktionen, wie z.B. Lob, Anerkennung, Respekt, bei anderen Menschen hervorruft. Die Vermutung, Faking diene der Sicherung eines positiven, privaten Selbstbildes liegt daher nahe (Kanning, 2011). Auch Schmit et al. (1995) berufen sich auf die Selbstpräsentationstheorie von Hogan (1982), die besagt, dass Menschen Persönlichkeitsinventare nicht ehrlich bzw. realistisch beantworten, da sie sich auch immer so beurteilen, wie sie (hoffen) von anderen eingeschätzt zu werden. Einige Menschen verfügen demzufolge über realistischere Selbstbilder, was ihre Einschätzung durch andere betrifft. Andere sind demgegenüber weniger gut in der realistischen Einschätzung eigener Stärken und Schwächen. Allgemeine Persönlichkeitstests, die in der Eignungsdiagnostik verwendet werden, sind den Autoren zufolge anfälliger für Fehler in der Selbstpräsentation, da bei vielen Items nicht klar sei, ob und auf welchen Arbeitskontext oder welche Arbeitsrolle sich diese beziehen. Die Selbstpräsentation von Bewerbern könnte den Autoren zufolge durch eine arbeitsbezogene Formulierung der Items vereinfacht bzw. verringert werden.

Persönlichkeitsmaße sind nachweislich nützliche Auswahlinstrumente, dennoch haben sowohl Praktiker als auch Wissenschaftler Bedenken geäußert, dass diese Tests von Kandidaten grundsätzlich verfälscht werden können. Wenn sich jemand auf eine bestimmte Stelle bewirbt, ist diese Person in der Regel motiviert, diese auch zu bekommen und wird daher zeigen wollen, dass er gut für die zu besetzende Position geeignet ist. Da Persönlichkeitsmaße transparenter sind als andere Auswahlinstrumente, haben die Kandidaten bei diesen die Möglichkeit, sich selbst in einer unkorrekten Art und Weise zu präsentieren, die scheinbar positiver ist. Vor allem diese subjektiv als leicht empfundene Möglichkeit, Persönlichkeitstests verfälschen zu können, hat dazu beigetragen, dass diese Verfahren sowohl auf Seiten der Praktiker als auch seitens der Forscher im Kontext der Personalauswahl stark umstritten und teilweise wenig akzeptiert sind. Zwar hat sich Transparenz in der Vergangenheit als bester Prädiktor für die Akzeptanz eines psychologisch-diagnostischen Verfahrens erwiesen. Die beschriebene leichte Durchschaubarkeit von Verfahren zur Erfassung von typischem Verhalten geht dagegen häufig mit einer hohen Verfälschbarkeit einher, was auch als *Akzeptanz-Verfälschbarkeits-Dilemma* bezeichnet wird (Ortner et al., 2007).

Fast jeder psychologische Laie ist davon überzeugt, dass Bewerber grundsätzlich lügen und insbesondere Fragebogenverfahren kein zutreffendes Bild einer Person vermitteln können (Kanning, 2011; Ortner et al., 2007). Viele Forschungsarbeiten deuten darauf hin, dass Bewerber nicht nur ihre Antworten verfälschen, sondern dies auch im groß angelegten Stil tun, vor allem, wenn sie sich dadurch einen Vorteil erhoffen (Bühner, 2011). Forscher sprechen von 30 % bis 50 % verfälschten Testwerten (Lee et al., 2010). Sozial erwünschtes Antwortverhalten hat konkret zur Folge, dass beispielsweise der Mittelwert eines Persönlichkeitsmerkmals wie Gewissenhaftigkeit in einer Bewerbergruppe gegenüber einer neutralen Probandengruppe erhöht ist. Metaanalysen zufolge liegt dieser Effekt zwischen einer halben und einer ganzen Standardabweichung (Bühner, 2011; Hülshager & Maier, 2008; Kanning, 2011). Würden alle Personen ihre Testwerte gleichermaßen stark verzerren, würde es sich lediglich um eine Mittelwertsverschiebung (nach oben) handeln, was zu vernachlässigen wäre. Da dies jedoch nicht der Fall ist, da Personen in unterschiedlich starkem Ausmaß dazu neigen, sich sozial erwünscht zu präsentieren und dies darüber hinaus situationsabhängig ist, könnte sich dieser Effekt auch auf die psychometrischen Eigenschaften wie beispielsweise die Reliabilität oder Validität eines Tests auswirken (Bühner, 2011; Lee et al., 2010). Für die Praxis der Personalauswahl bedeutet dies, dass es möglich ist, dass sozial erwünschtes Antwortverhalten die Auswahlentscheidung eines Unternehmens in dem Sinne beeinflusst, dass nicht diejenige Person eingestellt wird, die die größte Passung auf die Anforderungen aufweist, sondern die Per-

son, die am besten die Erwartungen des Unternehmens antizipieren und ihre Antworten bzw. ihr Verhalten an diese anpassen kann. Es würde somit die anpassungsfähigste und nicht die am besten passende Person eingestellt. Das Ergebnis dieser Unterschiede in der Selbstbeschreibung kann also dazu führen, dass Unternehmen Einstellungsentscheidungen treffen, die nicht mit den Unternehmenszielen übereinstimmen (O'Brien & LaHuis, 2011). Sozial erwünschtes Antwortverhalten verändert Kritikern zufolge dabei allerdings nicht die durchschnittliche Vorhersagegüte einer Skala, sondern die individuelle Eigenschaftsausprägung einer Person, weshalb es aufgrund von Faking zu Rangreihenveränderungen unter den Bewerbern kommen kann. Dieser Effekt tritt umso stärker zutage, je weniger Personen selektiert werden (Bühner, 2011; Hülshager & Maier, 2008).

Crowne und Marlowe (1960) entwickelten die erste Skala zur Erfassung sozial erwünschten Antwortverhaltens (Crowne Marlowe Social Desirability Scale; CMSDS). Die 33 Items dieser Skala erfassen sozial erwünschtes Verhalten, das eher selten ist, sowie sozial unerwünschtes Verhalten, das eher häufig vorkommt. Die CMSDS wurde die weltweit am häufigsten verwendete Skala zur Erfassung sozial erwünschter Antworttendenzen und fungiert im Rahmen psychologischer Messungen als „Standardinstrument“ zur Erfassung dieser (Stöber, 2001). Stöber (1999) kritisierte allerdings die mittlerweile veralteten Items der Skala. Diese spiegelten dem Autor zufolge soziale Standards der 1960er wider, die nicht mehr auf heute geltende Normen bzw. Werte anwendbar und somit nicht mehr für das Konzept der sozialen Erwünschtheit repräsentativ seien. Analog der Konstruktion der CMSDS erstellte Stöber (1999) eine aktualisierte Version der Skala zur Erfassung sozial erwünschten Antwortverhaltens (Social Desirability Scale-17; SDS-17), die auch in der vorliegenden Studie verwendet werden soll.

Um die Sensitivität seines Instruments zur Erfassung von Faking in Situationen mit erhöhtem Potenzial für sozial erwünschtes Antwortverhalten zu überprüfen, verglich Stöber (2001) in einer Laborstudie zwei Subgruppen ($N_{ges} = 76$), die die SDS-17 zusammen mit anderen Fragebogenverfahren bearbeiteten. Die eine Hälfte der Teilnehmer wurde instruiert, sich vorzustellen, dass sie Bewerber in einem Personalauswahlverfahren und ihre Fragebogenantworten für die Bewerber sehr wichtig seien. Die andere Hälfte der Studienteilnehmer bearbeitete die Fragebogenverfahren unter vergleichbaren Bedingungen, nur ohne die Bewerber-Instruktion. Die Ergebnisse der Studie zeigen, dass die Probanden, die sich vorstellen sollten, sich in einer Personalauswahlsituation bewähren zu müssen, signifikant höhere Werte in der SDS-17 aufwiesen als die Probanden der neutralen Kontrollgruppe. In diesem Zusammenhang wurde auch der Einfluss der Persönlichkeitsdimension Gewissenhaftigkeit untersucht. Auch hier zeigten sich für die Gruppe mit der Instruktion „Personalauswahl“ höhere Werte für die Persönlichkeitsdimension. Dieser Befund ist konsistent zu vielen Laborstudien, die von höheren, potenziell sozial erwünschteren Werten für Persönlichkeitsmaße im Kontext von Leistungssituationen berichten.

Der Einfluss sozial erwünschten Antwortverhaltens wurde auf die eben geschilderte Art und Weise in vielen Untersuchungen überprüft. Oftmals wurden Laborstudien durchgeführt, in denen sich Studenten vorstellen sollten, dass sie sich in einer Auswahl-situation befinden und sich möglichst gut darstellen müssen, um einen Job zu bekommen. Danach wurden die Studenten aufgefordert, verschiedene Selbstbeschreibungsinventare zu bearbeiten. Die Ergebnisse aus diesen Studien deuten an, dass Personen in Bewerbungssituationen sich selbst höhere Eigenschaftsausprägungen zuschreiben als in neutralen Situationen. Diese Befunde wurden als Beweis für den negativen Einfluss sozial erwünschten Antwortverhaltens genommen und pauschal aufgrund der leichten Verfälschbarkeit von Persönlichkeitstests von deren Verwendung abgeraten. Diese Schlussfolgerung ist jedoch übereilt, da in eben diesen Untersuchungen ein negativer Einfluss der verzerrten Selbsteinschätzungen auf die spätere Leistung

im Beruf selbst ausblieb. Es blieb meist unklar, inwiefern die prognostische Validität der Verfahren durch das sozial erwünschte Antwortverhalten verringert wurde.

In einer Studie zur Ermittlung der Konstrukt- und Kriteriumsvalidität verglichen Blickle, Schnitzler, von Below und Johannes (2011) die Validität eines Selbsteinschätzungsinstrumentes mit einem Maß zur Erfassung der Big 5 sowie einer Vorgesetztenbeurteilung. Hierzu verglichen sie Berufstätige ($N=207$), die sie zwei Bedingungen zuordneten: einer Bewerbungssituation und einer „ehrlichen“ Situation. Die Ergebnisse der Studie besagen, dass auch in Auswahl-situationen eine kriteriumsbezogene Validität vorliegt. Die Autoren empfehlen daher Persönlichkeitstests in der Praxis der Personalauswahl einzusetzen. Diese Ergebnisse sind kongruent zu den Befunden von Hogan, Barrett und Hogan (2007), die anhand einer großen aggregierten Bewerberstichprobe ($N=250.000$) Faking als „Scheinproblem“ identifizierten. Von den sechs Monaten zuvor abgelehnten Bewerbern dieser Stichprobe bewarben sich über 5.000 Bewerber erneut bei dem Unternehmen. Den Ergebnissen von Hogan et al. (2007) zufolge zeigten 96,4 % der „Wiederholer“ kein signifikant besseres Testergebnis bei der zweiten Bewerbung. Lediglich 3,6 % der wiederkehrenden Bewerber wiesen höhere Testergebnisse auf. Diese verfügten jedoch gleichzeitig auch über deutlich bessere soziale Fertigkeiten und höhere Integritätswerte. Die Forscher schlussfolgerten, dass Personen mit hohen sozialen Fertigkeiten sich gemäß den Annahmen der sozioanalytischen Theorie sehr situationsensibel und adaptiv verhalten und aus diesem Grund ihre Testwerte den angenommenen Situationsanforderungen anzupassen versuchten.

Denkbar ist auch, dass in einer realen Auswahl-situation die implizite Annahme seitens der Bewerber, die eben gemachten Aussagen würden später überprüft, zu ehrlichen Antworten führt (Schuler & Höft, 2006).

Eine Metaanalyse von Birkeland, Manson, Kisamore, Brannick und Smith (2006) zeigte, dass Bewerber ihre Testwerte in der Gewissenhaftkeitsdimension im Vergleich mit den anderen Big 5 am meisten nach oben korrigierten. Die Autoren erklärten diesen Effekt damit, dass Bewerber diesen Faktor als die wichtigste Dimension für die (potenziellen) Arbeitgeber erkennen und aus diesem Grund ihre Antworten in dieser Dimension besonders stark verfälschen bzw. an die Erwartungen anpassen. Die tatsächliche Verfälschungstendenzen von Bewerbern in Auswahl-situationen fällt ihren Ergebnissen zufolge jedoch wesentlich geringer aus.

Auch Stöber (2001) konnte eine signifikante Korrelation von sozialer Erwünschtheit und Gewissenhaftigkeit nachweisen ($r=.38, p<.01$). Der Autor stellt diesbezüglich die Frage auf, ob diese Zusammenhänge als konvergente oder diskriminante Validität interpretiert werden könnten. So sei Gewissenhaftigkeit einerseits ein Teil sozialer Erwünschtheit, andererseits erfassten Maße wie die SDS-17 eine sogenannte Bias, d.h. eine einseitige Neigung, die unabhängig von der wahren Gewissenhaftigkeit einer Person dazu führt, Fragebögen in Richtung sozial erwünschter Normen und Wertvorstellungen zu verzerren. Da es hierzu jedoch bisher meist nur korrelative Daten gibt, ist weiterhin unklar, worauf diese Antwortmuster letztlich zurückzuführen sind.

Verschiedene Forscher führen die Zusammenhänge zwischen sozialer Erwünschtheit und Gewissenhaftigkeit auch auf die in der Sprache allgemein enthaltenen bewertenden Implikationen zurück (Borkenau & Ostendorf, 1993, 2008; Stöber, 2001). So lassen sich kulturübergreifend in verschiedenen Sprachen vor allem bewertende Eigenschaftsadjektive finden. Nicht-wertende Adjektive sind hingegen kaum anzufinden. Dies könnte darauf zurückzuführen sein, dass insbesondere solche Begriffe Eingang in die Sprache gefunden haben, die nützliche Hinweise darauf liefern, ob man mit einer Person in Kontakt treten möchte oder nicht.

Lee et al. (2010) untersuchten anhand einer Studentenstichprobe ($N = 694$ College-Studenten) den Zusammenhang von Gewissenhaftigkeit und Selbstwirksamkeit in Bezug auf ihre Tendenz, sozial erwünscht zu antworten. Die Studienteilnehmer beantworteten hierzu u.a. den NEO-FFI und eine Selbstwirksamkeitsskala. Die Untersuchung fand zu zwei Zeitpunkten statt. Während der ersten Testung wurde den Studenten erzählt, sie könnten sich bei einem Unternehmen bewerben. Bei der zweiten Testung wurden die Studenten über den wahren Zweck der Untersuchung aufgeklärt. Lee und Kollegen (2010) klassifizierten Teilnehmer, die eine signifikant höhere Ausprägung auf der Gewissenhaftigkeitsdimension aufwiesen als „Faker“. Die Ergebnisse der Studie zeigen, dass das Ausmaß, indem Faker ihren Gewissenhaftigkeitswert veränderten, signifikant mit Selbstwirksamkeit korrelierte. Das heißt, Personen, die ihre Testwerte verfälschten, wiesen eine geringere Selbstwirksamkeit auf als ehrliche Kandidaten. Eine mögliche Erklärung für diesen Effekt ist den Autoren zufolge, dass Menschen mit nur gering ausgeprägter Selbstwirksamkeitserwartung nicht daran glauben, dass sie eine gewünschte Position allein aufgrund ihrer (ehrlichen) Eigenschaften erhalten (können) und daher ihre Werte in eine sozial erwünschte Richtung verzerren.

Die Ergebnisse der Feldstudie von Sutherland et al. (2007) zeigten eine signifikante Korrelation von Gewissenhaftigkeit mit sozial erwünschtem Antwortverhalten, jedoch keine Zusammenhänge zwischen Gewissenhaftigkeit und Arbeitsleistung. Die Forscher schlussfolgerten, dass Faking als medierende Variable agierte und hoch gewissenhafte Menschen bei der Arbeit allgemein weniger kontraproduktives Verhalten zeigen würden.

König, Jansen und Lüscher Mathieu (2011) überprüften, inwiefern Bewerber dazu tendieren, Persönlichkeitstests bzw. deren Ergebnisse zu fälschen, wenn sie wissen, wie diese ausgewertet werden. Hintergrund dieser Studie war die Beobachtung, dass viele Bewerber teilweise sehr komplexe Annahmen über Persönlichkeitstests haben. Die Forscher fragten sich, inwiefern Faken forciert wird, wenn Bewerber wissen, wie einfach Persönlichkeitstests ausgewertet werden. Hierzu untersuchten sie 187 Personen, die berufsbegleitend an verschiedenen Fachschulen für Wirtschaft in der Schweiz studierten. Sie ließen die Probanden die deutsche Version des NEO-FFI (Borkenau & Ostendorf, 2008) sowie eine Skala zum Impression Management und einen Test zur Erfassung kognitiver Fähigkeiten bearbeiten. Die Studienteilnehmer beantworteten die Fragebogen zuerst in einer „Bewerber“-Bedingung und danach noch einmal unter einer "Ehrlich"-Instruktion. Sie wurden randomisiert einer der beiden Gruppen zugewiesen, die entweder mit oder ohne Informationen zur Auswertung von Persönlichkeitstests versehen wurden. Die Ergebnisse deuten einen kleinen Effekt dahingehend an, dass informierte Bewerber ihre Werte etwas sozial erwünscht verfälschten, jedoch nicht stark. Die Autoren folgerten daher, dass Bewerber ihre Testwerte mehr faken würden, wenn sie wüssten, wie einfach Persönlichkeitstests auszuwerten sind. Da dies bei den meisten Bewerbern jedoch nicht der Fall ist, spricht dies für die (weitere) Verwendung von Selbstbeschreibungen in der Praxis der Eignungsdiagnostik.

Zwar wird das Konstrukt der sozialen Erwünschtheit seit mehr als 50 Jahren beforscht, allerdings herrscht immer noch nur wenig in diesem Bereich und viele Befunde sind widersprüchlich (Kanning, 2011; Lee et al., 2010). Es ist nach wie vor umstritten, inwiefern durch die Verfälschbarkeit von Selbstbeschreibungsinventaren die Aussagekraft von Persönlichkeitstests überhaupt herabgesetzt ist (Hülsheger & Maier, 2008). So gibt es auch eine Reihe von Forschern, die besagen, dass sozial erwünschtes Antwortverhalten ein Zeichen der Anpassungsfähigkeit eines Menschen ist und daher kein Grund zur Sorge sein sollte (z.B. Marcus, 2003; Viswesvaran, Deller & Ones, 2007). Marcus (2003) nimmt beispielsweise an, dass – auch im Hinblick auf verschiedene metaanalytische Ergebnisse in Bezug auf die Kriteriumsvalidität sozial erwünschter bzw. verfälschter Fragebogenergebnisse – die Fähigkeit zu situationsangemessenem Verfälschen eventuell sogar selbst ein guter Prädiktor für bestimmte spätere berufliche Leistungen sein könnte. Es gibt Hinweise, dass es eine gewisse Intelligenz

erfordert, das Konstruktionsprinzip eines Tests und dessen genauen Ziele zu erfassen, um einen Fragebogen adäquat verfälschen zu können (Bing, Whanger, Davison & VanHook, 2004). Insofern könnten sozial erwünschte Antworttendenzen als Indikator für die Intelligenz einer Person dienen und folglich den Zusammenhang der Eigenschaften mit Berufserfolg eventuell sogar noch erhöhen (Borkenau & Ostendorf, 2008).

Insgesamt wurde bisher vor allem allerdings nur auf der Dimensionsebene zum Zusammenhang zwischen Gewissenhaftigkeit und sozial erwünschtem Antwortverhalten geforscht. Nur wenige Studien beschäftigten sich bisher mit der Interaktion der untergeordneten Persönlichkeitsfacetten bzw. -aspekten und deren Zusammenhang mit sozial erwünschtem Antwortverhalten (Ziegler, 2011; Ziegler, Dreier & Moldzio, 2010).

Ziegler et al. (2010) vermuteten, dass sozial erwünschtes Antwortverhalten unterschiedlich stark mit den einzelnen Facetten der Big 5-Dimensionen zusammenhängen kann – je nach Anforderung einer Situation. Dies sei auch der Grund dafür, dass die Befundlage zum Zusammenhang von sozial erwünschtem Antwortverhalten und der Persönlichkeit im Kontext der Berufseignungsdiagnostik noch immer umstritten ist. Sie untersuchten die Ausprägungen von Psychologiestudenten im Grundstudium ($N = 145$) mittels NEO-PI-R zur Vorhersage von Studienerfolg. Die Probanden bearbeiteten das NEO-PI-R erst einmal neutral und danach mit einer Faking-Instruktion. Hinterher verglichen die Forscher die intraindividuellen Mittelwerte der Probanden. Die Autoren konnten einen Einfluss von Faking nur auf der Facettenebene, nicht aber auf der Dimensionsebene finden. Die signifikanten bzw. moderaten Testwertverschiebungen zeigten sich auf den meisten Facetten, allerdings waren nicht alle Verschiebungen auf den Facetten einer Dimension in die gleiche Richtung. Faking beeinflusste zwar die Selbsteinschätzungen der Studenten, hatte allerdings keinen Einfluss auf die Validität der Persönlichkeitsdimensionen. Für die Vorhersagekraft der Facetten waren die Ergebnisse allerdings weniger eindeutig. Ein weiteres Ergebnis der Studie war, dass eine Tendenz zu sozial erwünschtem Antworten die interindividuellen Unterschiede zwischen den Studenten verringerte. Daraus kann gefolgert werden, dass die meisten Studenten ein einheitliches Konzept des „idealen Bewerbers“ hatten, was somit die Abnahme der Varianz auf den Facetten verursacht haben könnte.

In einer Felduntersuchung verglich Ziegler (2011) eine Bewerberstichprobe mit einer Schülergruppe und fand einen Einfluss von sozial erwünschtem Antwortverhalten in Bezug auf die Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie Selbstwirksamkeit. Allerdings handelte es sich hierbei lediglich um Mittelwertsvergleiche. Eine Untersuchung der Unterschiede in Bezug auf die prognostische Validität der verfälschten Testwerte blieb aus.

Aufgrund der mehrdeutigen Befundlage zur Existenz des Einflusses von sozial erwünschtem Antwortverhalten, haben Forscher in zahlreichen Studien Faking häufig als Kontrollvariable mit erfasst. Dies ist vor allem darauf zurückzuführen, dass die Wissenschaftler nachweisen möchten, dass die von ihnen gefundenen Ergebnisse nicht auf sozial erwünschtes Antwortverhalten zurückzuführen sind (Stöber, 1999). Das Konstrukt dient somit vorrangig als Ausschlussverfahren möglicher konfundierender Effekte.

Zusammenfassend lässt sich sagen, dass die Befundlage zur sozialen Erwünschtheit weiterhin ungeklärt ist. Auch wenn Laborstudien gezeigt haben, dass Faking zu Testwertveränderungen führen kann, so ist die ökologische Validität bzw. die Generalisierbarkeit dieser Ergebnisse auf die Praxis der Personalauswahl fraglich und bisher nicht erwiesen (Marcus, 2003). Die Tatsache, dass Persönlichkeitseigenschaften – auch wenn sie in Situationen erhoben wurden, die sozial erwünschtes Antwortverhalten begünstigen – deutlich mit den interessierenden Kriteriumswerten korreliert sind, zeigt, dass der Einsatz von Persönlichkeitsmaßen in organisationalen Settings sinnvoll ist. Man sollte Fragebogenverfahren nicht als eignungsdiagnostische Methode verbannen, sondern diese vielmehr zu einem frühen Zeitpunkt der Auswahl –

vor allem bei hohen Bewerberzahlen – zur Vorselektion heranziehen. Hierfür könnte ein eher niedriger Cut-off-Wert gewählt werden, sodass potenzielle Rangplatzvertauschungen weniger stark ins Gewicht fallen (Bühner, 2011). Weitere Forschung ist notwendig, um im Anwendungskontext den Einfluss von sozial erwünschtem Antwortverhalten auf die Leistungsvorhersage von Mitarbeitern aufklären zu können. Dies soll auch in der vorliegenden Untersuchung geschehen.

2.5 Berufsbezogenes Testen in der Eignungsdiagnostik

Im vorigen Abschnitt (2.4) wurde auf Vor- und Nachteile von Persönlichkeitstests in der beruflichen Eignungsdiagnostik eingegangen. Zwei der genannten Probleme dieser Verfahren, die vermutlich die geringe Verwendung von Persönlichkeitstests in der Praxis der Personalauswahl und -entwicklung begründen, sind die geringe Akzeptanz sowie die geringe Validität bzw. als gering wahrgenommene Vorhersagekraft von Persönlichkeitsmerkmalen. Entsprechend forderten De Fruyt und Salgado (2003) mehr Anwendungsbezug in der Arbeits- und Organisationspsychologie. Eine Möglichkeit, einen solchen Anwendungsbezug für den Bereich der Persönlichkeitstestung herzustellen bzw. zu vertiefen, ist die Entwicklung und Verwendung von arbeitsbezogenen Persönlichkeitsmaßen.

Arbeitsbezogene Persönlichkeitsmaße erfassen Traits, die für Arbeitsverhalten als wichtig erachtet werden. In der Regel enthalten sie arbeitsbezogen formulierte Instruktionen und Items (Reddock, Biderman & Nguyen, 2010; Reddock et al., 2011). De Fruyt und Salgado (2003) beschreiben diese als „*inventories that are specifically designed to assess traits operating in the work place*“ (S. 125). Beispielsweise können allgemeinen bzw. *generischen* Items arbeitsbezogene Formulierungen hinzugefügt werden (z.B. „Bei der Arbeit...“) oder die Testpersonen in den Instruktionen angewiesen werden, einen Fragebogen zu beantworten, indem sie sich an eine bestimmte Situation erinnern oder hineinversetzen.

Die meisten verfügbaren Persönlichkeitstests sind jedoch aus Items zusammengesetzt, die sich auf allgemeine Verhaltenstendenzen, Einstellungen, Beziehungen, Präferenzen und soziale Fertigkeiten beziehen, die über Situationen hinweg generalisiert werden können. Die Grundannahme ist hier, dass Personen Items beantworten, indem sie an ihre allgemeinen Verhaltensweisen, Gefühle, Beziehungen usw. denken – oder zumindest daran, wie sie diese selbst bei sich wahrnehmen oder gern wahrnehmen würden. Es ist jedoch wahrscheinlich, dass dies vor allem bei Bewerbern in Auswahl Situationen nicht der Fall ist. Es ist möglich, dass diese einen arbeitsbezogenen *Referenzrahmen* bei der Beantwortung von Persönlichkeitsfragebögen bzw. -items verwenden, weil sie davon ausgehen, dass sich die Items auf den Arbeitskontext beziehen bzw. dies Situationen sind, die den (potenziellen) Arbeitgeber interessieren. Andere Bewerber wiederum denken bei der Itembeantwortung vielleicht an Situationen aus ihrer Freizeit oder dem Familienleben. Aus diesem Grund kann es zu Unterschieden bei der Beantwortung von allgemein gehaltenen Fragebogenitems kommen (Schmit et al., 1995).

Die Verwendung arbeitsbezogener Persönlichkeitstests in der Personalauswahl wurde von verschiedenen Forschern empfohlen (Bing et al., 2004; Lievens et al., 2008; Reddock et al., 2010; Schmit et al., 1995), da z.B. angenommen wurde, dass bei Bewerbern, die Items arbeitsbezogen beantworten, bessere Vorhersagewerte erzielt werden könnten, da Arbeitsverhalten ein besserer Indikator für Berufserfolg ist als allgemeines Verhalten. Die zugrundeliegende Logik ist, dass eine höhere (inhaltliche) Übereinstimmung von Prädiktor und Kriterium zu einer höheren prädiktiven Validität führen sollte. Auch eine Erhöhung der Reliabilität aufgrund der Reduzierung der interpersonellen Antwortinkonsistenz wurde vermutet (Lievens et al., 2008). Ein Referenzrahmen bzw. *Frame-of-Reference* (FOR) standardisiert somit die Iteminterpretation und reduziert auf diese Weise den Messfehler, während er gleichzeitig die Validität im Vergleich zu nicht kontextualisierten Items erhöht (Bing et al., 2004).

Auch Barrick und Mount (2004) empfahlen in ihrer Metaanalyse, der situationalen Komponente mehr Beachtung zu schenken, um vor allem auch den Zusammenhang von Gewissenhaftigkeit und Arbeitsleistung bzw. -erfolg weiter aufklären zu können, da die aktuelle Situation von Bewerbern einen Einfluss auf die Ausprägung von Persönlichkeitsmerkmalen habe bzw. diese moderieren könne. Die Autoren vermuteten ein reziprokes Verhältnis, in dem Sinne, dass Persönlichkeitseigenschaften einerseits beeinflussen würden, welche Situationen eine Person aufsucht und wie sie diese wahrnimmt, andererseits diese Situationen wiederum als Referenzrahmen für die Beantwortung von Selbstbeschreibungen dienen würden.

Darüber hinaus wird vermutet, dass mit kontextualisierten Items die Akzeptanz von Persönlichkeitstests sowohl auf Seiten der Bewerber als auch auf Seiten der Personaler erhöht werden kann, da diese Items augenscheinlich berufsrelevant sind.

Die Idee, dass Verhalten nicht nur abhängig ist von stabilen Dispositionen und intrapersonellen Vorgängen, sondern auch in Abhängigkeit von Situationen variiert, hat u.a. als „Person-Umwelt-Debatte“ eine lange Tradition in der Psychologie. Bereits Allport (1937) und Lewin (1936) argumentierten, dass Verhalten eine Funktion einer Interaktion zwischen einer Person und der sie umgebenden Situation ist. Kurt Lewin (1936) nahm an, dass das Verhalten (V) eines Menschen eine Funktion aus Merkmalen von Person (P) und Umwelt (U) ist, die sich wechselseitig bedingen:

$$V = f(P, U)$$

Allerdings ist es schwer, stabile Interaktionen aufzuzeigen und zu replizieren. Funder (2006) gab der Interaktion aus Persönlichkeitseigenschaften, Umgebungsfaktoren und gezeigtem Verhalten einen neuen Rahmen, indem er in seiner Persönlichkeitstriade die Elemente Person, Situation und Verhalten zueinander wie folgt in Beziehung setzt:

Situation = Verhalten x Person

Person = Verhalten x Situation

Verhalten = Situation x Person

Der Autor zeigte in einem Experiment, dass eine objektiv gleiche Situation (= man sollte sich mit jemandem Unbekanntem unterhalten) auf der psychologischen Ebene zu völlig verschiedenen Verhaltensweisen führen kann. In einer für die Probanden unbekannten Umgebung begegneten die Probanden der unbekannten Person eher reserviert und zurückhaltend, während die gleichen Probanden in einer bekannten Umgebung sich der unbekannten Person gegenüber entspannt und locker verhielten. Ähnliche Effekte sollten dem Autor zufolge in Bewerbungssituationen wirken. In der eher unbekannten Situation, z.B. während eines Assessment-Centers, sollte sich ein Bewerber anders gegenüber den Mitbewerbern und Beobachtern verhalten, als ein Mitarbeiter, der eine Beobachtungsfunktion in diesem Verfahren übernimmt.

Nicht jede Situation wirkt sich somit gleichermaßen auf die Ausführung bestimmter Verhaltensweisen aus. Lord (2011) unterscheidet zudem zwischen *starken* und *schwachen Situationen*. Sie definiert starke Situationen als solche, in denen es eindeutige Hinweisreize aus der Umgebung für angemessenes Verhalten gibt. In starken Situationen verhalten sich der Autorin zufolge Menschen gleich trotz individueller Unterschiede in ihren Persönlichkeitseigenschaften, die für die Situation relevant sind. Wenn es starke Hinweisreize gibt, sich auf eine bestimmte Weise zu verhalten, werden individuelle Persönlichkeitsunterschiede demzufolge eher nivelliert. Hierzu zählen beispielsweise Bewerbungssituationen. Demgegenüber haben Persönlichkeitseigenschaften in schwachen Situationen eine größere Vorhersagekraft, da in diesen der Situationsdruck, sich auf eine bestimmte Weise zu verhalten, gering ist.

Die unterschiedliche Stärke verschiedener Situationen wird von Funder (2006) u.a. nicht beachtet. So ist die von dem Autor postulierte Persönlichkeitstriade eine zu starke Vereinfachung.

chung der Zusammenhänge, die beispielsweise weitere Interaktionsmöglichkeiten sowie medierende, moderierende oder andere Einflüsse unberücksichtigt lässt. Bereits Mischel (1968) postulierte, dass jede objektive Stimulussituation eine Vielzahl von Effekten hervorrufen könnte, die je nach individuellem Wahrnehmen unterschiedlich ausfallen. Daher müsste man jede Person um eine persönliche Bewertung einer interessierenden Situation bitten, um interindividuell vergleichbare Ergebnisse zu erlangen. Gemeinsam mit seinen Kollegen zeigte er als erster Forscher überhaupt auf, dass Menschen allgemein eher stabile Verhaltensmuster zeigen, diese jedoch in Abhängigkeit von den sie umgebenden situationalen Bedingungen variieren. Aus diesem Grund führten Wright und Mischel (1987) das Konzept der *konditionalen Verhaltensdispositionen* ein.

Die Theorie der konditionalen Verhaltensdispositionen (Wright & Mischel, 1987) besagt, dass die Manifestation eines Persönlichkeitsmerkmals situationsabhängig ist. Das heißt, je nach Kontext beantwortet eine Person ein Item in der einen oder anderen Weise, also unterschiedlich. Individuen verhalten sich also innerhalb gleicher Situationen ähnlich und somit vorhersehbar, während sie sich über verschiedene Situationen hinweg nicht notwendigerweise konsistent verhalten.

Diese Theorie hat Implikationen für die Messung von Persönlichkeit im Arbeitskontext, denn allgemeine Items sind offen für Interpretationen seitens der Befragten. Unterschiedliche Antworten bzw. Iteminterpretationen führen zu erhöhten Messfehlerraten und somit einer Verringerung der Validität. Daher können diese allgemein gehaltenen Fragebogenverfahren insbesondere im Kontext der Berufseignungsdiagnostik Probleme bereiten, da sie nicht die konditionelle Natur von Persönlichkeitseigenschaften berücksichtigen (Bing et al., 2004).

Das Konzept der konditionalen Dispositionen wurde von Mischel und Shoda (1995) in ihrer *kognitiv-affektiven Systemtheorie der Persönlichkeit* wieder aufgegriffen. Diese Theorie besagt, dass situationsübergreifend stabiles Verhalten nur dann erwartet werden kann, wenn die Situationen die gleichen psychologischen (An-) Reize und Signale aussenden, da das Verhalten eines Individuums von diesen Bedingungen abhängig ist (Lievens et al., 2008).

Auch Hogan (1991) spricht sich in seiner *Self-presentation theory of item responding* für eine kontextbezogene Formulierung von Testitems aus. Seiner Theorie zufolge ist die Beantwortung von Testitems auch ein Aspekt des Sozialverhaltens, da man durch die Beantwortung der Items zu beeinflussen versucht, wie man von anderen Menschen eingeschätzt bzw. bewertet wird – ähnlich wie in „klassischen“ sozialen Interaktionen wie z.B. der face-to-face-Kommunikation. Da generische Items situational mehrdeutig und offen für Interpretationen sind, können Unsicherheiten beim Probanden hinsichtlich der Art der Beantwortung bzw. relevanten sozialen Situation hervorrufen und somit eine zuverlässige Selbsteinschätzung verhindert werden. Eine Standardisierung der Items durch eine kontextbezogene Formulierung könnte personenübergreifend zu einer zuverlässigeren Selbsteinschätzung und somit höheren Vergleichbarkeit sowie Validität der Selbstbeschreibung führen.

Lievens und Kollegen (2008) erweiterten den traditionellen Ansatz des Referenzrahmens, demzufolge es nur aufgrund der höheren inhaltlichen Übereinstimmung zwischen Prädiktor und Kriterium zu einer höheren Kriteriumsvalidität kommt, um einen weiteren Aspekt. Ihnen zufolge verringern kontextualisierte Items nicht nur die interpersonelle Variabilität, wenn alle Testpersonen an die gleiche Situation denken, sondern auch die intrapersonelle Antwortvariabilität. Sie betten die Effekte in die Schema-Theorie ein, der zufolge eine Testperson den Inhalt jedes Items gegen ein (internes) kognitives Schema prüft, das den Kontext für die Itemverarbeitung bzw. relevanten Selbstinformationen darstellt. Dabei wählen die Testpersonen v.a. diejenigen Informationen aus ihrem autobiografischen Gedächtnis aus, die am besten in das bestehende Schema hineinpassen. Da generische Fragebogen aus verschiedenen Items bestehen, ist zu vermuten, dass nicht alle Items dieselben Schemata aktivieren und auf diese

Weise bei der Itembeantwortung allgemeiner Fragebogenverfahren auf verschiedene Referenzrahmen zurückgegriffen wird. Wenn man jedoch von verschiedenen Referenzrahmen ausgeht, muss man Lievens und Kollegen (2008) zufolge die Kovarianz zwischen den Items als Kovarianzen zwischen „verschiedenen“ Items interpretieren. Da die Testpersonen in diesem Falle eine maximale (intrapersonelle) Inkonsistenz in ihrem Antwortverhalten zeigen, sollte dies auch die Reliabilität eines Fragebogens (negativ) beeinflussen. Verwendet man jedoch für die Beantwortung der Items immer den gleichen Referenzrahmen, sollte dies die intrapersonelle Inkonsistenz im Antwortverhalten verringern und somit die Reliabilität eines Fragebogens erhöhen, was wiederum ebenfalls positive Effekte auf die Validität des entsprechenden Instruments haben könnte.

Zusammenfassen lässt sich sagen, dass in der Theorie immer wieder betont wurde, dass Verhalten sowohl von individuellen Eigenschaften als auch Umgebungsfaktoren beeinflusst wird (Viswesvaran et al., 2007). Gemeinsam ist allen Ansätzen, dass sie eine kontextbezogene Formulierung von Testitems als Lösungsansatz vorschlagen, um situationale Verhaltensvariationen im Rahmen von Persönlichkeitstests mit zu erfassen. Die Einbettung der Items und/oder Instruktionen bietet in einen für den Kontext der Testung relevanten Referenzrahmen. Für die Praxis der Personalauswahl und -entwicklung, deren vorrangiges Ziel es ist, Schul- und Arbeitserfolg bzw. Leistung vorherzusagen, ist der relevante Referenzrahmen die Schul- bzw. Arbeitssituation. Allgemeine Items und Persönlichkeitsinventare sagen nur wenig über wirkliche Verhaltensweisen bei der Arbeit aus (Schmit et al., 1995). Der Einfluss kontextbezogener Items in Persönlichkeitsfragebogen wurde in einigen neueren Forschungsarbeiten untersucht, die im Folgenden dargestellt werden.

Eine der ersten Studien zum Vergleich kontextbezogener vs. generischer Items und Instruktionen führten Schmit et al. (1995) durch. Die Forschergruppe formulierte die generischen Items und Instruktionen des NEO-PI-R bzw. NEO-FFI um, sodass diese einen Kontextbezug erhielten, und verglichen die prädiktive Validitäten der Gewissenhaftigkeitskalen in Hinblick auf deren Vorhersagekraft von Schulerfolg. Sie führten die Untersuchung in zwei verschiedenen Studentenstichproben ($N_{S1}=100$ und $N_{S2}=200$) durch. Die eine Hälfte der Studenten bearbeitete die beiden Fragebogenversionen unter neutralen Bedingungen, während der anderen Hälfte der Versuchspersonen gesagt wurde, dass sie möglichst hohe Werte erreichen sollten, um zum Studium zugelassen zu werden (= „Bewerber-Bedingung“). Die Forschergruppe konnte nachweisen, dass kontextspezifisch formulierte Items vermehrt sozial erwünschtes Antwortverhalten hervorriefen und zusätzliche kontextbezogene Instruktionen diesen Effekt noch verstärkten. Allerdings ergab sich sonst kein signifikanter Effekt für die Formulierung der Instruktionen. Die Ergebnisse der Studie sind in Tabelle 2-8 abgebildet.

Tabelle 2-8: Ergebnisse zum Vergleich der generischen vs. kontextualisierten Gewissenhaftigkeitsdimension

		Itemformulierung			
		generisch		kontextualisiert	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
neutrale Bedingung	S1:	33,04	6,00	S1: 37,12	4,12
	S2:	32,31	6,60	S2: 37,50	5,56
Bewerber-Bedingung	S1:	37,48	6,30	S1: 42,00	4,28
	S2:	37,21	6,80	S2: 40,07	6,33

Anmerkungen: $N_{S1}=100$; $N_{S2}=200$; S1= Stichprobe 1; S2= Stichprobe 2; *M*= Mittelwert; *SD*= Standardabweichung.

Es zeigte sich, dass allgemeine Items allgemein geringere Mittelwerte aufwiesen, aber deutlich höhere Streuungen aufzeigten als kontextbezogene Items. Bezüglich der Validität ergab sich eine klare Rangordnung. Die höchste Validität wiesen die kontextualisierten Items in der Bewerber-Bedingung auf, gefolgt von den kontextualisierten Items in der neutralen Bedin-

gung, danach folgten die allgemeinen Items in der neutralen Bedingung. Die geringste Validität wiesen generische Items in der Bewerber-Bedingung auf. Laut Schmit et al. (1995) zeigte auch die Reihenfolge, in der die Tests instruiert wurden, Effekte. Der Unterschied zwischen den beiden Itemversionen war größer, wenn die allgemeine Skala zuerst gegeben wurde. Dies könnte dafür sprechen, dass der Referenzrahmen bei der Beantwortung der kontextualisierten Items zuerst wahrscheinlich übernommen wird und noch aktiv ist, wenn die generischen Items beantwortet wurden.

Aufgrund der Forschungsergebnisse von Dollinger und Orf (1991), dass Gewissenhaftigkeit ($r = .25$) bzw. die Facetten *Dependability* und *Achievement* ($r = .15$ bis $.30$) gute Prädiktoren schulischer Leistungskriterien wie z.B. GPA sind, nahmen Schmit et al. (1995) weiter an, dass die kontextualisierte Items der Gewissenhaftigkeitsskala Schulnoten (GPA) besser vorher-sagen können als allgemeine Items. Die Ergebnisse dieser zweiten Fragestellung entsprachen ihren Annahmen. So ergaben sich die höchsten Kriteriumskorrelationen für die kontextualisierten Items mit $r = .46$ ($p < .01$) für die Bewerber-Bedingung und $r = .41$ ($p < .01$) für die neutrale Bedingung. Die Kriteriumskorrelationen der generischen Items betrug demgegenüber nur $r = .25$ ($p < .01$) für die Bewerber-Bedingung und in der neutralen Bedingung ergab sich kein signifikanter Zusammenhang ($r = -.02$, n.s.).

Dies waren die ersten Ergebnisse, die einen positiven Einfluss kontextualisierter Items auf die Vorhersage von Schulleistung nachweisen konnten. Den Autoren zufolge sollten kontextualisierte Items den Bewerbern auch ermöglichen, sich selbst genauer bzw. korrekter darzustellen.

Allerdings verpassten es Schmit und Kollegen (1995), den Einfluss kognitiver Fähigkeiten, die ebenfalls prognostische Validität bezüglich Schulleistung bzw. GPA besitzen, mit zu erfassen bzw. zu kontrollieren.

Dies wurde von Bing und Kollegen (2004) nachgeholt. Die Forscher postulierten, dass ein Zusammenhang zwischen kontextualisierten, d.h. schulbezogen formulierten Gewissenhaftigkeitsitems, und kognitiven Fähigkeiten eher auf einen Zusammenhang mit kognitiven Fähigkeiten denn auf eine Verbesserung der Gewissenhaftigkeitsmessung per se zurückzuführen sei. Sie nahmen weiter an, dass das Beantworten einer Selbstbeschreibungsskala unter Bewerber-Bedingungen allein auch schon einen ausreichend großen Referenzrahmen erzeugen würde – unabhängig davon, ob Items kontextualisiert seien oder nicht. Sie untersuchten ebenfalls Studenten ($N_{T1} = 342$) in einer Langzeitstudie. Zu Beginn des Herbst-Semesters wurde die Gewissenhaftigkeit der Studenten erfasst und der Notendurchschnitt (GPA) zum Ende des Sommersemesters erhoben ($N_{T2} = 308$). Die Probanden wurden zu T1 wie auch bei Schmit et al. (1995) unter entweder neutralen oder Bewerber-Bedingungen getestet.

Die Bewerberinstruktionen riefen sozial erwünschtes Antwortverhalten hervor, allerdings fielen die Effekte sowohl für generische Items ($d = .20$) als auch für schulbezogene Items ($d = .23$) eher gering aus. Es ergab sich eine hohe Kollinearität zwischen generischen und schulspezifischen Items sowohl in der Bewerber-Bedingung als auch in der neutralen Bedingung ($r = .89$ bzw. $.90$). Die Kontextualisierung der Items führte in beiden Settings zu einer Erhöhung der Korrelation mit dem Schulleistungskriterium. Hierarchische Regressionsanalysen wurden berechnet, um die inkrementelle Validität der kontextualisierten Items zu überprüfen. Die Ergebnisse zeigten, dass diese zur Varianzaufklärung beitragen und somit inkrementelle Validität über allgemeine kognitive Fähigkeiten sowie allgemeine Gewissenhaftigkeit hinaus besitzen. Die Bewerber-Bedingung reduzierte die Validität der Items nur bei generischen Items, bei schulbezogenen Items hingegen erhöhte sie sich.

Hunthausen, Truxillo, Bauer und Hammer (2003) verglichen in einer Feldstudie mit Kundenberatern eine arbeitsbezogen formulierte Version des NEO-FFI mit der allgemeinen Version und konnten zeigen, dass die berufsbezogen formulierten Dimensionen stärker mit beruflichen

Leistungsvariablen zusammenhängen als die allgemeinen Persönlichkeitsdimensionen. Die Ergebnisse zeigten einen moderierenden Effekt des Referenzrahmens auf den Extraversion- und Offenheitsskalen in Bezug auf die Vorhersage von Berufsleistung. Die kontextualisierten Skalen besaßen gegenüber kognitiven Fähigkeitstests inkrementelle Validität.

Pace und Brannick (2010) verglichen eine neu entwickelte arbeitsbezogene Offenheitsskala mit der allgemeinen Offenheitsskala des NEO-PI-R zur Vorhersage von Vorgesetztenbeurteilungen miteinander. Da ein einfaches Hinzufügen des Zusatzes „*bei der Arbeit...*“ bzw. „*in der (Berufs-) Schule...*“ den Autoren zufolge keine adäquate Kontextualisierung darstellt, da diese beispielsweise inhaltlich oftmals keinen Sinn machten, entwickelten sie eine arbeitsbezogene Offenheitsskala für diese Untersuchung. Sie führten in einer Firma eine Fragebogenuntersuchung durch, indem sie 254 Mitarbeiter, die bei ihrer Arbeit technische Probleme lösen müssen, die generische sowie arbeitsbezogene Offenheitsskala beantworten ließen. Zudem erhielten sie von 83 Mitarbeitern zusätzlich eine Vorgesetztenbeurteilung.

Die Ergebnisse zeigten ebenfalls die Überlegenheit der kontextualisierten Skala ($r = .32$, $p < .01$) gegenüber der allgemeinen Offenheitsskala ($r = .09$, n.s.) bei der Vorhersage der Vorgesetztenbeurteilungen. Hierarchische Regressionen zeigten einen starken Effekt für die inkrementelle Validität der arbeitsbezogenen gegenüber der allgemeinen Skala. Bei der Auswertung der Facetten ergaben sich ähnliche Effekte. Die Facetten der allgemeinen Skala zeigten keinerlei Zusammenhänge mit der Vorgesetztenbeurteilung, während einige der arbeitsbezogen formulierten Facetten deutliche Zusammenhänge mit der Vorgesetztenbeurteilung aufwiesen. Auch wenn die Effekte nicht bei allen Facetten nachgewiesen werden konnten, stellt dies den Autoren zufolge keine Einschränkung der Ergebnisse dar, da die Effekte auf eine eingeschränkte Power zurückzuführen sein könnten. Sie empfehlen, in zukünftigen Studien näher zu untersuchen, welche Aspekte von Situationen für die Vorhersage von Verhalten besonders wichtig sind und welche nicht. Hier sollte man insbesondere auch die Facetten der Persönlichkeitsdimensionen näher in Betracht ziehen. Auch hier sollte auf die Übereinstimmung von Facette und Kontext zur Vorhersage von Verhalten geachtet werden.

In ihrem Beitrag verglichen Höft und Schuler (2001) in drei verschiedenen Stichproben ($N_1 = 106$ Studenten ohne Berufserfahrung, $N_2 = 77$ Studenten mit mehr als einem Jahr Berufserfahrung und $N_3 = 73$ Bankkaufleute aus dem Marktsektor) neutrale und berufsbezogene Items miteinander. Die Versuchspersonen füllten in zwei getrennten Erhebungen mit einem Abstand von mindestens drei Wochen sowohl einen verhaltensbezogenen als auch einen neutralen Persönlichkeitsfragebogen aus. Die Ergebnisse deuteten an, dass von den Big 5 die Dimensionen Neurotizismus und Gewissenhaftigkeit am ehesten für eine berufsbezogene Formulierung und Erfassung der Persönlichkeitseigenschaften geeignet sind.

Lievens und Kollegen (2008) verglichen in ihrer Untersuchung generische und kontextualisierte Items nicht nur hinsichtlich ihrer prognostischen Validität, sondern auch bezüglich ihrer Effekte auf die interpersonelle Variabilität (s.o.) miteinander. Sie ordneten 337 Studenten einer von drei Untersuchungsbedingungen zu: 1. generische Items in einer neutralen Untersuchungsbedingung, 2. kontextualisierte Items mit „Schule“ als Referenzrahmen und 3. kontextualisierte Items mit „Arbeit“ als Referenzrahmen. Als Kriterium diente auch in dieser Untersuchung die Durchschnittsnote der Studenten (GPA). Das heißt, das Kriterium war inhaltlich identisch mit dem Referenzrahmen „Schule“, während der Referenzrahmen „Arbeit“ inhaltlich inkongruent war. Dies stellte die Untersuchungsbedingung für das intrapersonelle Design der Studie dar. Eine weitere Stichprobe von 105 Studenten bearbeitete zwei Fragebogen mit einem richtigen (Schule) und einem falschen (Arbeit) Referenzrahmen. Als Schulleistungskriterium diente auch in dieser Stichprobe der GPA der Teilnehmer. Dieses Setting stellte die Untersuchungsbedingung für das intrapersonelle Design dar. Die Ergebnisse zeigen, dass nur die Gewissenhaftigkeitsdimension ($r = .19$, $p < .01$) sowie deren Facetten Achievement ($r = .22$, $p < .01$) und Self-Discipline ($r = .18$, $p < .01$) Schulleistung signifikant vor-

hersagen. Für die anderen Traits der Big 5 ergaben sich keine signifikanten Effekte. Weiter zeigte sich, dass die Kriteriumsvaliditäten am größten für die Gruppe mit dem korrekten Referenzrahmen „Schule“ waren (s. Tabelle 2-9).

Tabelle 2-9: Interne Konsistenzen und Kriteriumsvaliditäten unterteilt nach Referenzrahmen (nach Lievens et al., 2008)

	Referenzrahmen					
	keiner (n= 115)		Schule (n= 91)		Arbeit (n= 98)	
	α	Val.	α	Val.	α	Val.
Achievement	.79	.16	.82	.41	.85	.12
Self-Discipline	.89	.16	.90	.34	.91	.06
Neurotizismus	.91	.06	.90	.09	.86	-.13
Extraversion	.88	.04	.89	.06	.87	-.19
Offenheit	.82	.12	.76	.03	.70	-.06
Verträglichkeit	.84	-.02	.84	.05	.84	-.09
Gewissenhaftigkeit	.78	.09	.79	.37	.84	.16

Anmerkungen: α = Reliabilität (Cronbach's Alpha); Val.= Validität; n= Teilstichproben.

Es zeigte sich, dass die Probanden mit dem kongruenten Referenzrahmen „Schule“ deutlich höhere Kriteriumsvaliditäten aufwiesen als in der Bedingung ohne oder mit dem inkongruenten Referenzrahmen. So führte der Einsatz des Referenzrahmens „Schule“ im Vergleich zur neutralen Bedingung auf der Gewissenhaftigkeitsdimension zu einem Validitätszuwachs von .28. Vor allem die Gewissenhaftigkeitsskala sowie deren Facetten *Achievement* und *Self-Discipline* wiesen signifikante und vergleichsweise hohe prognostische Validitäten auf. Darüber hinaus sprechen die Ergebnisse dafür, dass die intrapersonelle Varianz durch das Hinzufügen eines Referenzrahmens verringert wird, d.h. zu höheren Reliabilitäten führt. Die positiven Effekte der Referenzrahmen kamen allerdings nur zum Tragen, wenn ein kongruenter Bezug mit dem Kriterium vorlag. Die Validität einer Selbstbeschreibung ist somit an die Interpretation der Items durch den Getesteten gebunden.

Auch Reddock et al. (2010) fanden eine Validitätssteigerung kontextualisierter Fragebogenverfahren von .07 im Vergleich zu generischen Selbstbeschreibungen. Der Mittelwert der kontextualisierten Gewissenhaftigkeitsdimension war hierbei etwas höher als in der generischen Version. Die gleiche Forschergruppe führte ein Jahr später eine Untersuchung zur Vorhersagekraft kontextualisierter Gewissenhaftigkeit in Bezug auf Schulleistung (GPA) durch und untersuchten ebenfalls die intrapersonelle Variabilität innerhalb der Big 5-Dimensionen. Intrapersonelle Variabilität (IPV) wurde als mittlere Standardabweichung pro Versuchsperson operationalisiert. Reddock und Kollegen (2011) untersuchten wiederum eine Studentengruppe ($N= 329$), die sich selbst anhand kontextualisierter und generischer Versionen eines Big 5-Inventars einschätzten. Darüber hinaus wurden die allgemeinen kognitiven Fähigkeiten der Probanden erfasst. Die Ergebnisse der Studie wiesen wiederum auf eine höhere prädiktive Validität der kontextualisierten Gewissenhaftigkeitsdimension ($r= .27$) im Vergleich zur allgemeinen Erfassung der Gewissenhaftigkeit ($r= .20$) hin. Zudem wiesen die Extraversion- und Verträglichkeitsskalen mit Kontextbezug eine signifikant höhere interne Konsistenz auf. Für die Neurotizismus- und Gewissenhaftigkeitsdimensionen konnte dieser Effekt allerdings nicht nachgewiesen werden. Demgegenüber zeigte sich IPV jedoch als konsistentes Maß, was den Autoren zufolge andeutet, dass es sich bei dieser Variable um eine Charaktereigenschaft handeln könnte, da dieser Effekt nicht auf eine Dimension beschränkt war. Die Forscher korrelierten die IPV-Variable daher mit dem Maß zur Erfassung der allgemeinen kognitiven Fähigkeiten. Sie fanden, dass diese negativ mit Intelligenz korreliert, auch nachdem Gewissenhaftigkeit herauspartialisiert wurde. IPV ist den Autoren zufolge somit ein signifikanter Prädiktor allgemeiner kognitiver Fähigkeiten und in der Lage, zusätzliche Validität über Ge-

wissenschaftlichkeit hinaus zu erklären. Zudem scheint IPV den Zusammenhang zwischen Gewissenhaftigkeit und Schulleistung zu moderieren. Die Autoren empfehlen, die intrapersonelle Antwortvariabilität in zukünftigen Studien stärker zu beachten, da diese sich auf die Höhe von beispielsweise Reliabilitäten auswirkt. Vor allem für Normierungsstichproben, die möglichst heterogen und somit variabel hinsichtlich der Antworttendenzen sein sollten, ist dies ein wichtiger Hinweis.

Trotz der eindeutigen Befunde zur Validitätssteigerung kontextualisierter Fragebögen gibt es auch einige Probleme, die mit der Erforschung dieser Ansätze einhergehen. Beispielsweise erhalten die kontextualisierten Skalen die gleichen Bezeichnungen wie die allgemeinen Skalen. Dies könnte einerseits zu Verwechslungen führen und andererseits ist dann auch die Konstruktvalidität der kontextualisierten Skalen fraglich.

Holtz, Ployhart und Dominguez (2005) untersuchten die Wahrnehmung und Akzeptanz verschiedener Testverfahren. Ihren Ergebnissen zufolge ergab sich allerdings für die kontextualisierten Versionen im Vergleich mit generischen Fragebogen-Versionen keine positivere Augenscheinvalidität. Einige Studien zeigen, dass arbeitsbezogene Skalen die Konstruktvalidität unterschätzen (Salgado et al., 2003). Zudem ist es möglich, dass kontextualisierte Skalen bestimmte Antwortmuster hervorrufen oder sozial erwünschte Antworttendenzen verstärken. Skalen, die sich inhaltlich auf bestimmte Kontexte, wie z.B. Arbeit, beziehen, könnten darüber hinaus Items enthalten, die sich auf Verhaltensweisen oder Arbeitsinhalte einer Person beziehen, die jedoch nicht zum Verhaltensrepertoire aller Menschen gehören (z.B. Kalendernutzung). Auch weisen die oben beschriebenen Befunde von Lievens et al. (2008) sowie Reddock et al. (2010) darauf hin, dass es für den Fall, dass der verwendete Referenzrahmen nicht mit dem zu prognostizierenden Kriterium oder zu messenden Konstrukt übereinstimmt, zu Validitätseinbußen kommen kann. Insofern kann sich ein falscher Referenzrahmen, beispielsweise eine arbeitsbezogene Skala zur Vorhersage von Schulerfolg, sogar schwächend auf die Zusammenhänge zwischen Prädiktor und Kriterium auswirken.

Bezüglich des Selbstwirksamkeitskonstruktes ist diese Diskussion schon weiter vorangeschritten. Bereits Bandura (1977) postulierte, dass Kontextfaktoren wie beispielsweise soziale, situationsspezifische oder zeitliche Umstände, denen das Verhalten unterliegt, die Basis für Selbstwirksamkeitserwartungen seien. Schwarzer et al. (1997) fanden kulturell bedingte Unterschiede bezüglich der Ausprägung der Selbstwirksamkeit in einer internationalen Vergleichsstudie bei Menschen aus China, Costa Rica und Deutschland, die sie auf die unterschiedlichen Sozialisationshintergründe zurückführen. Demzufolge ist Selbstwirksamkeit nicht unabhängig von Kontextfaktoren und sollte daher im Kontext der Personalauswahl und -entwicklung ebenfalls berufsbezogen erfasst werden. Unterschiede zwischen Selbstwirksamkeitserwartung und gezeigtem Verhalten sind vor allem auf die Zweideutigkeit von Situations- und Aufgabenfaktoren zurückzuführen. Die Forschungsergebnisse belegen darüber hinaus, dass situationsspezifische Selbstwirksamkeit Verhalten präziser Vorhersagen kann als ein Globalfaktor.

Diese Sichtweise sollte nun auch in der Erfassung und Beurteilung von für die berufliche Eignungsdiagnostik relevanten Persönlichkeitsvariablen wie z.B. Gewissenhaftigkeit Einzug halten. Denn selbst nach Korrekturen hinsichtlich Prädiktor- und Kriteriumsunreliabilität und Varianzeinschränkungen bleibt die Kriteriumsvalidität des Gewissenhaftigkeitskonstrukts mäßig bis gering. Durch das Einführen eines situationsspezifischen Bezugsrahmens kann dieser Effekt verhindert bzw. nachweislich die Kriteriumsvalidität des Gewissenhaftigkeitskonstrukts als Vorhersagemaß erhöht werden (Bing et al., 2004; Hunthausen et al., 2003; Schmit et al., 1995). Eine Verschiebung des wissenschaftlichen Fokus' bzw. die Einführung eines gemeinsamen *Bezugsrahmens* oder auch *Frame-of-Reference* (FOR) sollte daher zur Erhöhung der prädiktiven und Kriteriumsvalidität von in der Personalauswahl und -entwicklung eingesetzten Persönlichkeitsmaßen vorgenommen werden. Eine solche arbeitsbezogene Formu-

lierung brächte außerdem einen zusätzlichen Vorteil durch eine höher wahrgenommene Augenscheinvalidität und daraus resultierende höhere wahrgenommene Relevanz und Fairness sowohl auf Seiten der Bewerber als auch bei den Personalern.

Zusammenfassend kann gesagt werden, dass die Idee, dass der Kontext bei der Vorhersage von Verhalten eine große Rolle spielt, nicht neu ist. Schon Mischel (1968) betonte, dass der Einfluss des Kontexts auf das Verhalten von Menschen groß sei. Insbesondere der Unterschied zwischen beruflichem und privatem Kontext ist gravierend. Dieser Unterschied beeinflusst auch die Persönlichkeit eines Menschen, was zu Unterschieden bei der Vorhersage von Verhalten führen kann. Indem man Personen, die einen Fragebogen bearbeitet, einen bestimmten Referenzrahmen vorgibt, hilft man bei der Klärung der Itembedeutung und reduziert auf diese Weise Messfehler (vgl. Bing et al., 2004; Lievens et al., 2008). Sämtliche Studien, die die Validitäten von Persönlichkeitsinventaren mit und ohne Referenzrahmen verglichen, kommen übereinstimmend zu dem Ergebnis, dass kontextualisierte Items zu höheren prädiktiven Validitäten führen, wenn der Kontext von Test und Kriterium identisch sind (z.B. Bing et al., 2004; Hunthausen et al., 2003; Lievens et al., 2008; Schmit et al., 1995). Da ein einfaches Hinzufügen der Phrase „bei der Arbeit“ nicht genügt, um eine arbeitsbezogene Skala zu entwickeln und bei vielen Items auch keinen Sinn macht (Lievens et al., 2008; Pace & Brannick, 2010; Schmit et al., 1995), sollten berufsbezogene Gewissenhaftigkeitsmaße neu entwickelt werden, denn *„the construction of context-specific personality items provides an opportunity to improve the prediction of performance and to add precision and depth to our understanding of personality“* (Bing et al., 2004, S. 157).

2.6 Vorliegende Instrumente zur Erhebung von Gewissenhaftigkeit

Es gibt eine Vielzahl an Instrumenten zur Erfassung von Persönlichkeitsmerkmalen. Vor allem aufgrund der oben geschilderten Diskussion, welches Faktorenmodell die Persönlichkeit am besten beschreibt (vgl. Abschnitt 2.3), sind viele verschiedene Inventare entstanden (Backhaus, 2004b). Jedoch gibt es nicht nur aufgrund der Uneinigkeit bezüglich der zugrundeliegenden Struktur der Persönlichkeit viele verschiedene Testverfahren. Goldberg (1971) formulierte diesbezüglich, dass *„the most potent source of variance in the determination of the constructs for past scales and inventories has been sheer historical accident“* (S. 335). So gibt es – neben den anderen Persönlichkeitsmerkmalen – eine Vielzahl an Verfahren, die Gewissenhaftigkeit (mit) erfassen. MacCann et al. (2009) geben in ihrer Untersuchung der Facettenstruktur von Gewissenhaftigkeit einen kurzen Überblick über die prominentesten Fragebogenverfahren zur Erfassung von Gewissenhaftigkeit. Ihnen zufolge gehören hierzu u.a. das Jackson Personality Inventory (Jackson, 1994), das NEO-PI-R bzw. NEO-FFI (Costa & McCrae, 1992) und der Six Factor Personality Questionnaire (Jackson, Paunonen & Tremblay, 2000).

Die zunehmende Etablierung des Fünf-Faktoren-Modells der Persönlichkeit hatte zur Folge, dass die Big 5 vermehrt auch in Anwendungskontexten erhoben werden. Da die Persönlichkeitseigenschaften in diesen Kontexten oftmals aber nur als Zusatzvariablen mit erfasst werden, ist die Zeit, die für diese Tests aufgewendet wird, häufig stark beschränkt bzw. sind die herkömmlichen Verfahren zur Erfassung der Big 5 oft zu umfangreich (Rammstedt & John, 2005). So erfasst das NEO-PI-R die Big 5 mit 240 Items und die Kurzversion NEO-FFI wartet immer noch mit 60 Items auf. Auch De Raad und Perugini (2002) führen in ihrem Handbuch „Big Five Assessment“ kein Inventar an, das die fünf Faktoren mit weniger als 50 Items erfasst. Das bisher kürzeste standardisierte Verfahren zur Erfassung der Big 5 ist das 44 Items umfassende Big Five Inventory (BFI; John, Donahue & Kentle, 1991).

Darüber hinaus erfassen die meisten Persönlichkeitsinventare standardmäßig immer alle Persönlichkeitsfaktoren, die dem zugrundeliegenden Testmodell angehören. Keines der bisherigen

Verfahren konzentriert sich auf die alleinige Erfassung berufsrelevanter Persönlichkeitsmerkmale wie Gewissenhaftigkeit. Da diese Persönlichkeitsdimension jedoch die stärkste Vorhersagekraft in Bezug auf Arbeitsleistung hat, kann es hilfreich sein, diese in Bezug auf die Vorhersage von Arbeitsverhalten (allein) mit zu erheben (Minbashian et al., 2010).

Erschwerend kommt hinzu, dass die Konstruktion von Persönlichkeitstests augenscheinlich sehr einfach ist und viele Unternehmen oder selbsternannte Diagnostiker eigene „Persönlichkeitstest“ entwickelt haben, die jedoch jeglicher theoretischer Grundlagen entbehren und auch den Ansprüchen an eine wissenschaftliche Testkonstruktion nicht genügen (Ackerman & Heggestad, 1997; Sarges, 2010). Dieser Trend hält nach wie vor an. Für den uninformatierten Personaler ist es daher schwer, diesbezüglich „die Spreu vom Weizen“ zu trennen, zumal die unwissenschaftlichen Persönlichkeitstests oftmals mit einem nicht unerheblichen Werbeaufwand vertrieben werden, während die wissenschaftlich fundierten Testverfahren eher wenig adressatengerecht beworben und erklärt werden.

Eine umfassende Darstellung sämtlicher Persönlichkeitsinventare, die Gewissenhaftigkeit (mit) erfassen, ist nicht das Ziel dieser Studie und wäre zu umfangreich für diese Arbeit. Die im Folgenden beschriebenen Persönlichkeitsmaße stellen daher nur einen Ausschnitt bestehender Persönlichkeitsinventare dar und erheben keinen Anspruch auf Vollständigkeit. Vielmehr sind an dieser Stelle im Hinblick auf das Ziel dieser Arbeit nur einige der prominentesten wissenschaftlichen Verfahren, theoretisch bedeutsamen Neuentwicklungen und/ oder in der Praxis am meisten verwendeten Tests kurz und überblicksartig dargestellt. Hierzu zählen die NEO-Persönlichkeitsinventare (Abschnitt 2.6.1), die Personality Research Form (2.6.2), die Big Five Aspect Scales (2.6.3), das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (2.6.4), das Freiburger Persönlichkeitsinventar (2.6.5), das Big Five Inventory (2.6.6) und der START-P (2.6.7). Abschnitt 2.6.8 fasst den aktuellen Stand der Persönlichkeitsinventare, die Gewissenhaftigkeit (mit) erfassen, noch einmal zusammen.

2.6.1 NEO-Persönlichkeitsinventare (NEO-PI-R und NEO-FFI)

Das Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) ist ein fakto-renanalytisch konstruiertes Fragebogenverfahren, das auf den psycholexikalischen Ansatz zurückgeht (s. Abschnitt 2.3).

Der NEO-PI-R erfasst mit 240 Items auf einer 5-stufigen Likert-Skala die fünf Persönlichkeitsdimensionen Neurotizismus, Extraversion, Offenheit, Verträglichkeit und Gewissenhaftigkeit. Jedem der fünf Faktoren werden sechs Facetten zugeordnet (= 8 Items je Facette). Es gibt eine Selbst- und eine Fremdeinschätzungsversion.

Die Kurzform, das NEO Fünf-Faktoren Inventar (NEO-FFI; Costa & McCrae, 1992), erfasst lediglich die fünf groben Dimensionen des Fünf-Faktoren-Modells mittels 60 Items (zwölf Items je Faktor), nicht aber die Facetten der Persönlichkeitsdimensionen.

Sowohl die Lang- als auch die Kurzversion des NEO-Tests wurde ursprünglich im amerikanischen Sprachraum entwickelt. Sie liegen mittlerweile in verschiedenen Sprachen vor und werden in mehr als 50 Ländern weltweit eingesetzt. Für den deutschen Sprachraum übersetzten und normierten Borkenau und Ostendorf (1993) das NEO-FFI, wofür sie 15 Jahre später noch einmal neue Normwerte vorlegten (Borkenau & Ostendorf, 2008). Das Verfahren ist den Autoren zufolge eher als „Breitbandverfahren“ zur Erfassung der Big 5 zu verstehen. Die deutsche Version des NEO-PI-R wurde von Ostendorf und Angleitner (2004) erarbeitet und veröffentlicht.

Angaben zur empirischen Evidenz der beiden Verfahren, beispielsweise zur konvergenten und diskriminanten Validität sowie externen Kriterien wie akademischer und beruflicher Leistung, sind in den jeweiligen Manualen enthalten (Borkenau & Ostendorf, 1993, 2008; Costa &

McCrae, 1992; Ostendorf & Angleitner, 2004). Ostendorf und Angleitner (2004) berichten gute bis sehr gute interne Konsistenzen sowohl in der Selbst- als auch in der Fremdbeurteilungsversion für die Hauptskalen ($\alpha = .87$ bis $.95$) und teilweise geringere Werte für die Facetten ($\alpha = .49$ bis $.91$) des NEO-PI-R. Die Retest-Reliabilität wurde sowohl für die Faktoren als auch für die Facetten nach einem Monat ($r = .48$ bis $.91$) und nach fünf Jahren ($r = .53$ bis $.78$) erhoben. Die Daten wurden an einer Normstichprobe von $N = 11.724$ Personen erhoben, die hinsichtlich ihrer Altersstruktur sowie der Geschlechterverteilung und ihrer Bildung repräsentativ für Deutschland war. Die hohe Schwankung der Retest-Reliabilitäten deutet allerdings an, dass die gemessenen Faktoren nicht nur stabil sind im Sinne eines Traits, sondern die Beantwortung der Items bzw. Skalen Schwankungen unterlag.

Eine erste Untersuchung der deutschsprachigen Übersetzung des NEO-FFI an 578 Probanden ergab eine interne Konsistenz von $\alpha = .72$ bis $.86$. Eine Retestung an einem Teil der Stichprobe ($N = 146$) nach einem Zeitraum von zwei Jahren ergab eine gute Retest-Stabilität von $.77$ (Borkenau & Ostendorf, 1991). Es liegen Normdaten für zwölf Geschlechts- und Altersgruppen aus einer bevölkerungsrepräsentativen Quotenstichprobe ($N = 11.724$) vor, die befriedigende bis gute interne Konsistenzen ($\alpha = .72$ bis $.87$) und Retest-Reliabilitäten ($r = .71$ bis $.82$) für die fünf Skalen berichten.

Eine Vorhersage spezifischer Kriterien war mit den NEO-Verfahren ursprünglich nicht vorhergesehen (Borkenau & Ostendorf, 1993), auch wenn mittlerweile über eine Vielzahl von korrelativen Zusammenhänge zwischen den einzelnen Persönlichkeitsdimensionen und beispielsweise beruflichen sowie schulischen Leistungsvariablen berichtet wurde (Borkenau & Ostendorf, 2008; Costa & McCrae, 1992; Hülshager & Maier, 2008).

Laut einer Umfrage von Hagemeister, Lang und Kersting (2010) gehören beide Verfahren – NEO-PI-R und NEO-FFI – mit zu den am häufigsten verwendeten bzw. genannten Verfahren in der psychologischen Diagnostik, wobei beide Testversionen gleich häufig genannt wurden. Auch Bühner (2011) bezeichnet NEO-PI-R und NEO-FFI als bedeutsame Vertreter der psychometrischen Persönlichkeitstests.

Allerdings sind die Items der NEO-Persönlichkeitsinventare allgemein gehalten. Eine berufsspezifische Erfassung der Persönlichkeit erfolgt nicht. In den Manualen sind nur wenige Angaben für die Anwender enthalten, was Lord (2011) mit ihrem Praxishandbuch zu den Verfahren nachzuarbeiten versucht. Eine Erfassung einzelner Skalen oder Facetten allein ist nicht vorgesehen, sodass die Anwendung vor allem der Langversion (NEO-PI-R) in der Praxis der Personalauswahl eher zu lang dauert, die Erfassung der fünf Merkmale mittels des NEO-FFI aber keine Rückschlüsse auf die Facetten erlaubt.

2.6.2 Personality Research Form (PRF)

Die Personality Research Form (PRF) wurde 1967 erstmals von D. N. Jackson eingeführt. Es handelt sich bei der PRF um ein multivariates Fragebogenverfahren, das auf der Persönlichkeitstheorie Murrays aufbaut und mit 234 Items insgesamt 20 Persönlichkeitsvariablen erfasst. Im anglo-amerikanischen Sprachraum war die PRF in den 1990er Jahren eines der am meisten verwendeten Persönlichkeitsinventare (Jackson et al., 1996). Dies ist darauf zurückzuführen, dass das Verfahren ein umfassendes Bild der Charaktereigenschaften einer Person abliefern und die verwendeten Konzepte leicht verständlich und kommunizierbar sind. Ein Schwerpunkt der PRF liegt auf der Erfassung von Aspekten des Leistungs- und Sozialverhaltens einer Person.

Die deutsche Version der PRF wurde von Stumpf, Angleitner, Wieck, Jackson und Beloch-Till (1985) übersetzt und angepasst. Es handelt sich hierbei um eine inhaltlich gekürzte Version der Originalausgabe, die lediglich 14 der 20 ursprünglichen Skalen enthält: *Achievement*

(Leistungsstreben), *Affiliation* (Geselligkeit), *Aggression* (Aggressivität), *Dominance* (Dominanzstreben), *Endurance* (Ausdauer), *Exhibition* (Bedürfnis nach Beachtung), *Harmavoidance* (Risikomeidung), *Impulsivity* (Impulsivität), *Nurturance* (Hilfsbereitschaft), *Order* (Ordnungsstreben), *Play* (Spielerische Grundhaltung), *Social Recognition* (Soziales Anerkennungsbedürfnis), *Succorance* (Anlehnsbedürfnis) und *Understanding* (Allgemeine Interessiertheit).

Es gibt eine Selbst- und eine Fremdeinschätzungsversion sowie zwei parallele Testversionen A und B, die trotz der inhaltlichen Kürzung der Skalenstruktur von 20 auf 14 dennoch genauso viele Items wie die Originalversion enthalten. Die 234 Items sind als Aussagen formuliert, die dichotom (*richtig* vs. *falsch*) beantwortet werden müssen. Zu jeder Skala gehören 16 Items. Es liegen Normen hinsichtlich verschiedener Altersgruppen sowie für jedes Geschlecht vor ($N = 4.327$). Die PRF kann bei Personen ab 17 Jahren eingesetzt werden.

Jackson et al. (1996) berichten von guten Reliabilitäten für die Selbsteinschätzungsskala ($\alpha = .81$ bis $.88$) sowie guten bis sehr guten internen Konsistenzen für die Fremdeinschätzungsversion ($\alpha = .85$ bis $.91$). Die Retest-Reliabilität des Verfahrens nimmt mit der Zeit etwas ab. So berichten die Testautoren nach ein paar Wochen von guten bis sehr guten Stabilitätskoeffizienten ($r = .85$ bis $.96$), nach einem ($r = .67$ bis $.85$) bzw. drei Jahren ($r = .69$ bis $.83$) verringerte sich diese auf ein ausreichendes bis gutes Niveau.

Bereits bei der Vorstellung des Verfahrens plädierte Jackson (1967) für eine Zweiteilung des Gewissenhaftigkeitsfaktors in zwei distinkte, aber korrelierte Faktoren *Achievement/ Industriousness* und *Methodicalness*. Dem ersten Faktor sind die Skalen Achievement, Endurance und Play zugeordnet, während dem zweiten Faktor die Skalen Cognitive Structure, Order und Impulsivity zugeordnet wurden. Diese aggregierten Faktoren untersuchten Jackson et al. (1996) hinsichtlich ihrer Kriteriumsvalidität. Sie fanden, dass der Achievement-Faktor mit Schulleistung zusammenhing ($r = .24$, $p < .02$), während Methodicalness dieses Muster nicht zeigte ($r = -.01$, n.s.). Demgegenüber korrelierte Methodicalness negativ mit der Aufnahme eines Kunststudiums ($r = -.22$, $p < .05$). Dieser Zusammenhang konnte demgegenüber für den Achievement-Faktor nicht gefunden werden ($r = -.07$, n.s.). Den Autoren zufolge spricht dies für die differentielle Vorhersagekraft der beiden aggregierten Faktoren und somit für deren Eigenständigkeit. Allerdings sind bis dato keine Metaanalysen zur Faktorenstruktur der PRF veröffentlicht worden. Auch gibt es keine eindeutigen Empfehlungen, inwiefern Skalen allein verwendet werden können. Vielmehr ist der Anwender auch bei diesem Fragebogen gezwungen, sämtliche Faktoren mit zu erheben, die in einer eignungsdiagnostischen Erhebung jedoch nicht immer relevant sind. Somit handelt es sich bei der PRF ebenfalls um ein – für die Praxis – aufwendig und unökonomisch zu handhabendes Verfahren.

2.6.3 Big Five Aspect Scales (BFAS)

Die Big Five Aspect Scales (BFAS) wurden von DeYoung et al. (2007) entwickelt. Die BFAS sind in das später veröffentlichte hierarchische Modell der Persönlichkeit (DeYoung, 2010) theoretisch eingebettet, das bereits in Abschnitt 2.3.1 erläutert wurde. Die Forschergruppe ließ Probanden einer bevölkerungsrepräsentativen Stichprobe in den USA (Eugene-Springfield Community; $N = 481$) zwei Selbstbeschreibungsinventare bearbeiten, die auf der Fünf-Faktoren-Theorie basieren und die Big 5 auch auf der Facettenebene erfassen. Die Ergebnisse ihrer Faktorenanalyse ergaben, dass jede der fünf Dimensionen über eine zweifaktorielle Subebene verfügt, die zwei distinkte, aber korrelierte Aspekte enthält. Für Neurotizismus benannten sie diese in *Volatility* (Sprunghaftigkeit) und *Withdrawal* (Rückzugsverhalten). Die Dimension Extraversion teilten die Forscher in die Aspekte *Enthusiasm* (Begeisterungsfähigkeit) und *Assertiveness* (Durchsetzungsfähigkeit) auf. Verträglichkeit setzt sich aus den beiden Aspekten *Compassion* (Mitgefühl) sowie *Politeness* (Höflichkeit) zusammen. Die Offenheits-

dimension unterteilten sie in *Intellect* (Intellekt) und *Openness* (Offenheit). Gewissenhaftigkeit besteht aus den Aspekten *Industriousness* (Fleiß) und *Orderliness* (Ordnung).

Aufbauend auf diesen Ergebnissen entwickelten DeYoung et al. (2007) die Big Five Aspect Scales (BFAS). Zu jedem der zehn Aspekte formulierten sie zehn Aussagen, sodass die BFAS insgesamt über 100 Items verfügt. Als Antwortformat schlagen die Autoren eine 5-stufige Likertskala vor. Diese Skala legte die Forschergruppe wiederum den Teilnehmern der Eugene-Springfield-Community sowie einer gleich großen studentischen Stichprobe ($N=480$) zur Bearbeitung vor. Tabelle 2-10 enthält die Ergebnisse der Skalenüberprüfung der BFAS.

Tabelle 2-10: Deskriptive Statistiken der BFAS (nach DeYoung et al., 2007)

	Stichprobe 1			Stichprobe 2		<i>r</i>
	<i>M</i> (<i>SD</i>)	α	<i>M</i> (<i>SD</i>)	α_1	α_2	
Neurotizismus	2,46 (.63)	.89	2,82 (.70)	.89	.89	.85
Volatility	2,48 (.70)	.85	2,72 (.82)	.87	.89	.85
Withdrawal	2,45 (.71)	.84	2,92 (.75)	.81	.80	.81
Agreeableness	4,11 (.45)	.84	3,70 (.56)	.85	.89	.79
Compassion	4,11 (.54)	.84	3,87 (.65)	.84	.91	.79
Politeness	4,10 (.53)	.75	3,52 (.67)	.76	.76	.74
Conscientiousness	3,76 (.51)	.84	3,06 (.56)	.81	.82	.86
Industriousness	3,80 (.61)	.81	2,84 (.70)	.79	.82	.82
Orderliness	3,73 (.62)	.80	3,28 (.64)	.72	.74	.79
Extraversion	3,48 (.60)	.85	3,37 (.63)	.88	.86	.83
Enthusiasm	3,59 (.72)	.81	3,52 (.73)	.81	.80	.73
Assertiveness	3,36 (.70)	.85	3,21 (.71)	.84	.88	.86
Openness/ Intellect	3,72 (.53)	.85	3,47 (.52)	.80	.82	.82
Intellect	3,70 (.68)	.84	3,39 (.67)	.79	.81	.86
Openness	3,74 (.61)	.78	3,52 (.64)	.72	.77	.79

Anmerkungen: Stichprobe 1= Eugene-Springfield Community ($N=481$); Stichprobe 2= College-Studenten ($N=480$); *M*= Mittelwert; *SD*= Standardabweichung; α = interne Konsistenz; α_1 = interne Konsistenz in der College-Stichprobe; α_2 = interne Konsistenz der Retest-Stichprobe ($N=90$); *r*= Retest-Reliabilität.

Der BFAS ist den Daten zufolge ein zuverlässiges Instrument zur Erfassung der zehn Persönlichkeitsaspekte, dessen Skalen befriedigende bis sehr gute interne Konsistenzen aufweisen sowie über eine befriedigende bis gute Retest-Reliabilität verfügen. Allerdings gibt es bis dato keine gesicherten Erkenntnisse zur Kriteriumsvalidität der BFAS. Dreier (2008) korrelierte die Items der Gewissenhaftigkeitsaspekte *Industriousness* und *Orderliness* mit Schulnoten (GPA) in einer US-amerikanischen Stichprobe ($N=217$ Collegestudenten). Ihren Ergebnissen zufolge korreliert *Industriousness* mit Schulleistung ($r=.15$, $p<.05$), während *Orderliness* nicht mit dem Kriterium zusammenhängt ($r=.07$, n.s.). Struckmeier (2009) setzte in einer vergleichbaren Studie eine ins Deutsche übersetzte Version der Gewissenhaftigkeitsaspekte in einer Feldstichprobe von Arbeitnehmern in verschiedenen Unternehmen in Deutschland ($N=340$) ein. Ihre Ergebnisse zeigen, dass weder die Gewissenhaftigkeitsaspekte noch die Gewissenhaftigkeitsdimension mit Arbeitsleistung zusammenhängen.

Studien zur Vorhersagekraft der anderen Persönlichkeitsaspekte liegen bis dato nicht vor. Eindeutige Empfehlungen, inwiefern Skalen auch allein verwendet werden können, gibt es nicht. Die Items sind zudem nicht berufsbezogen formuliert. Außerdem existieren die BFAS bisher nur in englischer Sprache. Eine Übersetzung für den deutschen Markt liegt noch nicht vor. Somit ist eine Verwendung der BFAS – trotz der guten theoretischen Skalenkonstruktion – in der deutschsprachigen Eignungsdiagnostik (noch) nicht angezeigt.

2.6.4 Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP)

Das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) wurde unter der Leitung von Rüdiger Hossiep in enger Abstimmung mit Personalpraktikern entwickelt. Das Ziel der Forschergruppe war es, die in der Praxis als wichtig erachteten Eigenschaften und Verhaltensweisen von Mitarbeitern zu identifizieren, die als überfachliche Kompetenzen eine bedeutsame Determinante beruflichen Erfolges darstellen, und darauf aufbauend eine Testbatterie zur standardisierten Erfassung dieser Merkmale zu entwickeln. Die Autoren erhoben den Anspruch, ein Verfahren zu entwickeln, das praxisrelevant, von Kandidaten und Personalern gleichermaßen akzeptiert und wissenschaftlich fundiert ist sowie ausschließlich beruflich relevante Kriterien erhebt (Hossiep & Paschen, 2003).

Basierend auf einer Befragung von Praktikern wurden von Hossiep und Kollegen 17 Persönlichkeitseigenschaften identifiziert, die neben der fachlichen Eignung einer Person als „überfachliche Kompetenzen“ als bedeutsam für das Berufsleben gelten. Diese Eigenschaften wurden von den Autoren vier Persönlichkeitsbereichen zugeordnet. Der erste Bereich *Berufliche Orientierung* setzt sich aus den Eigenschaften *Leistungs-*, *Gestaltungs-* und *Führungsmotivation* sowie *Wettbewerbsorientierung* zusammen. Zum Bereich *Arbeitsverhalten* zählen die Autoren die Eigenschaften *Gewissenhaftigkeit*, *Flexibilität*, *Handlungs-* und *Analyseorientierung*. Der dritte Bereich der *Sozialen Kompetenzen* setzt sich aus den Eigenschaften *Sensitivität*, *Kontaktfähigkeit*, *Soziabilität*, *Teamorientierung*, *Durchsetzungsstärke* und *Begeisterungsfähigkeit* zusammen. Der vierte Bereich wurde als *Psychische Konstitution* benannt und beinhaltet die Eigenschaften *Emotionale Stabilität*, *Belastbarkeit* sowie *Selbstbewusstsein* (Hossiep & Collatz, 2010). Die zweite Auflage des BIP beinhaltet demgegenüber nur 14 Eigenschaften (Hossiep & Paschen, 2003). Das Verfahren wird allerdings beständig weiter entwickelt und validiert, was zu Unterschieden in der Faktorenstruktur und Itemanzahl innerhalb der verschiedenen Testversionen führt.

Zu jedem Bereich formulierten die Autoren Aussagen, die auf einer 6-stufigen Likertskala von „trifft voll zu“ bis „trifft überhaupt nicht zu“ beantwortet werden sollen. Aus dem ursprünglich sehr großen Itempool wählten die Autoren diejenigen Items aus, die sie als inhaltlich und statistisch bedeutungsvoll erachteten. Derzeit umfasst die aktuelle Forschungsversion des BIP 251 Items (Hossiep & Collatz, 2010). Das Testverfahren liegt als Selbst- und Fremdbeschreibungsbogen vor.

Mittlerweile ist das Testverfahren an 9.303 Personen normiert. Es liegen Normen für Hochschulabsolventen, für verschiedene betriebliche Hierarchiestufen sowie für unterschiedliche Funktionsbereiche (z.B. Vertrieb) vor. Auch Normen für weibliche Fach- und Führungskräfte sind vorhanden (Hossiep & Paschen, 2003).

Die interne Konsistenz der Skalen ist als befriedigend bis sehr gut zu bewerten ($\alpha = .74$ bis $.91$). Die Retest-Reliabilitäten der einzelnen Eigenschaften liegen nach einem Zeitraum von zwei bis drei Jahren bei $r = .71$ bis $.79$ und sind somit als befriedigend zu bezeichnen. Zudem zeigen die Skalen des BIP mittlere Zusammenhänge mit beispielsweise beruflichem Erfolg (Einkommen und Hierarchiestufe) und beruflicher Zufriedenheit. Zudem konnte gezeigt werden, dass die Eigenschaften mit verwandten Konstrukten anderer Persönlichkeitsfragebogen (z.B. NEO-FFI) korrelieren (Hossiep & Paschen, 2003).

Das BIP liegt darüber hinaus in verschiedenen Sprachen, wie z.B. Englisch, Spanisch, Portugiesisch und Dänisch, vor und ist somit international einsetzbar.

Die Autoren empfehlen das BIP speziell für die Anwendung in der beruflichen Eignungsdiagnostik, für die Berufs- und Karriereberatung sowie für das Training bzw. Coaching von Berufstätigen. Auch in Feedbackprozessen könne man das BIP einsetzen.

In der Untersuchung von Hagemeister und Kollegen (2010), welche psychologischen Testverfahren in Deutschland am häufigsten verwendet werden, wurde das BIP als das am zweithäufigsten verwendete Verfahren genannt.

Allerdings ist das BIP mit einer Bearbeitungszeit von 45 Minuten bzw. 251 zu beantwortenden Items ein vergleichsweise aufwendiges Verfahren. Darüber hinaus liegt dem Fragebogen kein theoretischer Ansatz, sondern lediglich eine sogenannte Expertenbefragung zugrunde. Es ist möglich, dass andere Praktiker andere relevante Eigenschaften genannt hätten und es andere Dimensionen bzw. Eigenschaften als relevante Kriterien hätten identifiziert werden können. Laut Hülshager und Maier (2008) erklärte das berufsbezogen formulierte BIP wider Erwarten Berufserfolgskriterien nicht besser vorher als die neutral formulierten Skalen des NEO-PI-R. Die Autoren berichten im Gegenteil dazu, dass unter Hinzunahme der Facetten des NEO-PI-R diese sogar einen inkrementellen Beitrag über das BIP hinaus zur Varianzaufklärung hinsichtlich Berufserfolgskriterien leisteten. Als Ursache hierfür sahen Hülshager und Maier (2008) die mangelnde Konstruktvalidität des BIP an, dem keine eindeutige Persönlichkeitstheorie zugrunde liegt. Die 14 Dimensionen der zweiten Auflage des BIP würden inhaltlich zudem stark überlappen.

2.6.5 Freiburger Persönlichkeitsinventar (FPI)

Das Freiburger Persönlichkeitsinventar (FPI) von Fahrenberg, Hampel und Selg (2010) ist ein mehrdimensionaler Persönlichkeitstest. Er wurde in den 1960er Jahren in Deutschland entwickelt und gilt als das erste deutschsprachige Persönlichkeitsinventar. Das Verfahren basiert dabei nicht auf einer bestimmten Persönlichkeitstheorie, sondern die einzelnen Eigenschaftskonzepte wurden „nach den Interessen der Autoren mit Ergänzungen und Erweiterungen im Zuge dieser Forschungsvorhaben“ entwickelt (Fahrenberg, Hampel & Selg, 2001, S. 15). Auch eine statistische Datenreduktion, wie z.B. eine Faktorenanalyse, wurde nicht zur Skalenkonstruktion durchgeführt. Zwar wurden bei der Testentwicklung Faktoren-, Item- und Clusteranalysen verwendet, diese waren jedoch nur Hilfsmittel, um die Prägnanz der Skalentwürfe und theoretischen Konzepte zu verbessern.

Die Autoren wählten die Dimensionen des Fragebogens aufgrund ihrer Erfahrungen aus und reicherten diese mit in der Literatur als interessant und allgemein wichtig erscheinenden Persönlichkeitsbereichen an. Zu diesen Bereichen zählten *Emotionalität*, *Beanspruchung (Stress)*, *körperliche Beschwerden* bzw. *Gesundheitssorgen*, *Aggressivität* und *prosoziales Verhalten*. Diese Bereiche nahmen die Autoren in die Fragebogenkonstruktion auf und erhielten durch Faktorenanalysen die zehn Skalen *Lebenszufriedenheit*, *Soziale Orientierung*, *Leistungsorientierung*, *Gehemmtheit*, *Erregbarkeit*, *Aggressivität*, *Beanspruchung*, *Körperliche Beschwerden*, *Gesundheitssorgen* und *Offenheit*. Die zehn Skalen werden mit jeweils zwölf Items erfasst. Die zehn Skalen ergänzten die Autoren um die sogenannten Sekundärfaktoren *Extraversion* und *Emotionalität*, die jeweils mit 14 Items erfasst werden. Es ergeben sich somit 138 Items für den FPI, denen ein bipolares Antwortformat („stimmt“ vs. „stimmt nicht“) zugeordnet ist. Der FPI liegt als Selbsteinschätzungsinstrument vor, dessen Beantwortung laut Manual ca. 20 bis 30 Minuten dauert.

Die letzte Normierung des FPI wurde 1999 vorgenommen, da die ursprüngliche Version nur Daten von Probanden aus Westdeutschland enthielt. Die neue Normierungsstichprobe ($N=3.740$) war eine bevölkerungsrepräsentative Stichprobe, die sowohl Ost- als auch Westdeutsche berücksichtigte. Es gibt Normtabellen getrennt nach dem Geschlecht der Probanden sowie für sieben verschiedene Altersgruppen. Die Autoren empfehlen einen Einsatz des Verfahrens für Jugendliche ab 16 Jahren und Erwachsene.

Die interne Konsistenz der einzelnen Skalen liegt zwischen $\alpha=.73$ und $.83$ für alle Skalen, was als befriedigend bzw. gut zu interpretieren ist. Die Retest-Reliabilität für alle Skalen nach

einem vierwöchigen Zeitintervall ist mit $r = .57$ bis $.85$ nicht ausreichend bis gut (Fahrenberg et al., 2010). Die Konstruktvalidität wurde u.a. durch die konfirmatorische Faktorenanalysen belegt. Auch in neueren Normierungsverfahren konnte die Skalenstruktur immer wieder nachgewiesen werden. Die Autoren berichten zahlreiche Korrelationen der Skalen mit soziodemografischen Merkmalen der Probandengruppen sowie Interkorrelationen der Skalen. Auch Mittelwertunterschiede zwischen Patientengruppen und anderen Vergleichsgruppen werden berichtet. Zahlreiche Arbeiten belegen zudem die Kriteriumsvalidität des FPI.

Für die Verwendung des FPI spricht neben seiner langen Tradition, dass die Skalen des FPI gut nachvollziehbar sind und den psychologischen Selbstbeschreibungen der Durchschnittsbevölkerung entsprechen. Im Vergleich mit anderen Persönlichkeitsinventaren ist das FPI zudem schnell durchzuführen und auszuwerten. Fragebogenversionen in anderen Sprachen sind verfügbar.

Allerdings fehlt es dem FPI an einem theoretischen Konstrukt, das den Skalen zugrunde liegt. Außerdem wurde das Verfahren ursprünglich für den klinischen Kontext entwickelt. Viele Skalen beziehen sich daher auf gesundheitsrelevante Themen. Auch wenn den Autoren zufolge die Globaleinschätzung der Persönlichkeit durch den FPI nicht nur im (klinischen) Alltag, der psychologischen Forschung und der Rehabilitation, sondern auch in Beruf und Ausbildung verwendet werden kann, fehlt dennoch ein konkreter Arbeitsbezug in den Skalen. Der Anwendungsschwerpunkt des FPI liegt entsprechend in den eher klinischen Bereichen der psychologischen Diagnostik, wie der Psychosomatik und -therapie, Rehabilitation und der Gesundheitspsychologie. Auch in der Forschung wird das FPI häufig verwendet. In der Untersuchung zur Verwendungshäufigkeit von Persönlichkeitstests von Hagemann und Kollegen (2010) belegte das FPI den ersten Platz in deren Ranking. Allerdings bestand die von den Autoren befragte Stichprobe auch aus mehr klinischen Psychologen ($N = 95$) als Wirtschaftspsychologen ($N = 66$), weshalb hier von einer leichten Verzerrung hin zu klinischen Verfahren auszugehen ist. Darüber hinaus differenziert das FPI die Persönlichkeitsdimensionen nicht in untergeordnete Facetten.

2.6.6 Big Five Inventory (BFI)

Das Big Five Inventory (BFI) ist ein Selbstbeschreibungsverfahren von Oliver John, das die Big 5-Dimensionen möglichst kurz und präzise erfasst. Ziel des BFI ist eine möglichst robuste und reliable Erfassung der Grobstruktur der Big 5-Dimensionen *Neurotizismus*, *Extraversion*, *Offenheit*, *Verträglichkeit* und *Gewissenhaftigkeit*. Hierzu formulierte der Wissenschaftler kurze, leicht verständliche Aussagen, die er in einem vergleichsweise kurzen multidimensionalen Fragebogen mit nur 44 Items zusammenfasste (John & Srivastava, 1999). Die Dimensionen Neurotizismus, Extraversion, Verträglichkeit und Gewissenhaftigkeit werden mit jeweils vier Items erfasst, Offenheit mit fünf Items. Als Antwortformat dient eine 5-stufige Likertskala von „*strongly disagree*“ zu „*strongly agree*“.

Zudem gibt es eine Kurzversion (K-BFI), die mit nur elf Items ebenfalls die Big 5-Dimensionen erfasst. Das K-BFI wurde mit dem Ziel entwickelt, mit einer Bearbeitungsdauer von unter zwei Minuten extrem ökonomisch zu sein. Die Ergebnisse belegen zufriedenstellende psychometrische Kennwerte für die Skala. Neben ausreichenden Reliabilitäten konnten sowohl die faktorielle Validität des Verfahrens als auch hohe Übereinstimmungen mit Bekanntenurteilen und mit anderen etablierten Verfahren (z.B. NEO-PI-R) gezeigt werden (Rammstedt & John, 2005).

Das BFI wurde ursprünglich in den USA entwickelt, liegt jedoch in verschiedenen anderen Sprachen vor. Im deutschsprachigen Raum wird die übersetzte Version von Lang, Lüdtke und Asendorpf (2001) verwendet. Die deutsche Adaptation des BFI umfasst 45 Items, die ebenfalls auf einer 5-stufigen Likertskala von „*sehr unzutreffend*“ bis „*sehr zutreffend*“ zu beant-

worten sind. In der deutschen Version wurde ein zusätzliches Item zur Skala Verträglichkeit hinzugefügt. Studien zur Übereinstimmung der deutschen Adaptation und amerikanischen Originalversion (Rammstedt & John, 2005) deuten auf eine hohe Vergleichbarkeit der psychometrischen Kennwerte als auch hinsichtlich des nomologischen Netzwerks hin. Lang et al. (2001) veröffentlichten hierzu auch Vergleichswerte einer altersheterogenen Normstichprobe ($N=480$). Bis zu diesem Zeitpunkt lagen noch keine veröffentlichten Daten zu einer Normstichprobe des BFI vor. Auch die Autoren der Originalversion haben bis dato keine Vergleichswerte veröffentlicht.

Erste Daten zur Reliabilität sowie konvergenten und divergenten Konstruktvalidität der deutschsprachigen Testversion des BFI sowie K-BFI wurden von Lang und Lüdtkke (2005) veröffentlicht. Ihren Daten zufolge erwies sich das BFI als zuverlässiges und effizientes Instrument zur Erfassung der Big 5 auch in heterogenen Populationen. Tabelle 2-11 enthält

Tabelle 2-11: Kennwerte der Skalen des BFI und K-BFI (nach Lang & Lüdtkke, 2005)

Instrument	Skala	<i>N</i>	<i>M</i>	<i>SD</i>	α (korr.)
BFI	Neurotizismus	1452	2,54	.60	.74
	Offenheit	1447	3,20	.66	.82
	Extraversion	1450	3,37	.67	.81
	Verträglichkeit	1447	3,68	.60	.76
	Gewissenhaftigkeit	1445	3,81	.61	.81
K-BFI	Neurotizismus	2495	1,97	.99	.60 (.65)
	Offenheit	2496	3,30	.99	.52 (.62)
	Extraversion	2499	3,32	1.07	.56 (.65)
	Verträglichkeit	2490	4,19	.97	.65 (.68)
	Gewissenhaftigkeit	2493	4,46	.97	.59 (.62)

Anmerkungen: *N*= Stichprobengröße; *M*= Mittelwert; *SD*= Standardabweichung; α = Cronbachs Alpha für standardisierte Items; (korr.)= Cronbachs Alpha nach Korrektur der Items auf Akquieszenz (in Klammern).

einen Ausschnitt der Angaben zu Mittelwerten, Standardabweichungen sowie der internen Konsistenz des deutschsprachigen BFI und K-BFI. In dem Kapitel von Lang und Lüdtkke (2005) sind auch Angaben zur Konstruktvalidität des BFI bzw. K-BFI mit dem NEO-FFI enthalten.

Die Ergebnisse zeigen befriedigende bis gute interne Konsistenzen des BFI, während die Reliabilität der Kurzversion geringer ausfällt ($\alpha = .52$ bis $.65$ bzw. $\alpha_{korr} = .62$ bis $.68$). Lang und Lüdtkke (2005) berichten darüber hinaus, dass bei älteren Erwachsenen über einen Zeitraum von viereinhalb Monaten hinweg eine befriedigende Stabilität nachgewiesen werden konnte. Rammstedt und John (2005) geben für die Gewissenhaftigkeitsskala eine interne Konsistenz von $\alpha = .70$ und eine Retest-Reliabilität von $r = .85$ an.

Die Bearbeitungszeit geben die Autoren mit fünf Minuten für die 44-Item-Version an, was im Vergleich mit anderen Maßen zur Erfassung der Big 5 recht effektiv ist. Inwiefern diese Werte der Realität entsprechen, sollte vor der Durchführung des Verfahrens einmal überprüft werden. Eine noch schnellere Erfassung der Big 5 ermöglicht die Kurzversion des BFI. Allerdings sind die Skalen des BFI und vor allem des K-BFI aufgrund der wenigen Items weniger reliabel als vergleichbare, längere Verfahren zur Erfassung der Grobstruktur der Big 5, wie z.B. der NEO-FFI. Allerdings übertrafen die Alpha-Koeffizienten in allen Fällen die nach der Spearman-Brown-Formel für Testverkürzungen geschätzten Werte (Rammstedt & John, 2005). Den Autoren zufolge ist das BFI reliabel und valide genug, um die fünf Faktoren der Persönlichkeit zu erfassen. Lang und Lüdtkke (2005) empfehlen, die beiden Fragebogenver-

sionen eher dann einzusetzen, wenn nur sehr wenig Zeit für die Erfassung der Big 5-Dimensionen zur Verfügung steht. Auch die Items des BFI bzw. K-BFI sind wiederum nur allgemein gehalten und lassen einen Kontextbezug vermissen. Verlässliche Daten zur Kriteriumsvalidität, vor allem in eignungsdiagnostischen Untersuchungen, sind bislang noch nicht in ausreichendem Maße veröffentlicht.

2.6.7 START-P

Der START-P gehört in die Testbatterie für Berufseinsteiger (START), die von Detlev Liepmann herausgegeben wird. Diese Testbatterie erfasst grundlegende Kompetenzen, die beim Berufseinstieg relevant sind, wie z.B. Konzentrationsfähigkeit, EDV-Kenntnisse und den Fremdsprachenerwerb (Englisch) betreffend. Weitere Module sind derzeit in der Entwicklung. Hintergrund der Entwicklung dieser Testbatterie ist der Bedarf von Maßen, die sich explizit an den Fähigkeiten, Eigenschaften, Anforderungen an und Bedürfnissen von Jugendlichen und jungen Erwachsenen in der Berufsstartphase orientiert. Bislang haben sich Forscher bei der Entwicklung von Testverfahren implizit oder explizit an älteren Personen orientiert. Berufsbezogen konstruierte Tests setzen teilweise Berufserfahrung voraus, um Aufgaben bzw. Items adäquat bearbeiten zu können. Über solches Wissen verfügen Berufsanfänger in der Regel jedoch nicht (Beauducel & Kersting, 2010). Die START-Testbatterie ist entsprechend auf den Altersbereich von Ausbildungsbeginnern (19-27 Jahre) ausgerichtet. Für jeden Test der Batterie sind Normdaten für verschiedene Altersgruppen vorhanden. Unterschieden wird auch nach Schulart.

Der START-P ist folglich ein Fragebogen zur Beschreibung berufsbezogener Persönlichkeitsaspekte von Jugendlichen und jungen Erwachsenen. Beauducel und Kersting (2010) verwenden im START-P explizit situative Komponenten, die beispielsweise in Berufsschulsituationen und der praktischen Lehre relevant sind, um die allgemeinen Verhaltenstendenzen einer Person in der Ausbildung erfassen bzw. prognostizieren zu können. Hierzu integrierten die Forscher auch die Belohnungs- bzw. Bestrafungssensitivität der Berufsanfänger, da dieses den Autoren zufolge subjektiv relevant in Ausbildungs- und Lernsituationen ist. Dieses Konzept wurde in der Persönlichkeitstestung bislang zudem nur wenig berücksichtigt. Der START-P stellt somit das erste persönlichkeitsdiagnostische Verfahren dar, das Belohnungs- und Bestrafungssensitivität im Ausbildungs- und Arbeitskontext erfasst.

Der START-P lässt sich in zwei Module untergliedern: *Verhaltens-Input* und *Verhaltens-Output*. Dem Modul Verhaltens-Input sind die vier Dimensionen *Belohnungs- und Bestrafungssensitivität*, *Sensitivität gegenüber Vorgesetzten und Ausbildern* sowie *Sensitivität gegenüber dem Team* zugeordnet. Das Modul beschreibt den Autoren zufolge eher die Reaktionen einer Person auf eingehende Reize und situative Bedingungen. Dem Modul Verhaltens-Output werden die sechs Dimensionen *Kontaktorientierung*, *Durchsetzungsvermögen*, *Sorgfalt*, *Emotionale Stabilität*, *Empathie* und *Leistungsmotivation* zugeordnet. Die letzten drei Dimensionen stellen auch die Dimensionen des *Kurz-Moduls Verhaltens-Output* dar. Unter Verhaltens-Output fassen die Autoren selbstgesteuerte und spontane Verhaltenstendenzen zusammen. Zur Erstellung dieses Moduls orientierten sie sich an den Metaanalysen zum Zusammenhang von Persönlichkeit und Berufserfolg (vgl. Abschnitt 2.3) und Zufriedenheit im Beruf.

Aus den eben genannten zehn Dimensionen (*Grunddimensionen*) können wiederum drei *Globaldimensionen* (*Belastbarkeit*, *Soziale Orientierung* und *Gewissenhaftigkeit*) gebildet werden. Abbildung 2-3 veranschaulicht die Dimensionsstruktur des START-P.

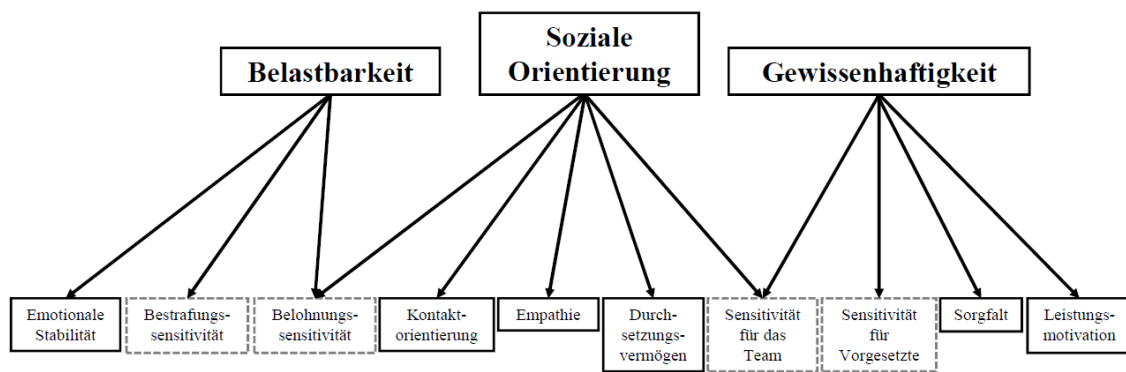


Abbildung 2-3: Grund- und Globaldimensionen des START-P (nach Beauducel & Kersting, 2010)

In der Abbildung 2-3 sind die vier Grunddimensionen des Moduls Verhaltes-Input in den grau-gestrichelten Kästchen enthalten. Die sechs Grunddimensionen des Moduls Verhaltens-Output sind in den schwarz umrandeten Kästchen enthalten.

Normiert wurde der START-P anhand einer Stichprobe von 2.063 Probanden. Im Manual sind alters-, geschlechts- und bildungsspezifische Normtabellen enthalten. Eine Besonderheit stellt eine Norm zur Korrektur potenziell sozial erwünschter Antworttendenzen in Bewertungskontexten dar.

Beauducel und Kersting (2010) berechneten zur Bestimmung der Reliabilität des START-P Split-Half-Reliabilitäten. Für die zehn Grunddimensionen erhielten sie befriedigende bis sehr gute Reliabilitätskennwerte ($r = .70$ bis $.94$). Die drei Globaldimensionen weisen ihren Ergebnissen zufolge durchweg sehr gute Reliabilitäten auf ($r = .94$ bis $.97$). Die Konstruktvalidität sowohl der zehn Grunddimensionen als auch der drei Globaldimensionen konnte den Autoren zufolge in Strukturanalysen nachgewiesen werden. So liegen die Korrelationen der verschiedenen Dimensionen mit verwandten Konstrukten aus z.B. dem NEO-FFI oder BFI zwischen $r = .50$ bis $.78$. Signifikante Zusammenhänge berichten Beauducel und Kersting (2010) auch mit relevanten Kriteriumsvariablen, wie z.B. Schulnoten oder nicht-kognitiven Aspekten in Schule und Studium.

Die Durchführung des START-P dauert laut Manual – je nach Modul – ca. zehn bis 40 Minuten. Je nachdem, welche Module also durchgeführt werden, handelt es sich beim START-P um ein mehr oder weniger ökonomisches und effizientes Verfahren. Ein Vorteil des Verfahrens ist, dass es eine anwendungsbezogene und zielgruppenspezifische Erfassung relevanter Persönlichkeitsmerkmale anstrebt. Allerdings ist es ein noch recht neues Verfahren. Unabhängige Untersuchungen zur Kriteriumsvalidität der Dimensionen lagen bis zur Abgabe dieser Arbeit noch nicht vor. Zukünftige Forschung sollte sich der unabhängigen Überprüfung der Vorhersagekraft des START-P widmen.

2.6.8 Zusammenfassung und Fazit zu den Gewissenhaftigkeitsmaßen

Wie zu Beginn dieses Abschnitts in 2.6 bereits beschrieben, gibt es eine Vielzahl an Persönlichkeitstests, die Gewissenhaftigkeit als Dimension und/ oder Facette erfassen. Die in den Abschnitten 2.6.1 bis 2.6.7 beschriebenen Testverfahren stellen lediglich einen Ausschnitt der verfügbaren Testverfahren zur Erfassung von Gewissenhaftigkeit dar.

Weitere Verfahren zur Erfassung von Gewissenhaftigkeit liegen vor, die jedoch weniger bekannt und erforscht sind. So stellten Howard und Howard (2001) mit ihrem Workplace Big Five (WB5) ein computerbasiertes Persönlichkeitsinventar vor, das auf Situationen und Ver-

halten bei der Arbeit bezogen ist. Es basiert theoretisch auf dem NEO-PI-R (Costa & McCrae, 1992) und enthält 114 Items zur Erfassung der Big 5-Dimensionen. Die Gewissenhaftigkeitskala ($\alpha = .89$) des WB5 enthält insgesamt 30 Items, die gleichmäßig mit jeweils sechs Items auf die fünf Subskalen *Perfektionismus* ($\alpha = .80$), *Organisation* ($\alpha = .81$), *Antrieb* ($\alpha = .68$), *Konzentration* ($\alpha = .74$) und *Methodisches Vorgehen* ($\alpha = .73$) verteilt sind. Allerdings liegt dieses Verfahren nur als PC-gestützte Version in englischer Sprache vor.

Auch das Abridged Big-Five Circumplex (AB5C; Goldberg, 1999) erfasst unter anderem *Industriousness*, *Orderliness* und *Cautiousness* als Gewissenhaftigkeitsskalen. Weitere Gewissenhaftigkeitsmaße werden von Jackson et al. (2010) beschrieben, die sich mit der Substruktur von Gewissenhaftigkeit auseinander gesetzt haben.

Feltham und Woods (1995) entwickelten für den Kontext eignungsdiagnostischer Untersuchungen von Berufsanfängern den Business Personality Indicator (BPI). Dieser besteht aus 128 dichotom zu beantwortenden Items, die elf Primärskalen bzw. fünf Faktoren zweiter Ordnung (*Dynamic*, *Extravert*, *Work Stamina*, *Worrying* und *Controlled*) zugeordnet werden. Auch im BPI sind einige Skalen enthalten, die eine hohe Übereinstimmung mit Gewissenhaftigkeit aufweisen: *Change Orientation*, *Perfectionism*, *Time Management* und *Work Orientation*.

Ein Verfahren, das von einem Unternehmen entwickelt wurde, ist der Occupational Personality Questionnaire (OPQ) von SHL (2012). Dieser wurde von der Unternehmensberatung für den Kontext der beruflichen Eignungsdiagnostik entwickelt und ist in verschiedenen Versionen verfügbar, u.a. auch für Studenten bzw. Bewerber mit wenig bis gar keiner Berufserfahrung. Der OPQ enthält sechs bis 32 Primärskalen (z.B. *Images*, *Customer Contact*, *Factor*, *Work Styles*), die mit 136 normativen Items auf einer 5-stufigen Likertskala erfasst werden. Einige der Primärskalen, wie z.B. *Achieving*, *Decisive* und *Methodical*, scheinen ähnlich der Gewissenhaftigkeitsdimension zu sein. Allerdings sind die Items nicht frei zugänglich und die Auswertung ist an das Unternehmen gebunden. Zudem sind Studien zur Reliabilität und Validität des OPQ ebenfalls nur schwer zugänglich. Als Testautor und einziger Vertriebsweg erscheint das Unternehmen, das bei Fragen zum Verfahren kontaktiert werden muss.

Alle (verfügbaren) Verfahren zur Erfassung von Gewissenhaftigkeit weisen somit verschiedene Mängel auf. So sind die meisten der existierenden und gut validierten Verfahren in der Regel nicht arbeitsbezogen formuliert (s. Abschnitt 2.5) und/ oder sie beruhen nicht auf einem theoretischen Modell. Einige Verfahren, die diese Merkmale aufweisen, gibt es nicht in deutscher Sprache oder es werden zu viele zusätzliche, für die Zielsetzung dieser Arbeit nicht interessierende Konstrukte oder Variablen mit erfasst. Oftmals sind diese Verfahren dann auch zu lang, d.h. es sind allgemein zu viele Items zu beantworten.

Ein berufsbezogen formuliertes Persönlichkeitsinventar, das lediglich Gewissenhaftigkeit bzw. dessen Aspekte Fleiß und Ordnung berufsbezogen erfasst und gleichermaßen für die Anwendung in jungen Stichproben mit wenig bis keiner beruflicher Vorerfahrung sowie die eignungsdiagnostische Untersuchung von Arbeitnehmern geeignet ist, liegt derzeit noch nicht vor. Die arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS; Dreier & Moldzio, 2010) könnten ein Ansatz sein, diese Lücke zu schließen.

2.7 Leistungs- und Erfolgskriterien in der Eignungsdiagnostik

Leistung und *Erfolg* sind hypothetische Konstrukte, die als solche nicht eindeutig beobachtbar oder messbar sind. Durch verschiedene Kriterien wird jedoch versucht, diese Konstrukte abzubilden bzw. zu operationalisieren. Diese Kriterien können – je nach Kontext – sehr unterschiedlich sein und werden allgemein den Ebenen *Potenzial*, *Verhalten* und *Ergebnis* zugeschrieben (Marcus & Schuler, 2006). Den Autoren zufolge handelt es sich bei Leistung und

Erfolg um eher einfache Konstrukte, die über beispielsweise verschiedene Arbeitsplätze hinweg oftmals sehr stark generalisierbar sind und deren Komplexität häufig überschätzt werde.

In der Regel wird *Erfolg* als Klammerbegriff für berufliche bzw. schulische Leistung verwendet. In der Theorie wird jedoch allgemein zwischen *Leistung* (*performance*) und *Erfolg* (*success*) unterschieden (Sutin, Costa, Miech & Eaton, 2009). Die beiden Konzepte stehen in Beziehung zueinander ($r = .40$), sind jedoch konzeptuell verschieden (Dudley et al., 2006; Marcus & Schuler, 2006).

Angelehnt an Hülshager und Maier (2008) kann man *Leistung* als Effektivität bei der Aufgaben- bzw. Arbeitserfüllung an einem bestimmten Platz bzw. in einer bestimmten Situation beschreiben. Leistung bzw. extrinsischer Erfolg bezieht sich auf Ergebnisse und Belohnungen, die eine Person während einer Tätigkeit ansammelt. Diese Ergebnisse sind in der Regel messbar und unabhängig von einem Beobachter. Im beruflichen Kontext sind objektive Leistungskriterien beispielsweise Stückzahlen bei Akkordarbeitern. Allerdings ist es schwer, außerhalb des produzierenden Gewerbes objektive Leistungsmaße zu finden. Ein Beispiel für ein solches Kriterium sind beispielsweise Fehlzeiten von Mitarbeitern, die jedoch nur einen kleinen Teilaspekt von Leistung erfassen und somit als Leistungsbeurteilungen nicht ausreichend sind. Teilweise sind diese Kriterien auch durch äußere Umstände außerhalb der Person, wie z.B. Defekte an Maschinen, beeinflussbar, so dass eine korrekte Leistungserfassung beeinträchtigt sein kann. Im schulischen Bereich gelten die Schulnoten als klassisches Leistungsmaß für Schüler.

Rode, Arthaud-Day, Mooney, Near und Baldwin (2008) definieren *beruflichen Erfolg* als die Akkumulation positiver, berufsbezogener Ergebnisse bzw. als Ergebnis der Arbeitserfahrungen einer Person. Dieser kann extrinsisch über objektive Kriterien wie z.B. die Gehaltshöhe eines Arbeitnehmers und intrinsisch beispielsweise mittels subjektiver Karrierebewertungen messbar gemacht werden. Dies geschieht im beruflichen Bereich meist durch einen Vorgesetzten oder Ausbildungsleiter, die Kollegen, den Arbeitnehmer selbst oder teilweise auch durch die Kunden einer Person. Selbst- und Vorgesetztenbeurteilungen sind die am häufigsten verwendeten Beurteilungsquellen (Marcus & Schuler, 2006). Im schulischen Bereich nehmen in der Regel die Lehrer diese Beurteilungsfunktion ein. Die Einschätzungen beziehen sich meist auf die subjektive Beurteilung des bisherigen Werdegangs einer Person vor dem Hintergrund persönlicher Erwartungen und Ziele, beispielsweise die Zufriedenheit mit der (eigenen) Schullaufbahn.

Zur Leistungsbeurteilung kann man verschiedene Verfahren anwenden. Bei *Einstufungsverfahren* werden Personen zunächst unabhängig voneinander auf metrischen Skalen eingeschätzt und anschließend miteinander verglichen. In der Praxis am stärksten verbreitet sind grafische oder verhaltensverankerte Einstufungsskalen. Bei *Rangordnungsverfahren* werden die Mitglieder eines Teams bzw. die Schüler einer Klasse direkt miteinander verglichen. Das Ausmaß der Leistungsunterschiede wird hierbei jedoch nicht erfasst. Man kann diesbezüglich direkte Rangreihen, Paarvergleiche und Verhaltensrangprofile voneinander unterscheiden. *Auswahl-* und *Kennzeichnungsverfahren* verlangen die Zustimmung oder Ablehnung bestimmter Aussagen, deren Wertigkeit zuvor bestimmt wurde und den Beurteilern selbst unbekannt ist. Bei *Zielerreichungsverfahren* wird der Erreichungsgrad zuvor vereinbarter Ziele beurteilt (Marcus & Schuler, 2006).

Zudem unterscheidet Schuler (2004) drei Beurteilungsebenen. Auf der ersten Ebene, dem *Day-to-day-Feedback*, finden sich Gespräche zur Verhaltenssteuerung und Unterstützung des Lernens zwischen Mitarbeiter und Vorgesetztem bzw. Schüler und Lehrer. Die zweite Ebene bilden die *systematischen Beurteilungen* bzw. *Beurteilungsgespräche* als Regelbeurteilung zur Leistungseinschätzung und Zielsetzung. Dies sind im organisationalen Kontext die (i.d.R. jährlich stattfindenden) Mitarbeitergespräche und im Schulkontext beispielsweise die Eltern-

sprechtag. Auf der dritten Ebene finden sich die sogenannten *Potenzialbeurteilungen* zur Fähigkeitseinschätzung und Leistungsprognose mittels eignungsdiagnostischer Methoden, wie z.B. einem Assessment Center. Diese Beurteilungsebene lässt sich meist nur in Unternehmen, nicht aber im schulischen Kontext wiederfinden.

Ein Vorteil von subjektiven Beurteilungen gegenüber objektiven Leistungsindikatoren ist, dass menschliche Beurteiler in der Regel in der Lage sind, situative Einflüsse in ihre Beurteilung der Leistung einer Person mit einzubeziehen. Diese „weichen“ Qualitätsmaßstäbe sind in der Praxis der Leistungsbeurteilung oftmals wichtiger und auch bei allen Beteiligten beliebter als objektive Maße bzw. technische Gütekriterien (Marcus & Schuler, 2006).

Allerdings sind subjektive Beurteilungen auch anfällig für Fehler. So haben nicht alle Personen, die einen Mitarbeiter oder Schüler einschätzen sollen, ausreichend Gelegenheit, diesen zu beobachten, was zu unkorrekten Beurteilungen führen kann. Auch eine mangelnde Fähigkeit bzw. mangelnder Wille zur Abgabe zutreffender Einschätzungen kann die Qualität von subjektiven Beurteilungen verringern. So verzerren *Urteilstendenzen*, d.h. „Abweichungen der Leistungsbeurteilung von einer theoretisch erwarteten Verteilung bzgl. Mittelwert, Streuung und Interkorrelation der Urteilsdimensionen“ (Marcus & Schuler, S. 455), oftmals die Beurteilungen. Obwohl vereinzelt auch Abweichungen nach unten auftreten (= *Strengetendenz*), sind die meisten Beurteilungen in der Praxis eher nach oben verlagert (= *Mildetendenz*), was auch als „beschönigende Urteile“ kritisiert wird. Auch eine geringe Streuung zwischen den Beurteilungen wird oft als Fehler interpretiert. Allerdings ist es möglich, dass dies auch einfach nur Leistungshomogenität innerhalb eines Unternehmens, eines Teams etc. widerspiegelt. Dies widerspricht dann allerdings der Forschung zur Variabilität individueller Leistung. Marcus und Schuler (2006) zufolge fallen für Forschungszwecke abgegebene Urteile allgemein korrekter aus als solche mit administrativen Folgen. Ein Problem bei Vorgesetztenbeurteilungen ist zudem, dass sie bei der Beurteilung von Arbeitsverhalten stark auch kontextuelles Arbeitsverhalten, d.h. das allgemeine Engagement eines Mitarbeiters über dessen eigentlichen Aufgaben hinaus, gewichten und nicht nur den reinen Output bzw. die Aufgabenerledigung beurteilen (Schmidt et al., 2008). Dies ist auch ein Grund, warum – auch nach angemessener Korrektur der Range Restriction (s. Abschnitt 6.3.2.1) – die subjektiven Beurteilungen nicht genau mit den objektiv gemessenen Arbeitsergebnissen korrelieren. Zusammenfassend muss gesagt werden, dass Vorgesetztenbeurteilungen höchst subjektiv sind.

Campbell und Kollegen (1993, zitiert nach Marcus & Schuler, 2006) entwickelten eine generelle Theorie beruflicher Leistung, die Leistung durch die Interaktion der drei Determinanten *deklaratives (statisches) Wissen* (DK), *prozedurales Wissen und Fertigkeiten* (PKS) sowie *Motivation* (M) bestimmt. Diese Determinanten sind multiplikativ verknüpft zu einem *Leistungsindikator* (PC).

$$PC = DK \cdot PKS \cdot M$$

Keine der Determinanten darf dieser Funktion zufolge einen Wert gleich null annehmen, damit Leistung zustande kommt. Allerdings gibt es zur Überprüfung dieser Theorie bisher kaum Studien (Marcus & Schuler, 2006).

Eine stärker auf den Inhalt beruflicher Leistung bezogene Theorie stammt von Borman und Motowidlo (1993). Kern der Theorie ist der Gedanke, dass neben der reinen Aufgabenerfüllung, die die Autoren als *aufgabenbezogene Leistung* bezeichnen, auch weitere Aspekte, die darüber hinausgehen und in Tätigkeitsanalysen oft übersehen werden, als *umfeldbezogene Leistung* zusammengefasst werden können. Hierzu zählen Tätigkeiten, die über die Aufgabenerfüllung und formalen Arbeitsabläufe hinausgehen, und indirekt ergebnisunterstützend sind, wie z.B. Commitment. Empirische Untersuchungen stützen diese Theorie (Marcus & Schuler, 2006).

Da sich die Leistungs- und Erfolgsmaße bzw. -definitionen je nach Kontext in Unternehmen, Schulen und der Berufsausbildung allerdings voneinander unterscheiden, werden diese in den folgenden Abschnitten noch einmal getrennt voneinander behandelt. Zuerst werden berufliche Leistungskriterien (Abschnitt 2.7.1) definiert und erläutert sowie Studienergebnisse hierzu berichtet. Da diese Arbeit vorrangig zum Ziel hat, Ausbildungserfolg bzw. schulische Leistungskriterien vorherzusagen, wird dieser Abschnitt aus ökonomischen Gründen jedoch eher kurz gehalten. In den nachfolgenden beiden Abschnitten wird danach differenzierter auf Leistungskriterien in der Schule (Abschnitt 2.7.2) und Berufsschule (Abschnitt 2.7.3) eingegangen.

2.7.1 Leistung im Beruf

Campbell et al. (1993) haben *Arbeitsleistung* als individuelles Verhalten definiert, das mit den organisationalen Zielen im Einklang ist. Leistung im Beruf gilt als Indikator, wie gut jemand seine Arbeit bzw. beruflichen Aufgaben erfüllt, eigeninitiativ vorgeht, Kreativität beim Lösen von Problemen zeigt und ressourcenschonend mit dem Material, der Zeit und der Energie eines Unternehmens umgeht (Sutherland et al., 2007). Die Autoren teilen berufliche Leistung in zwei konzeptuell unterscheidbare Komponenten auf.

Die erste Komponente ist die *Aufgabenleistung*, die sich auf die Erfüllung bzw. Leistung von Mitarbeitern in Bezug auf ihre vorgeschriebenen Tätigkeiten und Aufgaben bezieht. Diese unterscheidet sich je nach Beruf bzw. ausgeübter Tätigkeit stark und kann durch verschiedenste Leistungsmaße erfasst werden. Aufgabenleistung wird am besten durch geistige Fähigkeiten vorhergesagt.

Die zweite Komponente ist die *kontextuelle Leistung*, die durch Aktivitäten gekennzeichnet wird, die ein Mitarbeiter laut Arbeitsvertrag bzw. Tätigkeitsbeschreibung nicht unbedingt ausführen muss, die aber dennoch notwendig sind, damit ein Unternehmen seine Ziele erreicht. Diese Aktivitäten werden nicht unbedingt durch typische Leistungsmaße erfasst. Verhaltensweisen, die mit kontextueller Leistung erfasst werden, sind tendenziell eher diskret in ihrer Natur und beinhalten freiwillige Tätigkeiten, die formell nicht Teil der Arbeit sind, aber Einsatzbereitschaft, Mithilfe und Kooperation mit anderen signalisieren, indem organisationale Regeln eingehalten und Ziele unterstützt werden. Diese Tätigkeiten können auch als Extrarollenverhalten benannt werden. Es sind vermutlich diese Tätigkeiten, die die Basis dafür bilden, welche Mitarbeiter in einer Organisation ge- und befördert werden. Kontextuelle Leistung wird am besten durch dispositionelle Eigenschaften vorhergesagt. Hierbei erwies sich Gewissenhaftigkeit als der beste Trait-Prädiktor von kontextuellem Verhalten bei der Arbeit (Barrick & Mount, 1991).

Die Vorhersage von Arbeitsleistung war und ist eines der am meisten beforschten Themen in der Arbeits- und Organisationspsychologie. Schmidt und Hunter (1998) führten die erste Metaanalyse zum Vergleich der Vorhersagekraft von 19 verschiedenen Auswahl- und Assessmentmethoden durch. Sie fassten Studien aus 85 Jahren Forschung zur Vorhersage von Leistung zusammen. Als Kriteriumsvariablen wurden den Autoren zufolge in der Forschung bis dato vor allem der Geldwert von Leistung (*dollar value of output*) und anteilige Leistung (*output as percentage of mean output*) verwendet. Berufserfolg wurde in der Regel durch Vorgesetztenbeurteilungen erfasst. Auch die Selektionsrate bestimmte den praktischen Wert von Auswahlmethoden. Erst die Verwendung metaanalytischer Techniken brachte Klarheit in die Diskussion um die prädiktive Validität der verschiedenen Auswahl- und Assessmentmethoden. Vorherige Untersuchungen berichteten deutlich verschiedene prädiktive Validitäten, die jedoch meist auf Artefakte wie z.B. Stichprobeneffekte zurückzuführen waren. Tabelle 2-12 gibt einen Auszug der Ergebnisse der Metaanalyse von Schmidt und Hunter (1998) zur prädiktiven Validität verschiedener Auswahlmethoden sowie der inkrementellen

Validität der Verfahren über allgemeine kognitive Fähigkeiten hinaus wider. Als Kriterium diene allgemeine berufliche Leistung.

Tabelle 2-12: Prädiktive und inkrementelle Validität verschiedener Auswahlmethoden bzw. -variablen zur Vorhersage von beruflicher Leistung (nach Schmidt & Hunter, 1998)

Assessmentmethode	Prädiktive Validität	Inkrementelle Validität	Validitätszuwachs bzw. Nützlichkeitszuwachs
Allgemeine kognitive Fähigkeiten	.51		
Arbeitsprobe	.54	.63	24 %
Integritätstests	.41	.65	27 %
Gewissenhaftigkeitsmaß	.31	.60	18 %
Strukturierte Interviews	.51	.63	24 %
Unstrukturierte Interviews	.38	.55	8 %
Berufsbezogene Wissenstest	.48	.58	14 %
Überprüfung der Referenzen	.26	.57	12 %
Arbeitserfahrung (in Jahren)	.18	.54	6 %
Biographische Daten	.35	.52	2 %
Assessment Center	.37	.53	4 %
Interessentests	.10	.52	2 %
Graphologie	.02	.51	0 %
Alter	-.01	.51	0 %

Die Vorhersagekraft von Intelligenz bzw. allgemeinen kognitiven Fähigkeiten ist den Ergebnissen zufolge neben Arbeitsproben und strukturierten Interviews am höchsten. Dies ist v.a. darauf zurückzuführen, dass es zu allgemeinen kognitiven Fähigkeiten auch aufgrund der langen theoretischen Forschungstradition die meisten veröffentlichten Studien gibt. Zu Interviewverfahren gibt es im Gegensatz dazu vergleichsweise deutlich weniger Forschungsarbeiten (Schmidt & Hunter, 1998). Auch sind diese Verfahren vergleichsweise effizient und ökonomisch einsetzbar, weshalb sie viel in der Praxis der Personalauswahl eingesetzt werden. Graphologische Gutachten und Interessentests zeigen im Gegensatz dazu die geringste Vorhersagekraft. Diese Verfahren zeigen auch keinen bzw. kaum Zuwachs der Validität, wenn diese Verfahren zusätzlich zu einem Intelligenzmaß verwendet werden. Schmidt und Hunter (1998) empfehlen für die Praxis der Personalauswahl, grundsätzlich allgemeine kognitive Fähigkeiten zu erfassen und diese Tests mit einer Arbeitsprobe, einem Integritätstest oder einem strukturierten Interview zu kombinieren. Als einzige Persönlichkeitsvariable wurde auch die Validität von Gewissenhaftigkeit in der Metaanalyse mit erfasst. Die Vorhersagekraft von Gewissenhaftigkeit liegt bei $r = .31$ und die inkrementelle Validität über allgemeine Intelligenz hinaus beträgt .60. Somit ist Gewissenhaftigkeit auch eine sinnvolle Ergänzung der Methoden zur Leistungsvorhersage von Mitarbeitern.

Der Zusammenhang von Arbeitsleistung und Persönlichkeitseigenschaften war, wie bereits in Abschnitt 2.3 in Bezug auf die Vorhersage von Persönlichkeitseigenschaften beschrieben, im vergangenen Jahrhundert ebenfalls ein häufig untersuchtes Thema in der AO-Psychologie. Die Metaanalyse zweiter Ordnung von Barrick et al. (2001) fasste die Ergebnisse von 15 anderen Metaanalysen zusammen, die den Zusammenhang des Fünf-Faktoren-Modells mit Arbeitsleistung untersucht haben. Neben Neurotizismus korrelierte vor allem Gewissenhaftigkeit mit Arbeitsleistung ($r = .20$ bzw. $.23$). In Bezug auf spezifischere Leistungskriterien zeigte Gewissenhaftigkeit signifikante Zusammenhänge mit Vorgesetztenbeurteilungen ($r = .26$), objektiven Leistungskriterien ($r = .19$), Fortbildungsleistungen ($r = .23$) und Teamwork ($r = .23$). Gewissenhaftigkeit ist somit ein valider Prädiktor beruflicher Leistung über verschiedene Kriteriumsarten und Berufsgruppen hinweg. Der starke Zusammenhang von Gewissenhaftigkeit und beruflicher Leistung lässt sich inhaltlich so erklären, dass Mitarbeiter mit hohen Ausprägungen in Gewissenhaftigkeit eher dazu tendieren, zielorientiertes, moti-

viertes und hingebungsvolles Verhalten zu zeigen, was Vorgesetzte wiederum positiv bewerten (Sutherland et al., 2007).

Da das vorrangige Ziel dieser Arbeit allerdings die Vorhersage von Ausbildungsleistung, nicht die Vorhersage klassischer Leistung im Beruf per se. Die Vorhersagekraft von Persönlichkeitseigenschaften bzw. Gewissenhaftigkeit und deren Facetten wurde in den vorherigen Kapiteln für den beruflichen Kontext bereits ausführlich dargestellt. Aus ökonomischen Gründen wird daher an dieser Stelle die Vorhersage von Berufserfolg und -leistung nicht weiter vertieft, sondern in den folgenden beiden Abschnitten vor allem auf die Vorhersage von Schul- (2.7.2) und Berufsschulleistung (2.7.3) eingegangen.

2.7.2 Schulleistung

Eines der Hauptanliegen von Psychologen – neben der Vorhersage von Berufsleistung – ist auch die Vorhersage von akademischer Leistung. Sowohl historisch als auch international war und ist Schulleistung eines der wichtigsten Erfolgskriterien in der Gesellschaft, das seit vielen Jahren beforscht wird. Das Wissen um Faktoren, die schulischen Erfolg bedingen bzw. beeinflussen, hat wichtige Implikationen auf das Lernen und die Lehre. Lehrkräfte, Ausbilder und Mentoren interessieren sich dafür, wer in der Schule bzw. Ausbildung erfolgreich sein wird und wer nicht, da schulischer Erfolg maßgeblich auch beruflichen Erfolg bedingt (O'Connor & Paunonen, 2007; Poropat, 2011).

Einige Autoren haben sogar postuliert, dass Arbeits- und Schulleistung im Großen und Ganzen dasselbe seien (z.B. Lounsbury, Gibson, Sundstrom, Wilburn & Loveland, 2004). Diese Meinung ist teilweise darauf zurückzuführen, dass Schule Schüler auf die Arbeitswelt vorbereitet, indem sie „individuelle soziale Interaktionen strukturiert und Belohnungen bereithält, um so die Arbeitswelt zu replizieren“ (Bowles & Gintis, 1999, S. 3). Dennoch gibt es allerdings nur wenige Studien, die erfassen, welche Inhalte der formalen Schulbildung später tatsächlich in das Arbeitsleben eingebracht werden. Und auch wenn die tertiäre Bildung mit Arbeitsleistung korreliert ($r = .35$), ist dies dennoch kein Indikator dafür, dass diese beiden Leistungsindikatoren gleich sind (Poropat, 2009).

Schulleistung ist konzeptualisiert als (aktive) Teilnahme am Schulleben sowie als Durchschnitts- bzw. Abschlussnote. Dies geht einher mit der Annahme, dass sowohl kognitive als auch non-kognitive Komponenten eine Schlüsselrolle in der Schule, bei der Arbeit und in Bezug auf allgemeinen Lebenserfolg spielen (MacCann et al., 2009). Erfolg in der Schule beinhaltet das Lesen, Verstehen, Behalten, Reproduzieren und Interpretieren von Lernstoff und ist somit komplexer als „reine“ Intelligenz es erfordert (Furnham et al., 2003). Kriterien für schulischen Erfolg sind bspw. Schulnoten, Lehrerbewertungen, kognitive Tests, Anwesenheitszeiten, Disziplinarmaßnahmen, Teilnahme an besonderen Aktivitäten oder die Übernahme eines Ehrenamtes in der Schule (z.B. Schulsprecher).

Die am meisten verwendeten Kriteriumsmaße schulischer Leistung sind Schulnoten, insbesondere Abschlussnoten und Durchschnittsnoten – im englischsprachigen Raum als *grade point average* (GPA) operationalisiert (Poropat, 2009). Trotz ihrer vielfältigen Verwendung wurde die Reliabilität und Validität dieser Variablen jedoch auch häufig in Frage gestellt. Hierfür gibt es verschiedene Kritikpunkte, wie beispielsweise die Tendenz hin zu einer „Noteninflation“ in dem Sinne, dass für die gleiche Leistung auf verschiedenen Lernniveaus oder zu verschiedenen Zeitpunkten die gleiche Note vergeben wird. So ist eine sehr gute Abschlussnote auf einer Hauptschule anders zu bewerten als ein sehr gutes Abitur. Auch werden häufig „Deckeneffekte“ bemängelt. Dieser Effekt beschreibt ähnlich zur Noteninflation eine Tendenz dahingehend, dass eher gute Noten vergeben werden. Dies könnte wiederum zu verzerrten Normalverteilungskurven, Range Restriction oder Unterschieden in der Rangfolge und somit letztlich zu geringeren Korrelationen mit anderen relevanten Variablen führen (Poropat,

2009). Auch die PISA-Studie hat gezeigt, dass für gleiche Leistungen in unterschiedlichen Schulen unterschiedliche Noten vergeben werden (OECD PISA Deutschland, 2003, zitiert nach Görlich & Schuler, 2007). Daher sollten die Schulnoten für die Auswahl von Auszubildenden nicht die einzige Entscheidungsquelle sein.

Die Notenvergabe in Schulen bezieht sich generell auf einen Referenzrahmen, der sich in erster Linie an dem spezifischen Leistungsniveau der einzelnen Schule bzw. Lerngruppe und nicht an schulübergreifenden Kriterien orientiert. Für gleiche Leistungen werden demnach in unterschiedlichen Schulen unterschiedliche Noten vergeben. Das heißt, es gibt erhebliche Leistungsunterschiede auch zwischen Schulen derselben Schulform (Görlich & Schuler, 2007). Dies sollte man bei der Interpretation von Zusammenhängen mit Schulleistungskriterien bedenken. Es ist beispielsweise möglich, dass geringe Korrelationen nicht unbedingt auf einen mangelhaften Prädiktor, sondern auf Schwächen des Kriteriums (Schulnoten) zurückzuführen sind.

Auch wenn es die o.g. Probleme mit Noten und GPA als Leistungsindikatoren gibt, sind diese dennoch nützliche Maße zur Erfassung schulischer Leistung. So ist die interne Konsistenz bzw. Reliabilität von Schulnoten recht gut. Poropat (2009) berichtet von einer Stabilität von .94 nach vier Jahren, was im Vergleich mit Arbeitsleistung (.75) vergleichsweise hoch ist. Auch beeinflusst GPA nachhaltig die Zusammenhänge mit anderen Variablen. So berichtet der Autor in seiner Metaanalyse konsistente Zusammenhänge von GPA mit beispielsweise Intelligenz ($r = .56$), Arbeitsleistung ($r = .35$) oder beruflichem Status und Ansehen ($r = .37$). Das Wissen um Faktoren, die Leistung vorhersagen, kann somit bei der Trennung von Schülern helfen, die ein Lernprogramm erfolgreich absolvieren werden und denjenigen, die keinen Erfolg haben werden. Das Wissen um die zugrundeliegenden Prozesse und Erfolgsfaktoren von Schulleistung kann demnach dabei helfen, Schüler in die Richtung zu lenken, in der sie erfolgreich sein werden. Darüber hinaus können – basierend auf den Stärken und Schwächen der Schüler – Lehrpläne erstellt werden, um eventuelle Defizite aufzufangen und gezielt an diesen arbeiten zu können (O'Connor & Paunonen, 2007).

Lange Zeit war die Vorhersage von Schulleistung nur mit dem Intelligenzkonzept verbunden, da dieses explizit zur Vorhersage von Schulleistung bzw. potenziellen Lernschwierigkeiten entwickelt wurde. Metaanalysen (z.B. Kuncel, Hezlet & Ones, 2001) belegen entsprechend, dass allgemeine kognitive Fähigkeiten Schulleistung vorhersagen ($r = .21$ bis $.58$). Auch spezifische kognitive Fähigkeiten (z.B. verbale Fähigkeiten) sind valide Prädiktoren von Schulleistung ($r = .50$, Dollinger, Matyja & Huber, 2008). Allerdings erklären Intelligenztests kaum mehr als 50 % der Varianz bei der Vorhersage von Schulnoten, was darauf hinweist, dass auch andere Faktoren zur Schulleistung beitragen. Entsprechend belegen neuere Studien, dass beispielsweise Persönlichkeitseigenschaften prädiktive Validität in Bezug auf Schulleistung besitzen und kognitive Fähigkeiten allein keine hinreichenden Faktoren sind, um individuelle Leistungsunterschiede hinreichend erklären zu können (O'Connor & Paunonen, 2007). Vor allem bei der Vorhersage von akademischer Leistung, insbesondere für höhere Bildungsabschlüsse, scheinen Persönlichkeitsfragebögen ähnlich effektiv zu sein wie Intelligenztests.

Zu dieser Fragestellung führten Chamorro-Premuzic und Furnham (2008) eine Studie mit 158 College-Studenten durch, die bei Studienbeginn den NEO-PI-R ausfüllten, einen Lerntest sowie zwei Intelligenztests bearbeiteten. Ein Jahr später wurden die Kursnoten der Studenten als Kriterium zur Leistungsvorhersage verwendet. Zusammen klärten die Variablen ca. 40 % der Vorhersagevarianz von akademischer Leistung auf. Die Kursnoten korrelierten u.a. signifikant mit Gewissenhaftigkeit ($r = .37$, $p < .01$) und allgemeiner Intelligenz ($r = .24$, $p < .05$). Auch Gewissenhaftigkeit und allgemeine Intelligenz korrelierten signifikant miteinander ($r = -.19$, $p < .05$). Insgesamt klärten die kognitiven Fähigkeiten aber nur 6 % der Vorhersagevarianz auf, während Gewissenhaftigkeit demgegenüber zusätzliche 27 % der Varianz aufklärte. Offenheit für Erfahrungen besaß ebenfalls inkrementelle Validität über die allgemeinen

kognitiven Fähigkeiten hinaus und klärte zusätzliche 4 % Varianz auf. Lernverhalten erklärte zusätzliche 3 % der Varianz. Die Effekte von Intelligenz bei der Vorhersage wurden teilweise durch die Persönlichkeitseigenschaften und das Lernverhalten der Studenten mediert, was die Autoren mittels Pfadanalysen ermittelten. Die beiden Persönlichkeitsdimensionen Gewissenhaftigkeit und Offenheit sagten den Ergebnissen zufolge somit Schulleistung signifikant vorher und besaßen auch über die allgemeinen kognitiven Fähigkeiten hinaus inkrementelle Validität. Vor allem Gewissenhaftigkeit klärte einen großen Anteil an zusätzlicher Varianz in Bezug auf das Kriterium auf, auch wenn eine Kombination aus Persönlichkeitseigenschaften, Lernverhalten und allgemeiner Intelligenz insgesamt die beste Vorhersagekraft besaß.

Untersuchungen, die Persönlichkeitseigenschaften und schulische Leistung miteinander in Verbindung bringen, liegen bereits länger vor. Schon frühe Studien von Gough und Kollegen (1964) zeigten, dass die gewissenhaftkeitsbezogenen Skalen des California Psychological Inventory höhere Leistungswerte sowohl in der Highschool als auch im College vorhersagten (Gough, 1964; Gough & Hall, 1964) und Schüler, die für besonders gute Leistungen ausgezeichnet wurden, deutlich höhere Werte auf der Gewissenhaftigkeitsskala des Inventars aufwiesen als durchschnittliche Schüler.

Da Persönlichkeitstests jedoch ursprünglich nicht mit dem Ziel der Vorhersage von Schulerfolg entwickelt wurden, sollte überlegt werden, aus welchen Gründen Persönlichkeitsvariablen und Schulleistung zusammenhängen. Wenn nämlich das Fünf-Faktoren-Modell und dessen Annahmen auch in diesem Kontext zutreffen, sollten sich die fünf Dimensionen im Verhalten von Schülern und anderen schulbezogenen Leistungsvariablen widerspiegeln.

Eine konkrete, komplexe Theorie zur Vorhersage von Schulleistung hat sich bisher allerdings nicht durchgesetzt. Hierzu bestehen lediglich einige Ansätze. Beispielsweise die Idee, dass nicht nur Intelligenz, sondern auch sozioökonomischer Status und Persönlichkeit die Schulleistung beeinflussen. Dieser Ansatz ist konsistent zu den bereits in Abschnitt 2.7 erläuterten Annahmen von Campbell et al. (1993), dass Leistung durch drei Faktoren bedingt wird: Der *Fähigkeit zu leisten*, der *Möglichkeit zu leisten* und dem *Willen zu leisten*. Angewandt auf die Schule beinhaltet die *Fähigkeit zu leisten* allgemeines Wissen, Fähigkeiten und Intelligenz. Die *Möglichkeit zu leisten* wird von Umgebungsfaktoren und Ressourcen sowohl innerhalb als auch außerhalb der Schule bedingt. Der *Wille zu leisten* reflektiert die individuelle Motivation eines Schülers, allgemein geltende kulturelle Normen und individuelle Persönlichkeitseigenschaften (Poropat, 2009).

Verhaltenstendenzen, die sich in Persönlichkeitseigenschaften widerspiegeln, beeinflussen bestimmte Gewohnheiten, die wiederum schulischen Erfolg beeinflussen (können). Kognitive Fähigkeiten sagen etwas über die Fähigkeit zu leisten aus, während Persönlichkeitseigenschaften etwas darüber aussagen, was eine Person leisten will. In der Schule ist der Wille zu leisten beispielsweise assoziiert mit Anwesenheitszeiten, Initiative, Engagement in außer-curricularen Aktivitäten und Einstellungen zu lernen. Diese Faktoren haben nachweislich einen inkrementellen Effekt auf den Lernerfolg bzw. die Schulleistung einer Person über allgemeine Intelligenz hinaus und sind vor allem der Gewissenhaftigkeitsdimension des Fünf-Faktoren-Modells zuzuordnen (O'Connor & Paunonen, 2007). Gewissenhaftigkeit ist assoziiert mit ausdauerndem Einsatz und dem Setzen höherer Ziele (Barrick, Mount & Strauss, 1993). Auch andere Faktoren wie z.B. das Erledigen von Hausaufgaben sowie die Konzentration bei den Hausaufgaben (Trautwein, Lüdtke, Schnyder & Niggli, 2006) oder lernbezogenes Zeitmanagement hängen mit der Persönlichkeitsdimension zusammen (Bidjerano & Dai, 2007). Zudem ist es auch möglich, dass gewissenhaftere Schüler bessere Strategien und Handlungsmuster parat haben, um komplexe Aufgaben zu lösen.

Forscher haben auch Zusammenhänge zwischen Fremdratings der Gewissenhaftigkeit durch Lehrer und Schulnoten gefunden (Digman, 1989; Graziano & Ward, 1992). Allerdings könn-

ten diese Ergebnisse auch stark durch die subjektive Wahrnehmung der beurteilenden Lehrer konfundiert sein. Die Zusammenhänge könnten aufgrund der Bewertungstendenzen der Lehrer, vermeintlich gute Schüler konsistent zu ihrer fachlichen Einschätzung auch eine höhere Gewissenhaftigkeit zuzuschreiben, entstanden sein.

Chamorro-Premuzic und Furnham (2003a) untersuchten den Zusammenhang zwischen Persönlichkeitseigenschaften und akademischer Leistung in zwei studentischen Stichproben in Großbritannien. Die Ergebnisse zeigten, dass die Persönlichkeitsausprägungen, die in der ersten Uniwoche erhoben wurden, hoch prädiktiv für die Studienabschlüsse der untersuchten Studenten drei Jahre später waren. V.a. Gewissenhaftigkeit war ein signifikanter Prädiktor der universitären Leistung. In einer ähnlichen Studie untersuchten Chamorro-Premuzic und Furnham (2003b) auch die prädiktive Validität der Facetten der Big 5. Wiederum klärten die Big 5-Dimensionen 15 % der Varianz der Abschlussnoten auf. V.a. Gewissenhaftigkeit korrelierte positiv mit der akademischen Leistung. Auf der Ebene der Gewissenhaftigkeitsfacetten korrelierten Dutifulness und Achievement Striving signifikant positiv mit akademischer Leistung. Weitere Subfacetten der Big 5-Dimensionen korrelierten ebenfalls mit Leistung. Insgesamt konnten die Subfacetten gemeinsam 30 % der Varianz der Abschlussnoten aufklären.

Furnham et al. (2003) untersuchten zudem den Effekt, dass Schüler mit höheren Gewissenhaftigkeitsausprägungen oftmals geringere Ausprägungen in der Intelligenz aufweisen. Die Autoren nahmen an, dass es sich hierbei um eine kompensatorische Fähigkeit handle. Demzufolge würden Schüler in hoch kompetitiven Settings gewissenhafter werden bzw. eine höhere Gewissenhaftigkeit entwickeln, um in diesen Settings konkurrenzfähig zu bleiben. Eine geringere Fähigkeit (Intelligenz) würde den Autoren zufolge durch eine hohe Gewissenhaftigkeit ausgeglichen werden, während intelligendere Schüler diesen Annahmen zufolge keine hohe Gewissenhaftigkeit entwickeln müssten. Bereits Snow (1992, 1995) hatte entsprechend postuliert, dass Fähigkeiten, Einstellungen, Persönlichkeitseigenschaften und vorheriges Wissen interagieren und gemeinsam Lernprozesse und akademische Leistung beeinflussen. Furnham et al. (2003) untersuchten in einem längsschnittlichen Design 93 Studenten, die zum Zeitpunkt ihrer Universitätsbewerbung neben dem NEO-PI-R auch eine Einschätzung ihrer eigenen Intelligenz bzw. der Veränderbarkeit von Intelligenz abgeben mussten. Jeweils nach einem, zwei und drei Jahren wurde die Schulleistung der Probanden erhoben. Diese Einschätzung korrelierte signifikant mit der Gewissenhaftigkeit der Probanden ($r = .29$, $p < .05$). Weder die Einschätzung der eigenen Intelligenz, noch Intelligenz selbst waren signifikante Prädiktoren der Schulleistung. Geschlecht hingegen korrelierte signifikant mit Schulleistung ($r = .26$, $p < .05$), in dem Sinne, dass weibliche Studentinnen bessere Noten erhielten als männliche Studenten. Erst nachdem Geschlecht und die Annahmen über die Veränderbarkeit der Intelligenz auspartialisiert wurden, korrelierten Gewissenhaftigkeit und Schulleistung signifikant miteinander ($r = .34$ bis $.44$). Das heißt, Personen, die glaubten, die Intelligenz sei unveränderbar und zusätzlicher Einsatz keinen Mehrwert bringe, wiesen auch eine geringere Gewissenhaftigkeit auf. Gewissenhaftere Studenten zeigten allgemein höhere Anwesenheitszeiten und bessere Noten. Diese Effekte bestätigen auch die Ergebnisse von Chamorro-Premuzic und Furnham (2008), die ebenfalls einen negativen Zusammenhang zwischen Intelligenz und Gewissenhaftigkeit ($r = -.19$, $p < .05$), aber eine positive Korrelation zwischen Schulleistung und der Persönlichkeitsdimension ($r = .37$, $p < .01$) berichten.

O'Connor und Paunonen (2007) sowie Poropat (2011) veröffentlichten Metaanalysen zum Zusammenhang der Big 5-Persönlichkeitsdimensionen und Schulleistung. Alle fünf Persönlichkeitsdimensionen korrelierten mit Schulleistung. Die Ergebnisse zeigen, dass Allgemeine Intelligenz im Vergleich mit den Big 5-Persönlichkeitsdimensionen der beste Prädiktor akademischer Leistung ist. Von den Persönlichkeitsdimensionen ist Gewissenhaftigkeit mit einer mittleren Effektstärke von $d = .46$ der valideste Prädiktor von Schulleistung. Die korrigierte

Korrelation von Gewissenhaftigkeit ($\rho = .22$) ist sogar ähnlich zu der von Intelligenz ($\rho = .25$). Dies spricht für die Wichtigkeit der Persönlichkeitsdimension Gewissenhaftigkeit in Bezug auf die Vorhersage von Schulleistung. Gewissenhaftigkeit und Intelligenz erwiesen sich dabei als unabhängige Konstrukte ($r = -.03$). Wenn Intelligenz kontrolliert wurde, konnten nur für die Gewissenhaftigkeitsdimension Effekte dahingehend gefunden werden, dass die Gewissenhaftigkeit mit zunehmender Schulbildung anstieg. Moderierende Effekte für Alter wurden nicht gefunden. Die Ergebnisse sind in Auszügen in Tabelle 2-13 dargestellt.

Tabelle 2-13: Zusammenhänge zwischen den Big 5 und Intelligenz mit GPA (nach Poropat, 2009)

	<i>k</i>	<i>N</i>	<i>r^a</i>	<i>ρ^a</i>	<i>d</i>
Emotionale Stabilität	114	59554	.01	.02	.03
Extraversion	113	59986	-.01	-.01	-.02
Offenheit	113	60442	.10	.12	.24
Verträglichkeit	109	58522	.07	.07	.14
Gewissenhaftigkeit	138	70926	.19	.22	.46
Allgemeine Intelligenz	47	31955	.23	.25	.52

Anmerkungen: *k*= Anzahl der Studien; *N*= aggregierte Stichprobengröße; *r*= an der Stichprobengröße gewichteter Korrelationskoeffizient; *ρ*= an der Stichprobengröße gewichteter und anhand der Skalenreliabilität korrigierter Korrelationskoeffizient; *d*= Effektstärke; ^a $p < .001$ für alle Koeffizienten.

Auch im schulischen Kontext erweist sich somit wieder die Gewissenhaftigkeitsdimension als der Persönlichkeitsfaktor, der am meisten mit Schulleistung zusammenhängt ($r = .24$; $k = 16$; O'Connor & Paunonen, 2007).

Neben früherer Schulleistung ist den Ergebnissen von Poropat (2009) zufolge allgemeine Intelligenz der Prädiktor, der am meisten zur Vorhersage von Schulleistung herangezogen werden sollte. Allerdings ist die Nützlichkeit dieses Faktors durch die hohe Interkorrelation mit vorheriger Schulleistung stark begrenzt.

Durch das dreigliedrige Schulsystem in Deutschland, das nach der Grundschule bzw. Orientierungsstufe Schüler bereits in drei allgemeine Leistungsgruppen unterteilt, entsteht nämlich eine starke Einschränkung der Varianz (Range Restriction), die zu geringeren Vorhersagewerten führt. In Auswahl-situationen treten diese Effekte aufgrund der Selbstselektion der Bewerber und der zusätzlichen Vorauswahl anhand der Bewerbungsunterlagen, v.a. den Abschlusszeugnissen, durch die Unternehmen noch deutlicher zutage. Daher ist das Ergebnis, dass Gewissenhaftigkeit über Intelligenz hinaus und unabhängig davon Schulleistung bzw. Berufserfolg signifikant vorhersagen kann, noch bedeutungsvoller. Sofern sich Gewissenhaftigkeit zuverlässig erfassen lässt, sollte diese Persönlichkeitsdimension somit immer mit zur Vorhersage von Schulerfolg herangezogen werden.

Gewissenhaftigkeit zeigt zudem signifikante Zusammenhänge mit z.B. Durchfallraten und ist nicht nur in Bezug auf die Vorhersage von Erfolg, sondern auch für die Vorhersage von Misserfolg interessant (Poropat, 2009). Denn auch das rechtzeitige Ergreifen bzw. Einführen von Gegenmaßnahmen zur Verhinderung von (schulischem) Misserfolg, wie z.B. das Bereitstellen von Lernhilfen, die Schaffung von Strukturen für Lernen oder Mentorenprogrammen ist wichtig.

O'Connor und Paunonen (2007) berichten, dass insbesondere langfristiger Erfolg zuverlässiger durch Persönlichkeitseigenschaften vorhergesagt wird, da kognitive Leistungsmaße ihre Vorhersagekraft auf höheren Bildungsebenen zunehmend verlieren. Zudem sind ihren Beobachtungen zufolge die Zusammenhänge zwischen kognitiven Leistungsmaßen und Schulleistung in Universitätsstichproben in der Regel deutlich kleiner im Vergleich zu z.B. Stichproben von Grundschulern. Eine mögliche Erklärung für diesen Verlust der Vorhersagekraft ist den Autoren zufolge nicht nur die Range Restriction der Intelligenzwerte innerhalb der stark vorselektierten Studentenstichproben und die mit dem Studium einhergehenden

ständigen Leistungsüberprüfungen und andauernde Selektion in diesen Personengruppen (s. hierzu auch Furnham et al., 2003), sondern auch die zunehmende Verschiebung von Leistungsfaktoren von eher allgemeinen Fähigkeitsüberprüfungen hin zu sehr konkreten Leistungsbeurteilungen im Laufe der Schulzeit, die auch persönliche Eigenschaften mit berücksichtigen (wie z.B. von „kritischem Denken“ hin zur Überprüfung von „Fachwissen“).

Ältere Forschungsarbeiten mit Instrumenten, die nicht die Big 5 erfassen, legen nahe, dass gewissenhaftigkeitsnahe Traits, wie z.B. *Achievement* und (arbeitsorientierte) *Resilienz* mit schulischer Leistung zusammenhängen (Tross, Harper, Osher & Kneidinger, 2000). Paunonen und Ashton (2001) fanden entsprechend, dass einige Gewissenhaftigkeitsfacetten stärker mit Schulnoten korrelierten als andere. Insbesondere die *Achievement*- und *Endurance-Facetten* der PRF (Jackson, 1984) waren moderate Prädiktoren der Noten, während die *Order*-Skala fast gar nicht mit den Noten korrelierte. Gray und Watson (2002) fanden heraus, dass GPA am höchsten mit den Facetten *Leistungsstreben* ($r = .39$) und *Selbstdisziplin* ($r = .36$) im NEO-PI-R korrelierte, aber nur gering mit der *Ordnungsfacette* ($r = .15$). Darüber hinaus haben Forscher Selbstkontrollmechanismen, einem Aspekt von Gewissenhaftigkeit, untersucht und Zusammenhänge mit Schulnoten gefunden (Tangney, Baumeister & Boone, 2004). Gray und Watson (2002) fanden unterschiedliche Korrelationsmuster der Gewissenhaftigkeitsfacetten mit Highschool-GPA und College-GPA. Bei gleichzeitiger Analyse aller Gewissenhaftigkeitsfacetten erschien Pflichtbewusstsein als bester Prädiktor für Highschoolnoten, während die *Achievement-Facette* die beste Vorhersagekraft bei Collegenoten hatte. Dies erscheint kongruent zu der o.g. Unterscheidung von Gough (1957) zwischen *Leistung via Konformität* und *Leistung via Unabhängigkeit*. Ähnlich könnten Highschool- und Collegenoten zusammenhängen, denn mit zunehmender Schulbildung steigen auch der Komplexitätsgrad der Aufgaben und die Anforderung an Schüler, eigenständig zu lernen bzw. sich Inhalte eigeninitiiert zu erarbeiten.

Neben der allgemeinen Gewissenhaftigkeitsdimension fassten O'Connor und Paunonen (2007) in ihrer Metaanalyse auch die prädiktive Validität der Gewissenhaftigkeitsfacetten aus verschiedenen Persönlichkeitsinventaren in Bezug auf Schulleistung zusammen. Jackson et al. (1996) berichteten in ihrer Studie bereits, dass einige Facetten von Gewissenhaftigkeit besser geeignet sind, Schulleistung (GPA) vorherzusagen als andere. So zeigte *Achievement* deutliche Zusammenhänge mit GPA ($r = .24$, $p < .02$), während *Methodicalness* dieses Muster nicht zeigte ($r = -.01$, $p > .05$). Positive Korrelationen zwischen den Gewissenhaftigkeitsfacetten haben auch weitere Forschungsarbeiten belegt (De Fruyt & Mervielde, 1996; Gray & Watson, 2002; MacCann et al., 2009). Die Stärke der Zusammenhänge variierte jedoch je nach untersuchter Facette stark, was andeutet, dass einige Facetten wichtiger für schulischen Erfolg sind als andere.

Im direkten Vergleich der prädiktiven Validität der Gewissenhaftigkeitsdimension und ihren Facetten konnten Rothstein und Kollegen (1994) zeigen, dass nur die PRF-Facette *Achievement* ($r = .21$) und eine Verträglichkeitsfacette desselben Inventars mit Schulleistung zusammenhängen. Chamorro-Premuzic und Furnham (2003a) nutzten einen ähnlichen Ansatz und verglichen die NEO-PI-R-Facetten mit der zugehörigen Gewissenhaftigkeitsdimension. Sie fanden, dass GPA signifikant mit der Gewissenhaftigkeitsdimension ($r = .36$) sowie den Facetten *Dutifulness* ($r = .38$), *Achievement-Striving* ($r = .35$) und *Self-discipline* ($r = .22$) korrelierte. Die Ergebnisse dieser Arbeiten demonstrieren, dass – zumindest in einigen Situationen – engere Persönlichkeitsfacetten Vorteile gegenüber den allgemeinen Dimensionen haben. Allerdings sind dies nur korrelative Zusammenhänge, die keine Aussage über Ursache und Wirkung der gefundenen Effekte erlauben.

Laut den Ergebnissen der Metaanalyse von O'Connor und Paunonen (2007) sind die NEO-PI-R-Facetten *Achievement-striving* (beinhaltet Ehrgeiz, Fleiß und Ausdauer) mit $r = .15$ bis $.39$ sowie *Self-discipline* (beinhaltet Motivation, Zielorientierung, Konzentration) mit $r = .18$ bis

.46 die stärksten und konsistentesten Prädiktoren schulischer Leistung. Aber auch *Dutifulness* (= moralische Verpflichtungen erfüllen) mit $r = .25$ bis $.38$ hängt mit Schulleistung zusammen. Alle weiteren Facetten des NEO-PI-R spielen den Ergebnissen zufolge eine eher untergeordnete Rolle. Von den Subskalen des PRF zeigte die Achievement-Facette signifikante Zusammenhänge mit College-GPA ($r = .21$ bis $.27$) und Kursnoten ($r = .21$ bis $.26$).

O'Connor und Paunonen (2007) bemängeln, dass die Vorhersagekraft und inkrementelle Validität der Gewissenhaftigkeitsfacetten über die allgemeine Dimension hinaus bisher allerdings nur unzureichend beforscht wurden. Paunonen (1998) untersuchte die inkrementelle Validität von Facetten über Dimensionen hinaus in Bezug auf die Vorhersage schulischer Leistung mit Hilfe der PRF. Den Ergebnissen zufolge erklärten die Dimension 6 % der Varianz bezüglich GPA auf, die PRF-Facette Achievement erklärte zusätzliche 7,2 % der Varianz. In einer zweiten Studie waren die Dimensionen nicht in der Lage, GPA vorherzusagen, aber die Facette Responsibility erklärte 5,8 % der Varianz. Diesen Befunden zufolge sind Facetten in der Lage, ca. 5 bis 7 % zusätzliche Varianz aufzuklären. Insgesamt sind engere Facetten bei der Vorhersage von Schulleistung somit geeignete Prädiktoren, die zur Varianzaufklärung beitragen.

Farsides und Woodfield (2003) zeigen, dass kognitive Fähigkeiten 4 % der Varianz bei der Vorhersage von GPA aufklärten. Die ebenfalls erhobenen Anwesenheitszeiten der untersuchten Schüler in der Klasse erklärten zusätzliche 7 % der Varianz. Als dritten Faktor führten die Forscher die fünf Persönlichkeitsfaktoren des FFM in die Regressionsgleichung ein, welche zusätzliche 5 % Varianz aufklärten. Lounsbury, Sundstrom, Loveland und Gibson (2003) untersuchten ebenfalls die inkrementelle Validität von Persönlichkeitsfaktoren über kognitive Eigenschaften hinaus. Die Forscher fanden, dass allgemeine Intelligenz 16 % der Varianz und die Big 5 weitere 7 % Varianz aufklären konnten. Furnham und Chamorro-Premuzic (2004) fanden, dass Persönlichkeitseigenschaften zusätzlich zu den 3 % Varianz, die durch kognitive Fähigkeiten aufgeklärt wurden, 12 % der Varianz aufklären konnten. O'Connor und Paunonen (2007) schlagen vor, dass darüber hinaus erforscht werden sollte, inwiefern Persönlichkeitseigenschaften Schulleistung anders vorhersagen als Intelligenz.

Wie bereits in Abschnitt 2.5 ausführlicher beschrieben fanden ebenfalls Lievens et al. (2008), dass Gewissenhaftigkeit ($r = .19$) sowie Fleiß ($r = .22$) und Ordnung ($r = .18$) Schulleistung signifikant vorhersagten. Die Kriteriumsvaliditäten waren höher, wenn die Persönlichkeitsitems kontextbezogen bzw. schulbezogen formuliert waren ($r = .34$ bis $.41$). Die gefundenen Zusammenhänge ließen sich allerdings nur für die Gewissenhaftigkeitsdimension finden, für die anderen Traits des FFM fanden die Autoren keine signifikanten Effekte.

Noftle und Robins (2007) korrelierten in ihrer metaanalytischen Studie ($N = 10.497$ College-Studenten) verschiedene Schulleistungskriterien mit verschiedenen Gewissenhaftigkeitsmaßen (u.a. BFI, NEO-FFI, NEO-PI-R). Die gefundenen Zusammenhänge zwischen Gewissenhaftigkeit und dem verbalen ($r = -.01$, $p < .01$) sowie mathematischen SAT-Score ($r = -.07$, $p < .01$) waren allerdings nur gering. Bis auf zwei Ausnahmen korrelierten auch die Gewissenhaftigkeitsfacetten nicht mit den SAT-Scores. In Bezug auf GPA zeigten sich demgegenüber differenzierte und signifikante Ergebnisse. So korrelierten Highschool-GPA und College-GPA über alle Studien hinweg signifikant miteinander ($r = .18$ bis $.26$). Auch hinsichtlich der differenzierten Vorhersagekraft von Gewissenhaftigkeit bzw. deren Facetten zeigten sich signifikante Ergebnisse in Bezug auf die zwei Kriterien College- und Highschool-GPA (s. Tabelle 2-14).

Die Gewissenhaftigkeitsfacetten des HEXACO und NEO-PI-R korrelierten mit Highschool- und College-GPA, was untermauert, dass diese Facetten von Gewissenhaftigkeit insbesondere für akademischen Erfolg wichtig sind – egal auf welchem Schulniveau. Im Gegensatz zu den anderen Gewissenhaftigkeitsfacetten waren *Organization* (HEXACO) und *Order* (NEO-PI-R)

unabhängig von College-GPA und korrelierten unter den Facetten am geringsten mit High-school-GPA.

Tabelle 2-14: Zusammenhänge zwischen Gewissenhaftigkeit bzw. deren Facetten und GPA (nach Nofle & Robins, 2007)

Faktor/ Facetten	<i>r</i> College GPA	<i>r</i> Highschool GPA
HEXACO:		
Gewissenhaftigkeit	.20*	.26*
Organization	.00	.11*
Diligence	.24*	.24*
Perfectionism	.19*	.20*
Prudence	.21*	.21*
NEO-PI-R:		
Gewissenhaftigkeit	.18*	.25*
Competence	.19*	.22*
Order	.08	.13*
Dutifulness	.13	.17*
Achievement-Striving	.21*	.22*
Self-Discipline	.15*	.18*
Deliberation	.07	.17*

Anmerkungen: *r* = Korrelationskoeffizient; * $p < .05$.

Diese Resultate passen zu den Ergebnissen von Roberts et al (2005). Die Forscher fanden, dass die Facetten *Organisation* und *Ordnung* bzw. allgemein eine Tendenz hin zu einer übermäßigen Ordentlichkeit einen effektiven Fortschritt oder die Effektivität in Bezug auf Ziele behindern können. In sehr hohen Ausprägungen können diese Eigenschaften auch in Rigidität und Penibilität übergehen, wie sie bei Zwangsstörungen aufzufinden sind. Dies könnte

ein zusätzlicher Hinweis darauf sein, warum die bisher berichteten Zusammenhänge zwischen Gewissenhaftigkeit und schulischen Leistungskriterien eher moderat waren.

Die Ergebnisse von Nofle und Robins (2007) zeigen, dass Gewissenhaftigkeit bzw. deren Facetten unabhängig und inkrementell prädiktive Validität bezüglich akademischer Leistungen besitzen, auch über die „traditionellen“ Prädiktoren in diesem Bereich hinaus. Teilweise wies Gewissenhaftigkeit sogar bessere prädiktive Validitäten als der SAT-Score auf, was weiter für die prädiktive Validität der Persönlichkeitsdimension im schulischen Kontext spricht.

Die Ergebnisse von Nofle und Robins (2007) sind konsistent zu den Befunden von Paunonen und Ashton (2001). Die Autoren fanden heraus, dass die Facetten des PRF die Abschlussnoten (GPA) von Studenten besser vorhersagten als allgemeine Gewissenhaftigkeit. Alle Facetten des PRF korrelierten mit GPA, aber in unterschiedlichem Ausmaß. Am höchsten korrelierte die Achievement-Facette, die die Motivation hohe Leistungen zu erbringen erfasst, mit der Schulleistung. Im Gegensatz dazu konnte die Ordnungs-Facette, die ordentliches und sauberes Verhalten vorhersagt, keine Schulleistung vorhersagen. Allgemein sind die Aspekte der aktiven Leistungsannäherung höher mit Leistung korreliert, während Vermeidungskomponenten niedrigere Leistungen vorhersagen. Gewissenhaftere Menschen leisten somit mehr, da sie Aufgaben mit Blick auf ihre Kompetenzen erfüllen wollen.

Zusammenfassend bewiesen O'Connor und Paunonen (2007), dass sowohl Gewissenhaftigkeit als auch dessen Facetten, vor allem die Facetten Achievement, Self-discipline und Diligence Schulleistung vorhersagen können. Darüber hinaus können die Facetten oftmals zusätzliche Varianz in Bezug auf das Kriterium aufklären. Dies kann daran liegen, dass die Big 5 zu allgemein sind, um präziser Verhalten vorhersagen zu können, v.a. in konkreten Situationen. Darüber hinaus besitzen die Facetten eine höhere Varianz als die Traits (bei diesen wurde die systematische Varianz entfernt), die u.U. in besonderen Situationen prädiktiv ist.

Nofle und Robins (2007) fanden zudem starke Zusammenhänge zwischen Schulleistung (GPA) mit akademischem Aufwand ($r = .25$ bzw. $.17$, $p < .05$). Dieses Konstrukt gehört zu den selbst eingeschätzten eigenen Fähigkeiten und kann inhaltlich ebenfalls dem *Willen zu*

leisten zugeordnet werden. Diese Ergebnisse deuten an, dass über Gewissenhaftigkeit hinaus weitere Konstrukte die schulische Leistung bzw. deren Vorhersage beeinflussen.

Auch Selbstwirksamkeit korreliert positiv mit Schulleistung (Robbins, Lauver, Le, Davis, Langley & Carlstrom, 2004). Corker et al. (2012) untersuchten $N=347$ Studenten über den Zeitraum von zwei Semestern hinweg und fanden heraus, dass Selbstwirksamkeit drei akademische Leistungskriterien vorhersagen konnte. Hierzu zählten die erreichte Punktzahl in einem Kurs ($r=.15$), Hausaufgabenenerledigung ($r=.14$) sowie die Leistung in der Abschlussprüfung ($r=.18$).

Zusammenfassend lässt sich sagen, dass Schulbildung als Grundlage für Erfolg nicht nur auf individueller Ebene, sondern auch gesellschaftlich-politisch einen sehr hohen Stellenwert hat. Die Mitglieder der „Organisation für Wirtschaftliche Zusammenarbeit und Entwicklung“ (OECD) geben jährlich etwa 6,2 % ihres Bruttoinlandproduktes für Bildung und Ausbildung aus. Die Heranwachsenden in diesen Ländern sind durchschnittlich bis zu ihrem 22. Lebensjahr in der Ausbildung (OECD, 2007 zitiert nach Poropat, 2009). Schulleistung ist nicht nur ein wichtiger Prädiktor für späteren Berufserfolg, sondern damit einhergehend auch für Zufriedenheit, sozioökonomischen Status etc. Die Vorhersage von Schulerfolg bzw. -misserfolg ist somit ein wichtiges Feld in der Psychologie. Die Forschung hierzu ist allerdings bisher vor allem auf die Analyse korrelativer Zusammenhänge zwischen GPA und verschiedenen Intelligenz- und Persönlichkeits- bzw. Gewissenhaftigkeitsmaßen beschränkt. Furnham et al. (2003) kritisieren diesbezüglich, dass die bisherigen Studien bislang versäumt haben, die Zusammenhänge in einen theoretischen Rahmen einzubetten und inhaltlich zu erklären. Zukünftige Forschung sollte dies nachholen und sich mit der spezifischeren Erforschung des Zusammenhangs zwischen den Gewissenhaftigkeitsfacetten und Schulleistung beschäftigen. Schulleistung ist ein komplexes Konstrukt. Dabei sollten auch weitere Aspekte der Kriteriumsvariablen, wie z.B. Anwesenheitszeiten und Prüfungsergebnisse, analysiert werden, um die Zusammenhänge besser verstehen zu können (O'Connor & Paunonen, 2007). Zudem sollten die Zusammenhänge zwischen den Prädiktoren und Kriterien mittels multivariater Techniken untersucht werden. Beispielsweise könnten multiple Regressionen Aufschluss über die Vorhersagekraft einzelner Persönlichkeitsvariablen bzw. Facetten im Vergleich mit anderen geben. Auch die prädiktive Validität im Vergleich mit anderen Maßen und Konstrukten könnte so ermittelt werden.

2.7.3 Berufsschulleistung

Im Gegensatz zu z.B. den USA und Großbritannien gibt es in Deutschland, Österreich und der Schweiz für Schulabsolventen die Möglichkeit, eine *Berufsausbildung im dualen System* bzw. eine *Duale Ausbildung* zu absolvieren, die durch eine parallele Ausbildung an den beiden Lernorten Berufsschule und Ausbildungsbetrieb gekennzeichnet ist. In der Berufsschule werden theoretische Kenntnisse zu dem angestrebten Beruf vermittelt, während der ausbildende Betrieb für die Vermittlung praktischer Fertigkeiten zuständig ist. Das heißt, die Duale Ausbildung beinhaltet einerseits eine breit angelegte berufliche Grundbildung und andererseits die Aneignung der zur qualifizierten Berufsausübung notwendigen Fertigkeiten und berufspraktischen Erfahrungen (Wirth, 2008).

Derzeit gibt es in Deutschland bundesweit etwa 380 anerkannte Ausbildungsberufe. Die örtlichen Industrie- und Handelskammern sowie Handwerkskammern betreuen den größten Teil dieser Berufe. Sie beraten beispielsweise ausbildende Unternehmen und deren Auszubildende, registrieren und verifizieren Ausbildungsverträge und führen die Zwischen- sowie Abschlussprüfungen für die Auszubildenden durch (IHK Ostwestfalen, 2012). Die Industrie- und Handels- sowie Handwerkskammern überprüfen ebenfalls die Ausbildungsinhalte.

Insgesamt gab es im Jahr 2010 in Deutschland 1.508.476 im Dualen System registrierte Auszubildende (Statistisches Bundesamt, 2011). Als *Auszubildende* werden laut dem Statistischen Bundesamt (2011) Personen bezeichnet, die „aufgrund eines Ausbildungsvertrages nach dem Berufsbildungsgesetz eine betriebliche Berufsausbildung in einem anerkannten Ausbildungsberuf bzw. in einer Regelung für Menschen mit Behinderungen durchlaufen“ (S. 130). Insgesamt gibt es in Deutschland derzeit ca. 3.000 Berufsschulen, die in verschiedenen Bereichen, z.B. Industrie- und Handel, Landwirtschaft und Handwerk ausbilden. Zu den am stärksten besetzten Ausbildungsberufen gehörten im Jahr 2009 für männliche Auszubildende vor allem technische Berufe wie Kraftfahrzeugmechaniker (6,8 %), Industriemechaniker (5,3 %) und Elektroniker (3,7 %). Weibliche Auszubildende ließen sich eher in kaufmännischen und gesundheitsbezogenen Berufen, z.B. als Einzelhandelskauffrau (6,8 %), Bürokauffrau (6,6 %) und Medizinische Fachangestellte (6,5 %) ausbilden. Insgesamt besaßen im Jahr 2010 50 % der deutschen Bevölkerung über 15 Jahren einen beruflichen Abschluss, den sie durch eine Lehre bzw. eine Berufsausbildung im dualen System erworben hatten (Statistisches Bundesamt, 2011). Gesamtwirtschaftlich gesehen ist das duale Ausbildungssystem in Deutschland somit sehr bedeutend.

Die Berufsausbildung folgt in der Regel im direkten Anschluss an die Schulausbildung bzw. den Erwerb eines Haupt-, Real-, Fachhochschul- oder Gymnasialabschlusses. Wirth (2008) zufolge setzt das Duale System eine „frühe Berufsfindungsphase voraus und steht für eine enge Vernetzung zwischen Ausbildung, Berufswahl und zukünftiger Laufbahn“ (S. 87). Mehr als 50 % der Auszubildenden verbleiben nach dem erfolgreichen Abschluss ihrer Berufsausbildung in dem von ihnen erlernten Beruf. Es kann von einem stabilisierten Übergang der Jugendlichen in das Erwachsenenalter gesprochen werden.

Die Dauer der Berufsausbildung beträgt je nach Ausbildungsberuf zwei bis dreieinhalb Jahre.

Die Voraussetzung für eine Zulassung an einer Berufsschule ist das Abschließen eines Ausbildungsvertrages mit einem von der IHK als Ausbildungsbetrieb anerkannten Unternehmen. Die zu besuchende Berufsschule wird dabei durch den Ort bestimmt, an dem der ausbildende Betrieb bzw. dessen Niederlassung ansässig ist.

Unter *Berufsschulleistung* kann man allgemein die Schulleistung einer Person in der Berufsschule bezeichnen. Laut Lievens et al. (2009) kann man Leistung in der Berufsschule allgemein als Aneignung und Anwendung von deklarativem und prozeduralem berufsbezogenem Wissen bezeichnen. In der schulischen Berufsausbildung werden berufsbezogene bzw. fachtheoretische Fächer von berufsübergreifenden bzw. allgemeinen Fächern unterschieden. Dieser berufsbezogene Lernbereich lässt sich wiederum grob in einen kaufmännischen und einen nicht-kaufmännischen Bereich differenzieren (Wirth, 2008). Zu den fachbezogenen Unterrichtsfächern gehören beispielsweise Materialkunde oder Rechnungswesen. Zu den allgemein unterrichteten, berufsübergreifenden Fächern gehören z.B. Deutsch bzw. Kommunikation, Politik und Gesellschaftslehre sowie Sport bzw. Gesundheitsförderung. Die schulische Leistung wird – genauso wie auch an allgemeinbildenden Schulen – mit Hilfe schriftlicher und mündlicher Leistungsüberprüfungen in Form von z.B. Klassenarbeiten überprüft. Jeder Berufsschüler erhält zum Halbjahr sowie Schuljahresende daher auch ein Berufsschulzeugnis als Indikator seiner (bisherigen) Leistung.

Nach etwa zwei Jahren werden die grundlegenden Kenntnisse, Fertigkeiten und Fähigkeiten im Rahmen einer Zwischenprüfung schriftlich geprüft. Teilweise fließen diese auch bereits als erster Teil in das Gesamtergebnis der Abschlussprüfung ein, die nach etwa zwei bis drei Jahren durch die jeweilige IHK erfolgt. Der zweite Prüfungsteil bzw. die Abschlussprüfung zum Ende der Ausbildung konzentriert sich dann v.a. auf die berufstypischen Handlungskompetenzen, deren erfolgreicher Erwerb letztlich das Ziel der Ausbildung ist (IHK Ostwestfalen, 2012).

Görlich und Schuler (2007) differenzieren Ausbildungserfolg in *fachliche* und *überfachliche Leistung* im Betrieb sowie *Berufsschulleistung*. Den Autoren zufolge bezieht sich die fachliche Leistung auf Leistungsergebnisse und sei das erste betriebliche Erfolgskriterium. Das zweite betriebliche Leistungskriterium der überfachlichen Leistung beziehe sich auf Leistung als Grad der Lernzielerreichung. Unter Berufsschulleistung verstehen die Autoren allgemeine Berufsschulleistungen sowie Noten im IHK-Zwischenzeugnis. Die berufliche bzw. betriebliche Leistung sei durch globale Leistungsbeurteilungen des betrieblichen Ausbilders messbar. Die Autoren berichten diesbezüglich auch Zusammenhänge verschiedener Arten von Leistungsbeurteilungen. Ihren Angaben zufolge korrelieren Selbstbeurteilungen von Auszubildenden bezüglich ihrer eigenen Leistung mit den tatsächlichen betrieblichen Leistungsbeurteilungen moderat ($r = .39$). Die Annahmen der Auszubildenden über die Leistungsbeurteilung des Ausbilders und der eigenen Leistungsbeurteilung ($r = .73$) sowie der tatsächlichen betrieblichen Leistungsbeurteilung ($r = .55$) stimmen ihren Daten zufolge demgegenüber stärker überein.

Unabhängig von der Differenzierung der Ausbildungs- und Berufsschulleistung ist es aus den o.g. Gründen bedeutungsvoll, gerade bei der Auswahl von Auszubildenden eine möglichst zuverlässige Leistungsvorhersage treffen zu können. Schulnoten sind klassische erste Selektionsfilter für die Auswahl von Auszubildenden. Gemeinsam mit anderen Kriterien, wie z.B. dem Aussehen der Bewerbungsunterlagen, entscheiden sie über die Zulassung des jeweiligen Kandidaten zu weiteren Methoden der Personalauswahl bzw. der Absage oder Vertragszusendung an einen Bewerber. Dies reduziert allerdings auch den Anteil der Bewerber, die psychologische Eignungstests durchlaufen, obwohl die prognostischen Mängel wie beispielsweise mangelnde Vergleichbarkeit von Schulabschlüssen oder die Wohlwollenspflicht bei der Zeugniserstellung bekannt sind (Spengler & Fintrup, 2011). Zudem ist es auch aufgrund der bereits in Abschnitt 2.7.2 geschilderten Probleme der Aussagekraft von Schulnoten wichtig, auch weitere Vorhersagemaße wie z.B. Eignungstests zur Leistungsvorhersage heranzuziehen (Görlich & Schuler, 2007).

Für die Auswahl von Mitarbeitern ohne Berufserfahrung wie z.B. Auszubildende empfehlen Schmidt und Hunter (1998) in ihrer Metaanalyse zur Vorhersage von Berufsleistung die Verwendung von Maßen zur Erfassung der allgemeinen mentalen Fähigkeiten bzw. Intelligenz einer Person. Aber auch weitere zusätzlich verwendete Maße zur Erfassung von beispielsweise Gewissenhaftigkeit oder strukturierte Interviews können zur Vorhersagequalität des Auswahlprozesses beitragen.

Viele der bereits in Abschnitt 2.7.1 aufgezählten Auswahl- und Assessmentmethoden sagen auch berufsbezogenes Lernen voraus (Schmidt & Hunter, 1998). Allerdings gibt es zu berufsbezogenem Lernen, insbesondere dem Dualen Ausbildungssystem, deutlich weniger veröffentlichte Studien als zu Berufs- und Schulleistung allgemein. Schmidt und Hunter (1998) fassten in ihrer Metaanalyse auch Forschungsarbeiten zu berufsbezogenem Lernen in vergleichbaren Stichproben zusammen. Die Ergebnisse sind in Tabelle 2-15 dargestellt.

Den Ergebnissen von Schmidt und Hunter (1998) zufolge erweisen sich Maße zur Erfassung allgemeiner kognitiver Fähigkeiten mit Abstand als beste Prädiktoren zur Vorhersage von Ausbildungserfolg. Dahinter folgen Integritätstests, Einstellungsinterviews und die Erhebung biografischer Daten. Auch Gewissenhaftigkeitsmaße weisen eine gute prädiktive Validität auf. Für die Kombination von Testverfahren erweist sich eine Hinzunahme von Integritätstests und Gewissenhaftigkeitsmaßen als gute Möglichkeit, die Vorhersagekraft zu verbessern. So sind Gewissenhaftigkeitsmaße mit 16 % Validitätszuwachs sogar bessere Ergänzungen im Auswahlprozess als das Durchführen von Einstellungsinterviews (5 % Validitätszuwachs) oder einer Überprüfung der Referenzen (9 % Validitätszuwachs).

Tabelle 2-15: Prädiktive und inkrementelle Validität verschiedener Auswahlmethoden bzw. -variablen zur Vorhersage von beruflichem Lernerfolg (nach Schmidt & Hunter, 1998)

Assessmentmethode	Prädiktive Validität	Inkrementelle Validität	Validitätszuwachs bzw. Nützlichkeitszuwachs
Allgemeine kognitive Fähigkeiten	.56		
Integritätstests	.38	.11	20 %
Gewissenhaftigkeitsmaß	.30	.09	16 %
Einstellungsinterviews	.35	.03	5 %
Überprüfung der Referenzen	.23	.05	9 %
Arbeitserfahrung (in Jahren)	.01	.00	0 %
Biographische Daten	.30	.00	0 %
Interessentests	.18	.03	5 %

Es ist allerdings beobachtbar, dass die Vorhersagewerte von Intelligenzmaßen bezüglich Ausbildungserfolges in Deutschland ($r = .47$) allgemein niedriger sind als beispielsweise in den USA ($r = .57$) oder im europäischen Durchschnitt ($r = .54$; Görlich & Schuler, 2007). Die Autoren erklären diese Effekte mit der starken schulischen Vorselektion der Studenten bzw. Schüler in Deutschland aufgrund des dreigliedrigen Schulsystems. Dies führe zu homogenen Stichproben und somit auch zu weniger Aussagekraft leistungsbezogener Vorhersagevariablen wie z.B. der Intelligenz als Auswahlkriterium. Auch aus diesem Grund sollte man vor allem auch in Deutschland weitere Maße zur Auswahl bei Auszubildenden einsetzen, wie beispielsweise Persönlichkeits- bzw. Gewissenhaftigkeitstests.

Zudem beruhen die Ergebnisse der Metaanalyse von Schmidt und Hunter (1998) lediglich auf der Analyse korrelativer Zusammenhänge aus verschiedenen Studien. Die Annahme hinter den Zusammenhängen ist, dass diese auf stabilen Zusammenhängen zwischen Prädiktoren und Kriterien basieren, die zeitlich überdauernd sein sollten. Das heißt, Leistungsunterschiede sollten langfristig auf hohe und niedrige Ausprägungen der Prädiktoren zurückzuführen sein. Diese Effekte können jedoch nicht durch einfache Korrelationsanalysen bestätigt werden, sondern bedürfen längsschnittlicher Untersuchungen, denn sämtliche (potenzielle) Änderungen eines Validitätskoeffizienten haben Auswirkungen auf die Nützlichkeit des Instruments in der Praxis und sollten daher unbedingt berücksichtigt werden. Auch ist es möglich, dass die Zusammenhänge darauf zurückzuführen sind, dass Schule bzw. Berufsschule mit zunehmendem Alter dem Arbeitsumfeld immer ähnlicher wird und es somit aufgrund der zunehmenden inhaltlichen Übereinstimmung zwischen Prädiktor und Kriterium zu höheren Zusammenhängen kommt. Längsschnittliche Untersuchungen mit mehreren Messzeitpunkten könnten hierüber Aufschluss geben.

Zudem kritisieren einige Forscher, dass in den meisten vorliegenden längsschnittlichen Studien zur prädiktiven Validität von kognitiven Fähigkeiten oder Persönlichkeitseigenschaften nur selten ein Zeitraum gewählt wurde, der länger als ein Jahr war (Lievens et al., 2009). Bereits Humphreys (1960) postulierte, dass man bei der Untersuchung der prädiktiven Validität eignungsdiagnostischer Instrumente nicht nach der ersten (erfolgreichen) Validierung – zeitlich gesehen – aufhören bzw. zufrieden sein sollte. Er stellte in seinen Untersuchungen fest, dass vergangenes Verhalten weniger gut mit zukünftigem Verhalten korrelierte. Je länger eine Testung zurücklag, desto geringer war der Zusammenhang des später gezeigten Verhaltens mit den Prädiktoren. Für die Abnahme der prädiktiven Validität eines Verfahrens gibt es einen logischen Erklärungsansatz dahingehend, dass sich einerseits eine Person über die Zeit hinweg ändert und andererseits sich auch die Anforderungen an bzw. Aufgaben von einer Person sich über die Zeit verändern. Beide Effekte können auch gemeinsam auftreten und sogar miteinander interagieren.

Judge et al. (1999) untersuchten den Einfluss von Big 5 Persönlichkeitseigenschaften und allgemeinen kognitiven Fähigkeiten auf späteren Berufserfolg in einer Längsschnittstudie, die

den Zeitraum von der Kindheit bis ins Erwachsenenalter umfasste. Dabei erwies sich die im Kindesalter erfasste Gewissenhaftigkeit als valider, positiver Prädiktor für späteren intrinsischen ($r = .40$) und extrinsischen ($r = .41$) Berufserfolg. Allgemeine kognitive Fähigkeiten sagten lediglich extrinsischen Berufserfolg positiv vorher. Auch wenn die im Erwachsenenalter erfassten Persönlichkeitseigenschaften höhere Zusammenhänge mit den Erfolgsmaßen zeigten, deuteten die Ergebnisse an, dass sich die prädiktiven Effekte der Gewissenhaftkeitsdimension über die Zeit zu akkumulieren schienen und der Einfluss der Persönlichkeit auf beruflichen Erfolg langfristig betrachtet substantiell ist, auch wenn die Zusammenhänge zwischen Persönlichkeit und unmittelbarer beruflicher Leistung eher moderat sind.

Lievens et al. (2009) untersuchten den langfristigen Einfluss von Persönlichkeitseigenschaften auf akademische Leistung in einer US-amerikanischen College-Stichprobe, die vom Alter und Ausbildungsstand her vergleichbar zu den Auszubildenden in Deutschland ist. Sie führten zwei Gründe an, warum die Vorhersagekraft von Persönlichkeitsmerkmalen über die Zeit hinweg nicht ab- sondern zunehmen sollte. Zum einen überwiegen den Autoren zufolge während der Gewöhnungsphase an eine neue Tätigkeit die Effekte der Neuheit und Anforderungen der neuen Tätigkeit, die die natürliche Motivation und individuellen Persönlichkeitsausprägungen eines neu eingestellten Arbeitnehmers maskieren bzw. reduzieren. Dieser Effekt lasse jedoch nach der Eingewöhnungsphase nach und die natürliche Motivation und individuellen Persönlichkeitsausprägungen des Individuums kommen zum Tragen. Daher zeigen Persönlichkeitsmerkmale auf längere Sicht hin höhere prädiktive Validitäten. Zum anderen sollte aufgrund der sich zeitlich verändernden Arbeitsanforderungen (vom reinen Wissens- und Fähigkeitserwerb hin zu eigentlicher Leistung) spätere Leistung besser durch Persönlichkeitsvariablen vorhergesagt werden. Diese Fragestellung wurde von den Autoren an einer Kohorte von Medizinstudenten ($N = 785$) über einen Zeitraum von sieben Jahren untersucht. Die Studenten füllten wiederholt den NEO-PI-R (Costa & McCrae, 1992) aus. Als Kriterium diente der Notendurchschnitt (GPA) der Studenten. Abbildung 2-4 gibt einen Überblick über den Verlauf der operationalen Validitäten der Gewissenhaftkeitsdimension und ihrer sechs Facetten hinsichtlich des Kriteriums über den Untersuchungszeitraum hinweg.

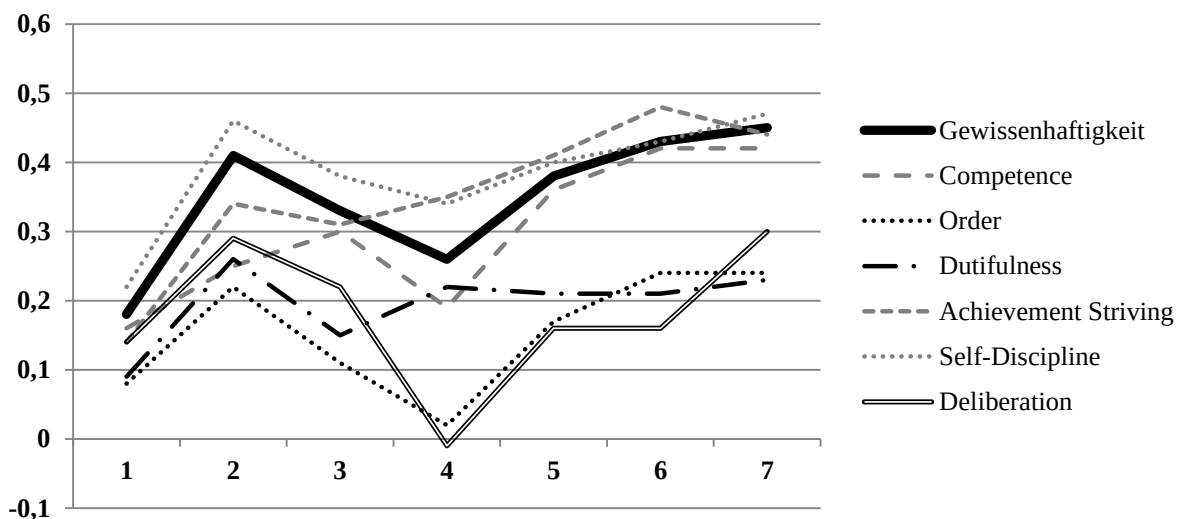


Abbildung 2-4: Verlauf der operationalen Validitäten der Gewissenhaftigkeitsskala und -facetten zur Vorhersage von Studienerfolg über einen Zeitraum von sieben Jahren (nach Lievens et al., 2009)

Der Abbildung ist zu entnehmen, dass die prädiktiven Validitäten der Gewissenhaftkeitsdimension sowie deren Facetten jedes Jahr tendenziell etwas größer werden, d.h. sie sagen die Leistung immer besser vorher. Für die Dimension Gewissenhaftigkeit kann ein Validitätszuwachs von 27 % von $r_{T1} = .18$ auf $r_{T7} = .45$ beobachtet werden. Ein ähnlicher Effekt (durchschnittlich 21 % Zuwachsrates) zeigte sich auch für die Gewissenhaftkeitsfacetten. Gewis-

senhaftigkeit ist den Ergebnissen von Lievens et al. (2009) zufolge somit eher für die Vorhersage späterer Schulleistung ein guter Prädiktor.

Zudem schienen die eher proaktiven Gewissenhaftigkeitsfacetten (= Self-Discipline, Achievement Striving und Competence) Leistung besser vorherzusagen als die eher inhibitorischen und regulatorischen Facetten (= Order, Deliberation und Dutifulness). Das heißt, Studenten, die frühzeitig mit dem Lernen beginnen, sich hierfür einen Plan machen und diesen auch unter widrigen Umständen einhalten, tendieren dazu, bessere Noten in der Schule zu erreichen als Studenten, die sich in diesem Bereich eher schlecht einschätzen.

Spengler und Fintrup (2011) ermittelten in ihrer Praxisstudie anhand multipler hierarchischer Regressionsanalysen die inkrementelle Validität psychometrischer Verfahren gegenüber der Vorhersagekraft von vorherigen Noten. Insgesamt sagen die eignungsdiagnostischen Verfahren ihren Ergebnissen zufolge 27 % der Leistungsunterschiede der Kandidaten voraus. Ohne Einsatz der psychologischen Verfahren hätten 37 % aller Kandidaten aufgrund ihrer (schlechteren) Noten eine falsche Leistungsprognose für ihr Examen erhalten. Ob ein Proband das anspruchsvolle Examen zum Wirtschaftsprüfer bestand, ließ sich besser anhand von psychologischen Tests als anhand von Noten vorhersagen. Insgesamt lässt sich aus diesen Ergebnissen folgern, dass psychologische Verfahren eine hohe Vorhersagekraft besitzen und für Unternehmen ein hoher monetärer Nutzen durch die Anwendung eignungsdiagnostischer Auswahlmethoden angenommen werden kann. Die geringe Übereinstimmung der vorherigen Schul- bzw. Ausbildungsnoten mit dem Bestehen des Wirtschaftsprüferexamens zeigt auf, dass die Vorhersagekraft von Noten bisher v.a. der hohen Basisrate geschuldet war und ihre Vorhersagekraft daher eher gering ist.

Lord (2011) zufolge ist der Zusammenhang bzw. die prädiktive Validität von Persönlichkeitseigenschaften in Bezug auf Leistung im beruflichen Kontext auch davon abhängig, in welchem Ausmaß die Verhaltensweisen, die sich aus den Persönlichkeitseigenschaften ergeben, im Arbeitsumfeld wertgeschätzt werden. Beispielsweise kann der Autorin zufolge die Feststellung, dass Gewissenhaftigkeit ein universeller Prädiktor für Arbeitsleistung ist, durch die Tatsache erklärt werden, dass Leistungsstreben, Zielorientierung und Zuverlässigkeit bei den meisten Berufstätigkeiten geschätzt werden.

Es ist darüber hinaus denkbar, dass Gewissenhaftigkeit Leistung noch zuverlässiger vorhersagen könnte, wenn man auch die moderierenden Effekte weiterer Variablen mit berücksichtigen könnte. Darüber hinaus postulieren die Autoren, dass je mehr Wissen und Expertise man sich in einem Bereich aneignet, desto höher der Einfluss der relevanten Persönlichkeitseigenschaften werden kann. Wenn dies der Fall ist, bedeutet dies für bisherige und zukünftige Validierungsstudien, dass Validitätsberechnungen mit früh erhobenen Kriterien die wahre Validität des jeweils untersuchten Persönlichkeitsmerkmals wohl eher unterschätzen. So weisen ältere Untersuchungen mit Studienabbrechern im Bereich der Medizinwissenschaften darauf hin, dass Ausbildungs- bzw. Studienabbrecher signifikant niedrigere Ausprägungen in der Leistungsorientierung hatten und ebenfalls geringere Ausprägungen u.a. in der Gewissenhaftigkeitsfacette Verantwortungsbewusstsein besaßen (Gough & Hall, 1964).

Sutin et al. (2009) analysierten in ihrer Langzeitstudie zur Vorhersage von Berufserfolg mittels Gewissenhaftigkeit bzw. dessen Facetten auch die möglichen Einflüsse weiterer Variablen wie z.B. Geschlecht, Alter, Ausbildung und Ethnizität. Sie untersuchten über einen Zeitraum von zehn Jahren eine Absolventenstichprobe, die an zwei Untersuchungen T1 ($N_{T1}=731$) und T2 ($N_{T2}=302$) teilnahmen und den NEO-PI-R beantworteten sowie verschiedene Maße zur Erhebung von Berufserfolg ausfüllten. Auf der Facettenebene war die Kompetenz-Facette mit Tätigkeiten in prestigeträchtigen Funktionen assoziiert. Allgemein gaben gewissenhaftere Menschen an, höhere Einkommen zu beziehen und zufriedener mit ihrer Arbeit zu sein. Dieser Effekt blieb auch nach der Kontrolle von Geschlecht, Alter, Ausbildung und

Ethnizität bestehen. Auf spezifischerer Ebene korrelierten die drei Gewissenhaftigkeitsfacetten *Competence* ($r = .20, p < .01$), *Dutifulness* ($r = .15, p < .01$) und *Achievement striving* ($r = .14, p < .01$) mit Gehalt. Wenn die Kontrollvariablen konstant gehalten wurden, korrelierten alle Gewissenhaftigkeitsfacetten außer Order und Deliberation signifikant mit Arbeitszufriedenheit ($r = .11$ bis $.18, p < .01$).

Zusammenfassend lässt sich sagen, dass Persönlichkeitseigenschaften, vor allem Gewissenhaftigkeit bzw. deren Facetten, wichtige Prädiktoren für die Vorhersage von Berufsschulleistung sind. Aufgrund der Unterschiede zwischen den verschiedenen Schulsystemen in den unterschiedlichen Ländern, vor allem der Besonderheit des dreigliedrigen Schulsystems in Deutschland, besteht weiter ein Bedarf an der Erforschung der Vorhersagekraft von Persönlichkeitseigenschaften in diesem Bereich. Wie bereits in Abschnitt 2.7.2 zuvor angemerkt, fehlt zudem bislang ein Vorhersagemodell, das die Zusammenhänge in einen theoretischen Rahmen einbettet und auch mögliche weitere Einflussvariablen mit berücksichtigt. Diese Mängel werden in der vorliegenden Studie aufgegriffen. Im nachfolgenden Abschnitt 2.8 soll daher basierend auf den bisher geschilderten Erkenntnissen ein theoretisches Modell zur Vorhersage von Schul- bzw. Berufsschulerfolg mittels verschiedener Persönlichkeitsvariablen entwickelt werden.

2.8 Theoretische Vorhersagemodelle in der Eignungsdiagnostik

Wie bereits in den vorherigen Abschnitten angemerkt, gibt es bislang kein anerkanntes theoretisches Modell, das der Erklärung der Zusammenhänge zwischen bzw. Vorhersagekraft von Persönlichkeitseigenschaften in Bezug auf Ausbildungsleistung dient und über die einfache, lineare Untersuchung von Zusammenhängen zwischen verschiedenen Persönlichkeitsvariablen und Leistungsindikatoren hinausgeht. Der große Anteil unerklärter Varianz bei der Vorhersage von Leistung spricht dafür, dass diese Beziehungen deutlich komplexer sind, als in der Forschung bisher untersucht wurde. Denn *„the fact of the matter is that human behavior at work is complex, and understanding the relation of personality traits and job performance is difficult. Personality traits are enduring, distal forces that influence behavior, but there are both mediating and moderating variables that must be accounted for to adequately explain the effects of personality on human behavior”* (Barrick & Mount, 2004, S. 369).

Es gibt einige Ansätze, die Zusammenhänge zwischen Persönlichkeit und Leistung erklären, welche im folgenden Abschnitt 2.8.1 erläutert werden. Im Vordergrund dieses Abschnittes steht allerdings die Entwicklung eines eigenen theoretischen Vorhersagemodells, welches in Abschnitt 2.8.2 dargestellt und erläutert wird.

2.8.1 Existierende Vorhersagemodelle

Ein Hauptanliegen von Unternehmen bei der Auswahl von Mitarbeitern ist die Vorhersage von Arbeitsleistung. Hierzu wird angestrebt, eine möglichst hohe Übereinstimmung zwischen den Anforderungen einer Tätigkeit und den fachlichen Fähigkeiten sowie persönlichen Eigenschaften eines Bewerbers herzustellen (vgl. Abschnitt 2.1 ff). Doch nicht nur für Unternehmen ist es wichtig, Erfolg und Leistung von Mitarbeitern vorherzusagen, denn auf Seiten der Mitarbeiter ist Arbeit auch mehr als nur eine Einkommensquelle. Die eigene Arbeit ist in der Regel ein wichtiger Aspekt der eigenen Identität, der die Entwicklung neuer Fähigkeiten und Fertigkeiten sowie von sozialen Bindungen ermöglicht (Sutin et al., 2009). Daher ist es von Interesse zu erfassen, wie die dynamischen Beziehungen zwischen Persönlichkeit und Karriereerfolg gestaltet sind.

Es wurden viele Zusammenhänge zwischen Persönlichkeitseigenschaften und relevantem Arbeitsverhalten gefunden, wie z.B. Abwesenheiten, gesundheitsbewusstem Verhalten und

Zufriedenheit. Diese Zusammenhänge werden zwar nicht immer auch von Intelligenz bzw. allgemeinen kognitiven Fähigkeiten vorhergesagt, sind für den Erfolg eines Unternehmens jedoch wichtig (Barrick & Mount, 2004). Der Erklärung dieser Zusammenhänge sollte daher auch aus theoretischer Sicht mehr Aufmerksamkeit gegeben werden. In ihrer Metaanalyse skizzierten Barrick und Mount (2004) den Zusammenhang zwischen Persönlichkeit, Selbstwirksamkeit und Leistungsmotivation zur Vorhersage von Arbeitsleistung bzw. -erfolg. Auf Seite der Prädiktoren sollten den Autoren zufolge persönliche und motivationale Eigenschaften, wie z.B. Gewissenhaftigkeit stehen. Die Autoren sehen Persönlichkeitseigenschaften als wesentlichen Bestandteil der distalen Motivation an, die wiederum die proximale Motivation einer Person beeinflusst. Zudem sprechen sie sich dafür aus, die motivationalen Prozesse, die Verhalten verursachen und beeinflussen, näher zu untersuchen, um auch Mediatoren sowie Moderatoren des Zusammenhangs von Persönlichkeitseigenschaften und Berufserfolg bzw. -leistung zu identifizieren.

Es sind zudem einige integrierende Theorien entwickelt worden, um die Zusammenhänge zwischen Persönlichkeitsvariablen und Leistung erklären zu können. Ackerman (1996) versucht mit seiner *PPIK-Theorie* (= *Process, Personality, Interests and Knowledge*) u.a. zu erklären, wie Persönlichkeitstraites und Interessen mit kognitiven Fähigkeiten interagieren, um Wissen aufzubauen. Hierzu knüpft er an eine Vielzahl theoretischer Ansätze zur Erklärung der intellektuellen Entwicklung einer Person an. Der PPIK-Theorie zufolge hängt das individuelle Kompetenzprofil einer Person davon ab, wie ein Individuum seine Ressourcen über die verschiedenen Bereiche verteilt bzw. wie groß der Anteil verfügbarer Ressourcen ist, die in das jeweilige Gebiet investiert werden. Er nimmt an, dass die wichtigste Ressource beim Aufbau von Fachwissen die individuelle Informationsverarbeitungskapazität, d.h. die allgemeine kognitive Leistungsfähigkeit, einer Person ist (Wild & Möller, 2009).

Eine Problematik bei der Erstellung eines theoretischen Vorhersagemodells für Berufs- bzw. Schulleistung mittels Persönlichkeitseigenschaften sind die verschiedenen theoretischen Annahmen, die den verschiedenen Persönlichkeitsmodellen zugrunde liegen. So bildet der genetische Traitansatz (vgl. Abschnitt 2.3) die Grundlage für die Big 5-Persönlichkeitsdimensionen, denen auch die Gewissenhaftigkeitsdimension angehört. Die Annahme des Traitansatzes ist, dass Persönlichkeit genetisch verankert und über das Leben hinweg stabil ist. Allerdings können diese Modelle die interpersonellen Unterschiede und Dynamiken im Verhalten zwischen Personen nicht hinreichend erklären (De Fruyt & Salgado, 2003).

Das Selbstwirksamkeitskonstrukt hingegen stammt aus der sozial-kognitiven Lerntheorie (vgl. Abschnitt 2.3.2), die annimmt, dass Selbstwirksamkeit sich durch Lernerfahrungen ausbildet und sich entsprechend der Erfahrungen, die ein Individuum macht, verändert. Die sozial-kognitive Theorie und ihre zentrale Variable Selbstwirksamkeit waren im Fokus vieler Untersuchungen. Die Anwendbarkeit der Theorie ist in vielen Kontexten gegeben. V.a. im arbeitspsychologischen Bereich wurde die Theorie viel angewendet bzw. überprüft. Es scheint, dass diese eine dominante Rolle in der Arbeitsmotivation spielt (Judge et al., 2007). Lent und Kollegen (1994) schnitten die Theorie explizit für den Berufskontext zu, um beispielsweise die Entwicklung beruflicher Interessen oder Karriereentscheidungen erklären zu können. Sie nahmen zudem das Konstrukt der Persönlichkeit als exogene, bestehende Variable an, die der Entwicklung von Selbstwirksamkeit vorausgeht.

Heutzutage rücken die Forscher von diesen entgegengesetzten extremen Ansichten jedoch etwas ab und integrieren die unterschiedlichen Ansichten in einem interaktionistischen Ansatz, der ein Zusammenspiel persönlicher Eigenschaften (Traits) mit verschiedenen Lernerfahrungen annimmt (Funder, 2006). Mittels des sog. *Person x Umwelt-Ansatzes* soll eine bessere Vorhersage von Berufs- und Schulerfolg möglich sein. Allerdings nimmt der Ansatz an, dass jede einzelne Persönlichkeitseigenschaft weniger als 10 % der Verhaltensvarianz aufklärt bzw. vorhersagt. Schimmack (2010) kritisiert, dass die meisten Studien zur prädiktiven Vali-

dität von Persönlichkeitseigenschaften jedoch nur auf Selbsteinschätzungen basieren. Wenn man die moderate Validität von Selbsteinschätzungen bedenke, sei anzunehmen, dass die auf diese Weise ermittelten Vorhersagewerte den eigentlichen prädiktiven Wert von Persönlichkeitseigenschaften unterschätzen. Es müsse somit ein weiteres bzw. erweitertes Kausalmodell geben, das diese Effekte besser erklären kann.

Barrick et al. (2001) schlagen drei allgemeine motivationale Konstrukte als Grundlage zur Modellbildung und weiteren Erforschung der Zusammenhänge zwischen Persönlichkeit und Leistung vor, die sie mit *Getting ahead*, *Getting along* und *Getting things done* bezeichnet haben. Unter *Getting ahead* oder auch *Statusorientierung* verstehen die Autoren ein allgemeines Machtstreben einer Person nach Leistung und Status in einer Organisation bzw. der Gesellschaft. *Getting along* bzw. die *Geselligkeitsorientierung* beschreibt das Streben einer Person nach Akzeptanz und Freundschaft mit anderen bzw. den Kollegen. *Getting things done* bzw. *Aufgabenorientierung* betont die nicht-soziale Motivationskomponente. Diese entsteht aus dem persönlichen Streben einer Person nach Exzellenz bzw. einem Expertendasein. Es beschreibt die allgemeine Motivation einer Person, eine Aufgabe zu erfüllen.

Ein theoretisches Modell, das sich explizit mit dem Zusammenhang von Persönlichkeitseigenschaften zur Leistungsvorhersage im Bereich der frühen Arbeitstätigkeit bzw. Aufnahme eines Berufes beschäftigt, liegt bis dato ebenfalls nicht vor. Dieser Aspekt wurde seitens der empirischen Forschung bisher eher ausgeklammert. Neuere Forschung deutet allerdings an, dass sich Karriereerfolg im Laufe des Lebens bzw. der Berufstätigkeit verändert. So ist Karriereerfolg im mittleren und späteren Alter laut Rode et al. (2008) eine Funktion aus Fähigkeiten, Persönlichkeit und Möglichkeiten sowie deren Interaktion. Die Faktoren, die frühen Karriereerfolg bestimmen, müssen den Autoren zufolge noch näher bestimmt werden. So bringen auch neue Mitarbeiter ihre Fähigkeiten und Persönlichkeit in ihre Arbeit ein, während jedoch ihre Möglichkeiten durch z.B. ein noch fehlendes, gut etabliertes soziales Netzwerk noch begrenzt sind. Zudem ist die frühe Berufstätigkeit durch eine hohe Mehrdeutigkeit charakterisiert. Bestimmte Persönlichkeitsausprägungen wie beispielsweise eine hohe Selbstwirksamkeit können demgegenüber von großem Vorteil sein, wenn es z.B. um den (erfolgreichen) Umgang mit neuen Situationen und Anforderungen geht. Es ist im Sinne der Prävention sehr wichtig, die Erfolgsfaktoren in der Phase des Arbeitsbeginns zu kennen, da Individuen, die zu Beginn ihrer Karriere nicht erfolgreich sind, auch im späteren Berufsleben mit bestimmten Nachteilen zu kämpfen haben (Dreher & Bretz, 1991). Allerdings gibt es bisher nur wenige Forschungsarbeiten, die das Verhältnis von Karriereerfolg und Fähigkeiten sowie Eigenschaften einer Person zu Beginn ihrer Arbeitstätigkeit untersucht haben. Neben der Studie von Rode et al. (2008) untersuchten O'Reilly und Chatman (1994) MBA-Absolventen drei bis vier Jahre nach ihrem Universitätsabschluss. Sie fanden, dass die Interaktion zwischen allgemeiner Intelligenz und Gewissenhaftigkeit das Lohnniveau und die Anzahl der Beförderungen in der frühen Karriere der Absolventen vorhersagen konnte. Allerdings konnten sie nicht zeigen, dass allgemeine Fähigkeiten oder Persönlichkeitseigenschaften Berufserfolg vorhersagen. Ebys, Butts und Lockwood (2003) fanden, dass eine proaktive Persönlichkeit und die Big 5-Dimension Offenheit positiv mit dem subjektiv wahrgenommenen Erfolg von Bachelor-Absolventen fünf Jahre nach deren Abschluss zusammenhing. Rode et al. (2008) fanden Korrelationen zwischen Gewissenhaftigkeit und extrinsischem Karriereerfolg ($r = .26$, $p < .01$) sowie proaktiver Persönlichkeit ($r = .41$, $p < .01$) in einer kleinen Absolventenstichprobe ($N = 59$). Zusammenhänge zwischen Erfolgsmaßen und kognitiven Fähigkeiten fanden sie jedoch nicht. Diese Ergebnisse widersprechen somit der Annahme der PPIK-Theorie von Ackerman (1996), dass insbesondere zu Beginn der Berufstätigkeit kognitive Fähigkeiten besonders wichtig sind und lernstarke Personen lernschwachen gegenüber einen Vorteil bei der Aneignung von beruflichem Wissen besitzen sollten. Zu ähnlichen Ergebnissen kamen Ng, Eby, Sorensen und Feldman (2005) in ihrer Metaanalyse. Laut ihren Ergebnissen war

Karriereerfolg (Gehalt) in mittlerem Alter konsistent und positiv mit kognitiven Fähigkeiten verbunden. Das heißt, je intelligenter ein Mitarbeiter war, desto besser wurde er auch bezahlt. Früher Karriereerfolg korrelierte demgegenüber nur mit Persönlichkeitseigenschaften positiv. Allgemeine kognitive Fähigkeiten scheinen zu dieser Zeit offenbar weniger wichtig für beruflichen Erfolg zu sein. Dies könnte den Autoren zufolge u.a. auf die besondere Supervision von Neulingen zurückzuführen sein. Judge et al. (1999) fanden, dass eine erhöhte Gewissenhaftigkeit im Kindesalter mit höherem Berufserfolg im Erwachsenenalter einherging.

Allgemein tendieren Gewissenhaftigkeit, Offenheit, Extraversion und Ausgeglichenheit dazu, sowohl mit extrinsischen (= Gehalt, Beförderungen) als auch intrinsischen Erfolgsmaßen (= berufliche Zufriedenheit) zu korrelieren (Barrick & Mount, 1991; Judge et al., 1999; Ng et al., 2005). Gewissenhaftigkeit ist im beruflichen Kontext neben der Tendenz allgemein gut organisiert zu sein, auch ein Trend dahingehend, effektiver die hohen Informationsmengen bzw. -dichte in neuen und unbekannten Arbeitsbedingungen bewältigen zu können.

Verschiedene Untersuchungen weisen darauf hin, dass je nach Kriterium unterschiedliche Persönlichkeitsdimensionen und -aspekte zur Vorhersage von beruflichem und schulischem Erfolg bzw. Leistung dienen. Vor allem die Gewissenhaftigkeitsdimension bzw. deren Subfacetten und das Selbstwirksamkeitskonstrukt haben sich als valide und zuverlässige Prädiktoren von beruflichem und schulischem Erfolg erwiesen (vgl. Abschnitte 2.3 ff und 2.7 ff). Persönlichkeitseigenschaften, v.a. Gewissenhaftigkeit und Selbstwirksamkeit, weisen darüber hinaus auch inkrementelle Validität über allgemeine kognitive Fähigkeiten hinaus in Bezug auf die Vorhersage von Leistung auf (Chamorro-Premuzic & Furnham, 2008). Diese Zusammenhänge wurden häufig dadurch versucht zu erklären, dass starkes Durchhaltevermögen, Selbstdisziplin und eine zielorientierte Natur sowie eine gute Selbsteinschätzungsfähigkeit erfolgreiche Schüler auszeichnen oder diese Fähigkeiten als „Kompensationsmöglichkeiten“ für weniger intellektuell begabte Schüler angesehen werden können (vgl. hierzu auch Chamorro-Premuzic & Furnham, 2004, 2006). Diese Ergebnisse beruhen jedoch v.a. auch auf korrelativen Zusammenhängen zwischen kognitiver Leistung und Persönlichkeitsdimensionen. Der Nachteil dieser Befunde ist, dass aus diesen Ergebnissen keine kausalen Schlüsse gezogen werden können, welches die beeinflussende Variable bzw. Ursache der Unterschiede ist und warum die Variablen miteinander korrelieren (Moutafi et al., 2004). Es sollten somit auch Prozessmodelle entwickelt werden, die explizit die Einflüsse definieren, die zwischen Schulleistung und Persönlichkeit wirken und die individuellen Unterschiede in der Leistung bewirken.

Die Mehrheit der Studien bezüglich der prädiktiven Validität von Persönlichkeitsdimensionen in Bezug auf berufliche Leistung hat, wie bereits beschrieben, mehrheitlich die Big 5-Persönlichkeitsdimensionen untersucht. Hurtz und Donovan (2000) haben eine Metaanalyse veröffentlicht, die unter anderem die prädiktive Validität der Big 5 in Bezug auf verschiedene Berufsgruppen analysiert. Ihren Ergebnissen zufolge zeigte die Gewissenhaftigkeitsdimension von allen Big 5 die höchsten Zusammenhänge sowohl bei Verkäufern ($r = .18$), Kundenberatern ($r = .17$) und Managern ($r = .11$) als auch bei angelernten Mitarbeitern ($r = .10$). Dudley et al. (2006) haben die Metaanalyse von Hurtz und Donovan (2000) mit den Facetten der Big 5 repliziert. Viele weitere Metaanalysen bestätigen die Vorhersagekraft von Persönlichkeitseigenschaften, vor allem von Gewissenhaftigkeit und dessen Facetten in Bezug auf berufliche und schulische Leistung (z.B. O'Connor & Paunonen, 2007).

In vielen Studien wiesen Persönlichkeitseigenschaften zudem insgesamt deutlich höhere prädiktive Validitäten auf, wenn sie gemeinsam zur Vorhersage der Leistungskriterien analysiert wurden (Chamorro-Premuzic & Furnham, 2008). Angelehnt an Brunswiks (1943) *Linsenmodell der optimalen Vorhersage*, haben sich zusammengesetzte Persönlichkeitsvariablen als vorteilhaft für die Vorhersage von allgemeinen Kriterien wie beispielsweise Arbeitsleistung herausgestellt. So fand Bipp (2010), dass verschiedene Selbsteinschätzungsstrukturen über

die Big 5 hinaus einen bedeutsamen Anteil der Varianz bei der Vorhersage von beruflichen Kriterien (intrinsischer Arbeitsmotivation) aufklären konnten. Die Gewissenhaftigkeitsdimension korrelierte mit fast allen in der Studie erfassten intrinsischen Faktoren von Arbeitsmotivation (außer Feedback). Allerdings war dieser Effekt nicht in allen Regressionen sichtbar.

Lord (2011) geht in ihrer Forderung nach einem umfassenden Erklärungsansatz noch einen Schritt weiter. Sie fordert ein differentielles diagnostisches Modell, das die verschiedenen Merkmale berücksichtigt, die einen Arbeitsplatz von anderen unterscheiden. Die Bedeutung eines solchen Modells liegt der Autorin zufolge darin, dass es verdeutliche, dass die Beziehung zwischen Persönlichkeit und Arbeitsleistung nicht einfach sei. Es genüge demzufolge nicht, pflichtbewusst und genau Persönlichkeitsvariablen mit Stellenanforderungen abzugleichen, denn am Ende werde das Ausmaß, in dem Mitarbeiter mit passenden Persönlichkeitseigenschaften tatsächlich den erwarteten Nutzen liefern, davon bestimmt, in welchem Maße die Arbeitsumgebung den Ausdruck eigenschaftsbezogener Verhaltensweisen, die für eine effektive Leistungen relevant sind, erleichtert oder gehemmt werde. Die Relevanz eines theoretischen Modells beziehe sich folglich darauf, wie Arbeitssituationen gestaltet werden könnten, um den Ausdruck erwünschter Verhaltensweisen zu erleichtern. Der Autorin zufolge wirke sich die individuelle Persönlichkeit einer Person auch darauf aus, wie jemand die Anforderungen einer bestimmten Rolle interpretiere. Dies führe in mehrdeutigen Situationen bzw. bei nicht eindeutig definierten Aufgaben dazu, dass Personen mit verschiedenen Merkmalsausprägungen diese ganz unterschiedlich interpretieren.

Zusammenfassend lässt sich sagen, dass weiterhin Einschränkungen der bisherigen theoretischen Ansätze durch beispielsweise die Verwendung unspezifischer, globaler Traits bei der Vorhersage von Leistung bestehen. Entgegen der nur geringen theoretischen Basis für die bessere Varianzaufklärung, bestärken unabhängige Forschungsarbeiten (z.B. Denis, Morin & Guindon, 2010) und Metaanalysen (z.B. Dudley et al., 2006) die Position, dass spezifischere Facetten besser zur Vorhersage von Leistungskriterien geeignet sind als globale Faktoren. Facetten sagen kontextuelle Leistung voraus und könnten somit auch Aufgabenleistung vorhersagen. Berufsbezogen formulierte Persönlichkeitsvariablen können aufgabenbezogene Leistung und Kontextleistung sogar noch präziser vorhersagen. Es gibt verschiedene Facetten-Leistungs-Zusammenhänge in Abhängigkeit vom Grad der Aufgabenkomplexität und -struktur. Auch Judge und Ilies (2002) bemängelten, dass es bisher keinen adäquaten theoretischen Rahmen gibt, der erklärt, wie Persönlichkeit und Motivation zusammen hängen und gemeinsam Leistung beeinflussen. Dudley et al. (2006) haben in ihrer Metaanalyse auch den Bedarf an genaueren, aussagekräftigeren Vorhersagemodellen thematisiert. Einer differenzierteren Analyse der Zusammenhänge der verschiedenen Persönlichkeitsvariablen haben sich bisher nur wenige Forschungsarbeiten gewidmet (z.B. Dreier, 2008; Korinteli, 2009; Struckmeier, 2009). Dies soll in dieser Studie nachgeholt werden. Der nachfolgende Abschnitt (2.8.2) beinhaltet daher die Modellbildung zur Vorhersage von Ausbildungserfolg mittels differenzierter Gewissenhaftigkeitsfacetten und Selbstwirksamkeit sowie weiteren Selbsteinschätzungsstrukturen.

2.8.2 Modellbildung zur Vorhersage von Ausbildungserfolg

Im Fokus dieses Abschnittes liegt die Entwicklung eines theoretischen Modells zur Vorhersage von Ausbildungserfolg mittels kontextualisierter und differenzierter Persönlichkeitseigenschaften unter Berücksichtigung möglicher weiterer Einflüsse durch beispielsweise Mediatoren oder Moderatoren.

Abbildung 2-5 veranschaulicht schematisch, wie mit Hilfe der Prädiktoren, den Gewissenhaftigkeitsaspekten Fleiß und Ordnung, das Kriterium Ausbildungserfolg vorhergesagt werden soll. Mögliche ergänzende Einflüsse, d.h. prädiktive bzw. medierende und/ oder moderierende

rende Effekte durch die Selbsteinschätzungsvariablen Selbstwirksamkeit sowie wahrgenommene eigene Fähigkeiten, werden ebenfalls berücksichtigt.

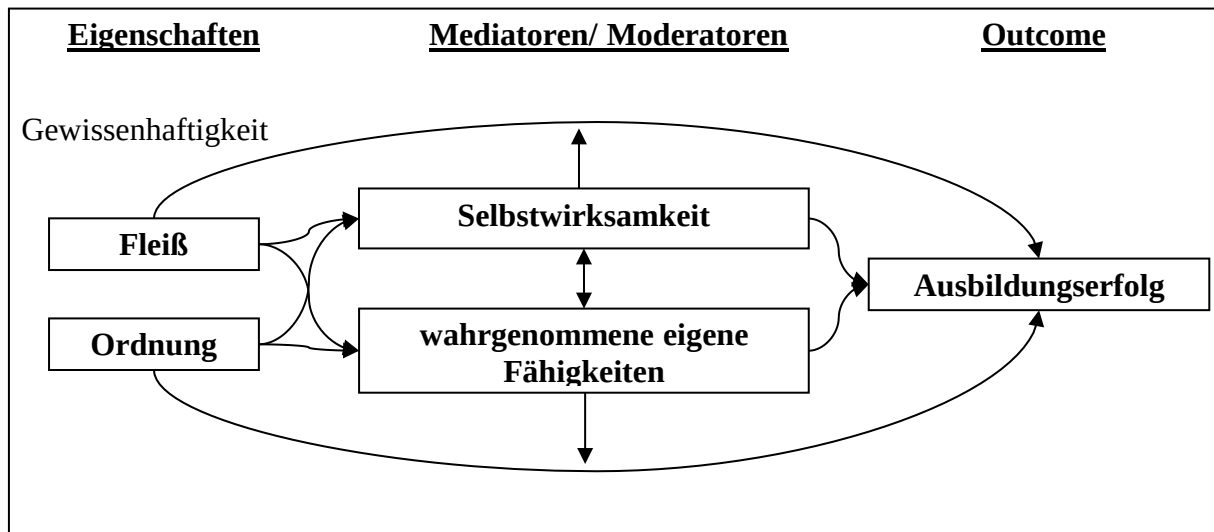


Abbildung 2-5: Schematisches Modell zur Vorhersage von Ausbildungsleistung

Die prädiktive Validität von Gewissenhaftigkeit bzw. dessen Facetten sowie Selbstwirksamkeit in Bezug auf berufliche und schulische Leistungskriterien wurde bereits in den Abschnitten 2.3.1 und 2.3.2 sowie 2.7 ff ausführlich dargestellt und wird daher an dieser Stelle nur ansatzweise noch einmal aufgegriffen.

Gewissenhafte Menschen bewältigen eine Menge an Arbeit, unabhängig davon, ob sie nun nach Gemeinschaft oder nach Status streben. Corker et al. (2012) benennen drei Arten, wie Gewissenhaftigkeit mit Leistung zusammenhängt. Einerseits korreliere Gewissenhaftigkeit mit Leistung im Sinne einer einfachen Vorhersage. Weiter hänge Gewissenhaftigkeit mit Leistungsmotivation zusammen und beeinflusse somit auch indirekt die Leistung. So beeinflusse die jeweilige Gewissenhaftigkeitsausprägung einer Person auch die Auswahl der persönlichen Leistungsziele, was zudem durch die individuelle Selbstwirksamkeitsausprägung sowie die wahrgenommenen eigenen Fähigkeiten einer Person beeinflusst werde. Gewissenhaftigkeit und Selbstwirksamkeitserwartungen korrelieren miteinander ($r = .22$; Schuler & Höft, 2006). Zudem werden Personen, die allgemein über eine höhere Ausprägung der Gewissenhaftigkeit verfügen, mit einer Reihe von Attributen beschrieben, die im Arbeitsbereich von Vorteil sind. Hierzu gehört auch eine höhere Ausprägung der Selbstwirksamkeit (Lee & Klein, 2002; Minbashian, et al., 2010).

Die Aspekte von Gewissenhaftigkeit sagen Leistung unterschiedlich gut voraus. So fanden beispielsweise Jackson et al. (1996), dass *Achievement* akademische Leistung vorhersagte ($r = .24$, $p < .02$), während *Methodicalness* dieses Muster in einer allgemeinen Stichprobe von Studenten nicht zeigte ($r = -.01$, $p > .05$). In einer anderen Stichprobe von Kunststudenten zeigte sich demgegenüber ein negativer Zusammenhang mit *Methodicalness* ($r = -.22$, $p < .05$). Dieses Muster konnte allerdings nicht für *Achievement* gezeigt werden ($r = -.07$, $p > .05$). Diese Ergebnisse sowie die bereits in Abschnitt 2.3.1 beschriebenen differenziellen Zusammenhänge bzw. Vorhersagekraft der Gewissenhaftigkeitsaspekte sprechen für eine differenzierte Erfassung der Persönlichkeitsdimension Gewissenhaftigkeit im Kontext der akademischen und beruflichen Leistungsvorhersage.

Allgemein lässt sich sagen, dass das Streben nach Leistung, wie es mit dem Fleißaspekt erfasst wird, einen Einfluss darauf hat, wie eine Person mit Aufgaben zurechtkommt, die unabhängig von anderen bewältigt werden müssen (Lord, 2011). Leistungsstreben ist der Autorin zufolge am engsten mit Gewissenhaftigkeit verbunden. „Fleißige“ Menschen tendieren dazu,

bei der Ausführung von Aufgaben zielstrebig und effizient zu sein. Hohe Ausprägungen im Aspekt Fleiß sollten daher mit einer produktiveren Arbeitsleistung und besseren Berufsschulnoten verknüpft sein, da Auszubildende mit hohen Fleißausprägungen beispielsweise über gute Lernstrategien verfügen und sich effektiver auf Prüfungen vorbereiten können. Höhere Gewissenhaftigkeits- bzw. Fleißausprägungen deuten zudem auf eine stärkere Ausprägung des Selbstvertrauens hin, was die Verbindung zum Selbstwirksamkeitskonstrukt rechtfertigt.

Für den Ordnungsaspekt ergeben sich gemäß den Ergebnissen von z.B. Jackson et al. (1996) teilweise gegensätzliche Annahmen. Allgemein beschreibt der Ordnungsaspekt von Gewissenhaftigkeit, wie strukturiert und planvoll eine Person bei der Arbeit vorgeht (für eine Definition des Aspekts s. Abschnitt 4.1). „Ordentliche“ Personen tendieren dazu, sich bei der Bearbeitung von Aufgaben an bestehenden Vorschriften und Regeln zu orientieren und diese auch dann einzuhalten, wenn ein Abweichen von diesen ggf. zu einer effektiveren Arbeitsweise beiträgt. Hohe Ausprägungen im Aspekt Ordnung sollten daher nicht unbedingt mit einer produktiveren Arbeitsleistung oder besseren Schulnoten verbunden sein, da dieses Verhalten auch in Pedanterie und Zwanghaftigkeit abrutschen kann. Personen, die Details zu viel Aufmerksamkeit schenken und sich ausschließlich mit diesen beschäftigen, verlieren den Überblick und somit die übergeordneten Ziele aus den Augen (Le et al., 2011). Entsprechend fanden Minbashian et al. (2010) in ihrer Feldstudie, dass hoch gewissenhafte Personen weniger in der Lage waren, ihr Verhalten in Bezug auf veränderte Aufgabenanforderungen anzupassen. Als Erklärung für diese Effekte vermuten sie, dass hoch-gewissenhafte Manager weniger flexibel und in höherem Maße rigide sind als ihre weniger gewissenhaften Kollegen. Die Autoren nehmen an, dass diese Rigidität sich auch im täglichen Verhalten von Personen niederschlagen könnte und auf diese Weise erklärt, warum Trait-Gewissenhaftigkeit inkonsistente Ergebnisse bzw. geringe Validitäten hinsichtlich der Vorhersage von Leistungsdaten zeigt, die Anpassungsfähigkeit erfordern. Demgegenüber könnten auch sehr geringe Ausprägungen von Ordnung negativ mit Leistung zusammenhängen, da das Fehlen von Strukturen ebenfalls zu Minderleistungen führen kann, wenn Personen beispielsweise Termine für Prüfungen vergessen oder keine adäquaten Lernstrategien entwickelt haben.

Eine Verbindung des Ordnungsaspektes mit dem Selbstwirksamkeitskonstrukt wurde bisher noch nicht überprüft. Es ist aber denkbar, dass das Wissen um das Vorhandensein und die Nutzung von beispielsweise Strukturierungshilfen mit Selbstwirksamkeit im Sinne eines Wissens um Bewältigungsmechanismen von Aufgaben in Zusammenhang stehen könnte. Dies könnte sich positiv auf die Ausprägung des Selbstvertrauens auswirken und zu höheren Leistungen führen.

Insgesamt bleibt festzuhalten, dass Fleiß und Ordnung unterschiedliche Aspekte von Gewissenhaftigkeit erfassen und somit bezüglich der untersuchten Kriterien unterschiedliche Vorhersagewerte besitzen könnten. Auch andere Forscher haben eine Abkehr vom starren Beharren auf den Big 5-Persönlichkeitseigenschaften gefordert, da die einzelnen Facetten der Eigenschaften teilweise eine größere Vorhersagekraft besitzen als die Dimensionen (z.B. Hülshager & Maier, 2008). Nun gilt es, diese Unterebene für Gewissenhaftigkeit eindeutig zu identifizieren und auf Basis eines theoretischen Fundaments mit Leistungskriterien in Verbindung zu bringen.

Selbstwirksamkeit ist im organisationalen Kontext ein wichtiges Konstrukt. Unabhängige Studien sowie zwei Metaanalysen haben gezeigt, dass Selbstwirksamkeitserwartungen positiv mit beruflicher Leistung zusammenhängen (Judge & Bono, 2001; Sadri & Robertson, 1993; Stajkovic & Luthans, 1998). Rigotti et al. (2008) zufolge kann Selbstwirksamkeit als Prädiktor für Arbeitszufriedenheit, Commitment und Leistungsvariablen dienen und ist aus diesen Gründen sowohl für die Personalauswahl als auch für die Personalentwicklung in Unternehmen von Bedeutung. Da diese Menschen länger an Aufgaben arbeiten, sich auch gegen

Widerstände durchsetzen (Bandura, 1997) und sich selbst höhere Ziele setzen (Bandura & Wood, 1989), ist es wahrscheinlich, dass hoch selbstwirksame Bewerber beruflich erfolgreicher sind als ihre weniger selbstüberzeugten Mitstreiter. Eine alternative Erklärung für diese Ergebnisse wäre, dass demgegenüber Mitarbeiter bei weniger vorhandenen Entwicklungsmöglichkeiten unzufriedener sind. In jedem Falle ist Selbstwirksamkeit aber ein wichtiges Konstrukt, das im Rahmen von Personalauswahl und -entwicklung stärker berücksichtigt werden sollte. Auch schulisches und berufsbezogenes Wahlverhalten hängt mit beruflicher Selbstwirksamkeit zusammen, ebenso wie Leistung und Persistenz (Hackett & Betz, 1995; Sadri & Robertson, 1993). Ebenso konnten Mediator- und Moderatoreffekte der Selbstwirksamkeit auf die Beziehung von Tätigkeitsmerkmalen und persönlicher Initiative nachgewiesen werden (Speier & Frese, 1997).

Selbstwirksamkeit soll in das Modell zur Vorhersage von beruflicher bzw. schulischer Entwicklung in der Berufsstartphase von Auszubildenden übernommen werden, da insbesondere in Situationen mit wenig Erfahrungshintergrund die Einschätzung des eigenen Fähigkeitspektrums und Verhaltens eine wichtige Rolle spielt. Entsprechend dem sozial-kognitiven Modell (Lent et al., 1994) sollte auch berufliches Selbstwirksamkeitsempfinden einen großen Einfluss auf den Berufsstart haben.

In der Forschung zu Selbstwirksamkeit haben viele Studien auch die Big 5-Persönlichkeitsdimensionen mit dem Selbstwirksamkeitskonzept in Zusammenhang gebracht (z.B. Judge & Illies, 2002). Beispielsweise fanden Judge et al. (1998) in ihrer Metaanalyse hohe Zusammenhänge zwischen Neurotizismus und Selbstwirksamkeit. Sie konnten auch einen positive Zusammenhänge mit Arbeitszufriedenheit ($r = .45$) sowie Arbeitsleistung zeigen ($r = .23$). Wie bereits Locke und Latham (2002) postulierten, ist Selbstwirksamkeit bei der Entwicklung von Aufgabenstrategien wichtig, die bei der Erreichung schwieriger Ziele essenziell sind. Bandura (1997) argumentierte, dass die Vorhersagekraft von Selbstwirksamkeit die prädiktive Validität anderer distaler Variablen immer übertreffen solle, da diese immer konditional, d.h. abhängig von der Arbeitserfahrung, sei. Er postulierte, dass allgemeine Traits nicht in der Lage seien, Verhalten vorherzusagen, wenn man diese um Selbstwirksamkeit kontrolliere. Judge et al. (2007) analysierten diesbezüglich ein komplexes Pfadmodell, das distale Eigenschaften wie allgemeine kognitive Fähigkeiten, Persönlichkeitseigenschaften und individuelle Erfahrungen mit Selbstwirksamkeit in Bezug auf Arbeitsleistung in Verbindung bringt. Ihren Ergebnissen zufolge sagten Gewissenhaftigkeit ($r = .26$, $p < .01$), allgemeine kognitive Fähigkeiten ($r = .52$, $p < .01$) und Erfahrungen ($r = .26$, $p < .01$) Leistung vorher. Wenn Selbstwirksamkeit in das Modell mit einberechnet wurde, ergab sich kein signifikanter Einfluss zwischen den distalen Variablen und Arbeitsleistung. D.h. obwohl Selbstwirksamkeit moderat mit Leistung korrelierte, nahm die Vorhersagekraft des Konstrukts stark ab, wenn individuelle Unterschiede mit berechnet wurden. Dies erklärten die Autoren damit, dass es viele direkte Verbindungen zwischen den distalen Variablen und Arbeitsleistung gibt, die unabhängig von Selbstwirksamkeit die Leistung einer Person beeinflussen können. So führe beispielsweise eine hohe Gewissenhaftigkeit dazu, dass sich Menschen ambitioniertere Ziele setzen und an der Erreichung dieser Ziele härter arbeiten. Je nachdem, welche weiteren Variablen als Moderatoren gewählt wurden, beeinflusste dies das Verhältnis zwischen Selbstwirksamkeit und Arbeitsleistung. Höhere Zusammenhänge ergaben sich beispielsweise, wenn Leistung objektiv gemessen wurde oder ein kleines Zeitintervall zwischen der Messung der Prädiktoren und Kriterien gewählt wurde. Die Ergebnisse dieser Studie besagen somit, dass allgemeine, individuelle Unterschiede genauso wichtig sind wie Selbstwirksamkeit. Selbstwirksamkeit moderierte nur teilweise den Einfluss der distalen Variablen auf die Leistung. D.h. die Vorhersagekraft der distalen Variablen bzw. eventuelle moderierende Effekte basieren eher auf der Situation oder sind davon abhängig. Selbstwirksamkeit scheint eher ein Prädiktor engerer Leistungsindikatoren zu sein, wie z.B. Aufgabenerledigung, wenn man die aufgabenspezifi-

sche Struktur des Konstrukts bedenkt. Insgesamt scheint Selbstwirksamkeit den Zusammenhang der distalen Variablen und Arbeitsleistung nicht zu moderieren. Dennoch ist es beispielsweise möglich, dass Selbstwirksamkeit mit Gewissenhaftigkeit interagiert, um Leistung vorherzusagen oder dass eine hohe Ausprägung in Selbstwirksamkeit v.a. für Menschen mit hoher Gewissenhaftigkeit von Vorteil ist. Demgegenüber ist es auch möglich, dass Selbstwirksamkeit für hoch gewissenhafte Menschen weniger wichtig ist, weil diese Menschen bereits über Eigenschaften verfügen, die Aufgabenerledigung fördern, wie z.B. eine starke Zielorientierung oder eine Tendenz, sich höhere Ziele zu setzen. Eine weitere Erforschung von Selbstwirksamkeit als Moderator sollte daher erfolgen.

Kanfer und Heggstad (1997) forderten, dass Forschung erfassen sollte, bis zu welchem Grad Traits und Selbstwirksamkeit miteinander Arbeitsverhalten vorhersagen können bzw. beeinflussen. Denn es gibt bisher nur wenige Versuche, die inkrementelle oder alleinige Vorhersagekraft von Selbstwirksamkeit zu bestimmen (Judge et al., 2007). Den Ergebnissen von Kanfer und Heggstad (1997) zufolge waren allgemeine kognitive Fähigkeiten ($\beta = .38$) und Gewissenhaftigkeit ($\beta = .28$) prädiktiver für Leistung als Selbstwirksamkeit. Nur wenn die untersuchten Personen bereits über Arbeitserfahrung verfügten, sagte Selbstwirksamkeit signifikant Leistung voraus ($\beta = .21$), aber auch in dieser Bedingung waren wieder die allgemeinen kognitiven Fähigkeiten ($\beta = .53$) und Gewissenhaftigkeit ($\beta = .30$) die valideren Prädiktoren.

Die bisherige Annahme in der Theorie war jedoch immer, dass Selbstwirksamkeit die Effekte zwischen Gewissenhaftigkeit und Leistung mediiert (z.B. Kanfer, 1992, Dreier, 2008). Entsprechend postulieren Martocchio und Judge (1997), dass „*Self-efficacy represents the mechanism through which the generalized tendencies of conscientiousness manifest themselves*“ (S. 766). Chen et al. (2001) überprüften ein metaanalytisches Modell, ob Selbstwirksamkeit den Zusammenhang zwischen kognitiven Fähigkeiten und Gewissenhaftigkeit zur Vorhersage von Berufsleistung mediiert. Sie fanden, dass dies von der Jobkomplexität abhing – je einfacher die Arbeit, desto höher war der Zusammenhang. Um das Verhältnis von Selbstwirksamkeit und Leistung zu verstehen und der sozial-kognitiven Theorie gerecht zu werden, muss man kontextuelle Faktoren mit bedenken, wenn man Berufsleistung vorhersagen will. Larson und Borgen (2006) sowie Ozer und Benet-Martinez (2006) sahen die distalen Persönlichkeitseigenschaften als „Treiber“ von Selbstwirksamkeit zur Vorhersage von Leistung an, in dem Sinne, dass sich zuerst die Persönlichkeit bei einer Person entwickelt und danach erst die Selbstwirksamkeit. Larson und Borgen untersuchten verschiedene Stichproben von College-Studenten ($N_{ges} = 1.173$) und fanden, dass die Gewissenhaftigkeitsfacette Achievement signifikant mit Selbstwirksamkeit korrelierte, während dies bei der Control-Facette nicht der Fall war. Die individuellen Persönlichkeitsmuster interagierten mit der Umgebung sowie den Fähigkeiten und Interessen einer Person, was in einer unterschiedlichen Qualität der Aneignung von Fähigkeiten resultierte. Auch wenn Selbstwirksamkeit und Persönlichkeit historisch betrachtet, getrennt voneinander entwickelt wurden, ist die Verbindung der beiden Konzepte miteinander mittlerweile deutlich.

Insgesamt tragen Gewissenhaftigkeit bzw. die Aspekte Fleiß und Ordnung sowie Selbstwirksamkeitserwartungen dazu bei, akademische Ziele zu planen und die Bemühungen sowie die Ausdauer zur erfolgreichen Erreichung dieser Ziele zu bündeln. Es ist wahrscheinlicher, dass Mitarbeiter bzw. Auszubildende mit hohen Ausprägungen in den Gewissenhaftigkeitsaspekten Fleiß und Ordnung eher zielorientiertes, gewilltes und hingebungsvolles Verhalten zeigen und dadurch auch durch Vorgesetzte wiederum positiv(er) bewertet werden (Sutherland et al., 2007).

Die Vorhersagevarianz wird durch Gewissenhaftigkeit bzw. Fleiß und Ordnung sowie Selbstwirksamkeit allerdings nicht hinreichend aufgeklärt (s.o.), weshalb der Einfluss weiterer Variablen als Prädiktoren bzw. Moderatoren oder Mediatoren angenommen wird. Angelehnt

an die Ergebnisse von Noffle und Robins (2007), dass Gewissenhaftigkeit und Leistung nicht nur mit den wahrgenommenen eigenen Fähigkeiten korrelierten, sondern der Zusammenhang der Variablen durch diese darüber hinaus mediert wurde, wird dies als Variable mit in das Vorhersagemodell aufgenommen.

Wichtig ist generell, die genannten Prädiktoren bzw. weiteren beeinflussenden Variablen arbeitsbezogen zu erfassen, da situationale Faktoren bei der Arbeit bzw. in der Berufsschule gewissenhaftes Verhalten fördern oder hemmen können, was wiederum Einfluss auf die Leistungsbeurteilung eines Auszubildenden haben kann. Verhalten ist neben der Persönlichkeit auch von situationalen Faktoren abhängig und basiert daher nicht nur auf den Charaktereigenschaften einer Person allein (Sutherland et al., 2007). Somit sollten die Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie die Selbstwirksamkeit kontextbezogen erfasst werden. Minbashian und Kollegen (2010) fanden entsprechend heraus, dass aufgabenbezogene Gewissenhaftigkeit auch adaptives Verhalten bei Managern positiv vorhersagte. Auf die Vorhersage von Ausbildungsleistung bezogen bedeutet dies, dass Auszubildende mit einer höheren Gewissenhaftigkeitsausprägung auch angesichts steigender Anforderungen und bei erhöhter Komplexität ihrer Aufgabenanforderungen ihr hohes Arbeitsniveau länger bzw. eher aufrechterhalten können, als weniger gewissenhafte Auszubildende.

Sarges (2010) fordert in seinem Artikel darüber hinaus explizit eine stärkere Bezugnahme zur tatsächlichen Arbeitstätigkeit bzw. der Erfassung so genannten „Real-Life-Kriterien“ z.B. mittels empirischer Untersuchungen in Feldstichproben. Diese bezeichnet er als ultimative, praxisrelevante Kriterien. Die Erfassung relevanter Personenmerkmale sollte valider vorgenommen werden, so der Autor, um beispielsweise effektiver „Hoch- und Niedrigleister“ voneinander unterscheiden zu können. Zudem weisen aufgabenbezogen erfasste Gewissenhaftigkeit inkrementelle Validität in Bezug auf die Vorhersage von relevantem Arbeitsverhalten auf (Schimmack, 2010), da durch den Kontextbezug auch individuelle Unterschiede berücksichtigt werden, die durch den traditionellen, generischen Trait-Ansatz nicht beachtet werden, aber dennoch wichtig für die Vorhersage von Leistung sind.

3. Forschungsfragen

Aufbauend auf den Inhalten des vorherigen Kapitels werden im Folgenden zunächst die Ziele der Untersuchung abgeleitet, indem aufgezeigt wird, worin der konkrete Forschungsbedarf besteht (Abschnitt 3.1), und anschließend inhaltliche Hypothesen formuliert. Danach wird auf mögliche konfundierende Variablen eingegangen (Abschnitt 3.2).

3.1 Ableitung der Ziele und inhaltlichen Hypothesen

Wie in den vorhergehenden Abschnitten (2.1 bis 2.7) dargestellt wurde, belegt eine Vielzahl von Primär- und Sekundärstudien die prädiktive Validität von Persönlichkeitsvariablen, allen voran Gewissenhaftigkeit bzw. dessen Aspekte Fleiß und Ordnung sowie Selbstwirksamkeit, in Bezug auf die Vorhersage schulischer und beruflicher Leistungs- sowie Erfolgskriterien. Bezüglich des Verständnisses des Zusammenhangs von Persönlichkeit und Arbeitsleistung wurden in den vergangenen 50 Jahren viele Fortschritte erzielt (Barrick et al., 2001). Allerdings sind Persönlichkeitsmaße trotz ihrer guten Vorhersagekraft sowohl in der Forschungsgemeinde als auch in der eignungsdiagnostischen Praxis stark umstritten. Die Zusammenhänge der untergeordneten Facetten im Zusammenhang mit Arbeitsleistung sind erst seit kurzem in den Fokus der wissenschaftlichen Aufmerksamkeit gerückt. Zudem existiert über die Annahme linearer Zusammenhänge hinaus bislang kein theoretisches Vorhersagemodell, das die verschiedenen Variablen miteinander in Beziehung setzt, um zu einer präziseren Vorhersage von Ausbildungsleistung zu gelangen bzw. die Zusammenhänge zwischen den Variablen zu erklären (Barrick et al., 2001).

Die vorliegende Studie versucht, sich dieser Kritikpunkte anzunehmen. Durch die (Weiter-) Entwicklung eines arbeitsbezogenen und differenzierten Gewissenhaftigkeitsmaßes (Abschnitt 3.1.1) zur Erfassung der Gewissenhaftigkeitsaspekte Fleiß und Ordnung wird versucht, den Problemen mit den bestehenden Persönlichkeits- bzw. Gewissenhaftigkeitsmaßen entgegen zu treten. Zudem wurde ein theoretisches Vorhersagemodell entwickelt, um die komplexen Zusammenhänge zwischen den verschiedenen Selbstbeschreibungskonstrukten erfassen zu können (Abschnitt 3.1.2). Abschließend sollen die prädiktiven Validitäten der arbeitsbezogen erfassten Gewissenhaftigkeitsaspekte Fleiß und Ordnung mit der Vorhersagekraft eines allgemeinen Gewissenhaftigkeitsmaßes verglichen werden (Abschnitt 3.1.3). Im Anschluss an den jeweiligen Abschnitt werden die inhaltlichen Hypothesen abgeleitet und tabellarisch aufgeführt.

3.1.1 Validierung eines kontextbezogenen Gewissenhaftigkeitsmaßes

Wie in Abschnitt 2.3.1 ausführlich dargestellt wurde, hat sich das Persönlichkeitsmerkmal Gewissenhaftigkeit hierarchie- und berufsgruppenübergreifend als valider Prädiktor von Berufserfolg erwiesen (Barrick et al., 2001; Schmidt & Hunter, 1998). Allerdings besitzen die spezifischeren Facetten bzw. Aspekte von Gewissenhaftigkeit eine größere und differenziertere Vorhersagekraft als die Gewissenhaftigkeitsdimension selbst in Bezug auf verschiedene berufliche und schulische Leistungskriterien (O'Connor & Paunonen, 2007).

Zudem besteht, wie in Abschnitt 2.5 erläutert wurde, sowohl seitens der Forschung (Lievens et al., 2008; Pace & Brannick, 2010) als auch seitens der eignungsdiagnostischen Praxis (Moldzio & Partner, 2009) ein Bedarf an kontextualisierten bzw. berufsbezogen formulierten Testverfahren zur Erfassung relevanter Persönlichkeitseigenschaften mit dem Ziel der Leistungsvorhersage im beruflichen Kontext.

Basierend auf dem erweiterten hierarchischen Modell der Big 5-Persönlichkeitseigenschaften von DeYoung und weiteren Forschern (DeYoung, 2010; DeYoung et al., 2007; Paunonen &

Ashton, 2001; Roberts et al., 2005) sowie den Befunden zur erhöhten prädiktiven Validität kontextbezogener Selbstbeschreibungsvorgahren, wurden im Rahmen mehrerer Forschungsarbeiten die Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS) entwickelt (Dreier, 2008; Dreier & Moldzio, 2010, 2012; Dreier, Kasper & Moldzio, 2010; Korinteli, 2009; Struckmeier, 2009), die die Gewissenhaftigkeitsaspekte Fleiß und Ordnung kontextualisiert für den Arbeitsbereich erfassen. Im Rahmen der vorliegenden Untersuchung sollen die AGS nun weiter entwickelt werden und hinsichtlich ihrer Zuverlässigkeit sowie Validität überprüft werden.

Die Entwicklung bzw. Überprüfung eines Fragebogenverfahrens stellt aus psychometrischer Sicht kein Novum dar. Die Überprüfung der Gütekriterien kann gemäß der klassischen Testtheorie erfolgen. Zu fordern ist somit einerseits, dass sich die Items der AGS über eine Itemtrennschärfe und -schwierigkeit verfügen sowie eine reliable und stabile Erfassung der latenten Variablen Fleiß bzw. Ordnung ermöglichen. Auch die faktorielle Validität sowie die Konstruktvalidität der Skalen soll erhoben werden. Eine umfassende Konstruktvalidierung ist notwendig, um die a priori postulierte Faktorenstruktur zu bestätigen.

Die Güteanforderungen lassen sich in Form inhaltlicher Hypothesen formulieren, die Tabelle 3-1 zu entnehmen sind.

Tabelle 3-16: Inhaltliche Hypothesen zur Skalengüte der Arbeitsbezogenen Gewissenhaftigkeitsskalen

Nr.	Annahmen	Kapitel
H1	Überprüfung der AGS auf Itemebene:	2.3.1
	H1a: Nach Kriterien der Itemtrennschärfe und Itemschwierigkeit bilden die AGS zwei Faktoren ab, die den definierten Subfaktoren Fleiß und Ordnung entsprechen.	6.2.1 6.2.2
	H1b: Die AGS, d.h. die beiden Skalen zur Erfassung der Gewissenhaftigkeitsaspekte Fleiß und Ordnung, erweisen sich als reliabel im Sinne einer hohen internen Konsistenz.	
H2	Die beiden Subskalen Fleiß sowie Ordnung erweisen sich als zeitlich stabil im Sinne einer hohen Retest-Reliabilität.	2.3.1 6.2.2.2
H3	Überprüfung des theoretischen Modells der AGS:	2.3.1
	H3a: Die Eindimensionalität innerhalb einer Subskala ist gegeben.	6.2.3
	H3b: Die Items einer Skala laden nur auf die korrespondierende latente Variable. Substantielle Ladungen auf die latente Variable der anderen Skala liegen nicht vor.	
H4	Überprüfung der Konstruktvalidität der AGS (MTMM-Analyse):	2.3.1
	H4a: Zur Messung der Homogenität der Subskalen sollen die mittleren Interitemkorrelationen zwischen .40 und .60 liegen.	6.2.3
	H4b: Es besteht eine mittlere, positive Korrelation zur konstruktkonvergenten Persönlichkeitsdimension Gewissenhaftigkeit.	
	H4c: Die Korrelation zum konstruktdivergenten Faktor Neurotizismus ist negativ und fällt niedrig aus.	
	H4d: Die Korrelation zum konstruktdivergenten Faktor Offenheit für Erfahrungen ist positiv und fällt niedrig aus.	
	H4e: Die Korrelation zu den konstruktdivergenten Leistungsmaßen der LPS-Skalen 3 und 4 ist nicht signifikant von null verschieden.	
H5	Überprüfung der Konstruktvalidität der AGS mittels konfirmatorischer Faktorenanalysen:	2.3.1
	H5a: Die AGS bzw. deren Subskalen Fleiß und Ordnung erweisen sich gegenüber einer weiteren Gewissenhaftigkeitsskala als faktoriell valide. Substantielle Ladungen auf die latente Variable der anderen Skalen liegen nicht vor.	6.2.3
	H5b: Die AGS bzw. deren Subskalen Fleiß und Ordnung erweisen sich gegenüber einer Neurotizismusskala als faktoriell valide.	
	H5c: Die AGS bzw. deren Subskalen Fleiß und Ordnung erweisen sich gegenüber einer Offenheitsskala als faktoriell valide.	
Anmerkungen:	Der Spalte „Kapitel“ kann man entnehmen, in welchem Teil des zweiten Kapitels die theoretischen Hintergründe zu den Hypothesen dargestellt sind, die die jeweilige Annahme untermauern. Zudem ist angegeben, in welchem Teil des sechsten Kapitels die zugehörigen statistischen Methoden dargestellt und erläutert sind.	

3.1.2 Entwicklung und Überprüfung eines Vorhersagemodells

Wie bereits in Abschnitt 2.8 dargestellt wurde, existiert bis dato kein anerkanntes Modell, das die Vorhersage von Ausbildungsleistung durch Persönlichkeitsvariablen erklärt. Zwar gibt es viele Primärstudien und Metaanalysen, die den Zusammenhang bestimmter Persönlichkeitsmaße, v.a. Gewissenhaftigkeit bzw. dessen Subfacetten und Selbstwirksamkeit, mit verschiedenen Leistungsvariablen und die prädiktive Validität dieser Maße in Bezug auf schulische und berufliche Leistungskriterien belegen (z.B. O'Connor & Paunonen, 2007; Schmidt & Hunter, 1998), aber ein komplexeres Vorhersagemodell wurde bisher nicht formuliert oder die genauen Ursache-Wirkungs-Zusammenhänge überprüft (Anderson & Ones, 2003; Barrick et al., 2001). Auch O'Connor und Paunonen (2007) bemängeln, dass vergleichsweise wenige Studien sich mit der komplexeren Erfassung von schulischer Leistung und dessen Indikatoren beschäftigen:

„Most studies have employed a single overall indicator of scholastic achievement as the criterion measure of academic performance, grade point average being the favorite. However, academic performance is not a unitary construct. (...) Thus, investigators are strongly encouraged to examine specific components of academic performance, as well as overall indicators of academic success, in future investigations.” (S. 986).

Das primäre Interesse der vorliegenden Arbeit war somit die Entwicklung und die Überprüfung eines Vorhersagemodells, das die zweifaktorielle Substruktur von Gewissenhaftigkeit in Bezug auf das Kriterium Ausbildungsleistung berücksichtigt und weitere moderierende bzw. mediierende Variablen zur Vorhersage von Ausbildungsleistung zusammenführt. Insbesondere soll hierbei beobachtet werden, inwieweit ein berufsbezogen formuliertes Gewissenhaftigkeitsmaß in der Lage ist, auch Leistungsvariablen beruflich noch unerfahrener Kandidaten vorherzusagen.

In bisherigen Studien zeigten sich eher moderate Vorhersagekennwerte für Gewissenhaftigkeit, die aber vor allem aufgrund überhöhter statistischer Maßstäbe eine zu geringe Beachtung fanden. Vergleicht man diese Kennwerte allerdings mit Zusammenhangsstärken aus anderen Forschungsbereichen wie beispielsweise der Medizin, so erscheinen die gefundenen Effekte deutlich klarer (Hülshager & Maier, 2008). Judge et al. (1999) untersuchten in drei Längsschnittstudien den Einfluss von Persönlichkeitsmerkmalen auf späteren Berufserfolg. Ihre Ergebnisse zeigen, dass sich die prädiktive Validität von Gewissenhaftigkeit über die Zeit akkumulierte. Eine möglichst frühe Erfassung und Berücksichtigung von Gewissenhaftigkeit empfiehlt sich also insbesondere bei Berufsanfängern. Auch Anderson und Ones (2003) forderten weitere Forschungsarbeiten, die untersuchen sollen, wie und warum Persönlichkeitseigenschaften berufliche Erfolgskriterien vorhersagen.

Gemäß der Annahmen der Person x Umwelt-Interaktion sollten zur besseren Vorhersage von Leistung nicht nur Variablen aus der Traittheorie verwendet, sondern auch aus der sozial-kognitiven Lerntheorie herangezogen werden, um eine möglichst gute Vorhersage des Kriteriums zu erreichen. Selbstwirksamkeit (Gore, 2006; Okech & Harrington, 2002) und wahrgenommene eigene Fähigkeiten (Dreier, 2008; Nofle & Robins, 2007) wurden bereits als solche Einflussfaktoren in diesem Prozess identifiziert und werden daher in das Modell mit aufgenommen.

Die genauen Wirkmechanismen dieser Variablen sollen im Rahmen der Überprüfung des theoretischen Vorhersagemodells, welches in Abbildung 3-1 konzeptuell dargestellt ist, überprüft werden. Einerseits steht eine Überprüfung der Wirkungsintensitäten der Prädiktorvariablen Fleiß und Ordnung unter Berücksichtigung eines potenziellen moderierenden Effekts durch Selbstwirksamkeit bzw. wahrgenommene eigene Fähigkeiten in Bezug auf die Kriteriumsvariable Ausbildungserfolg an. Andererseits soll auch überprüft werden, inwiefern

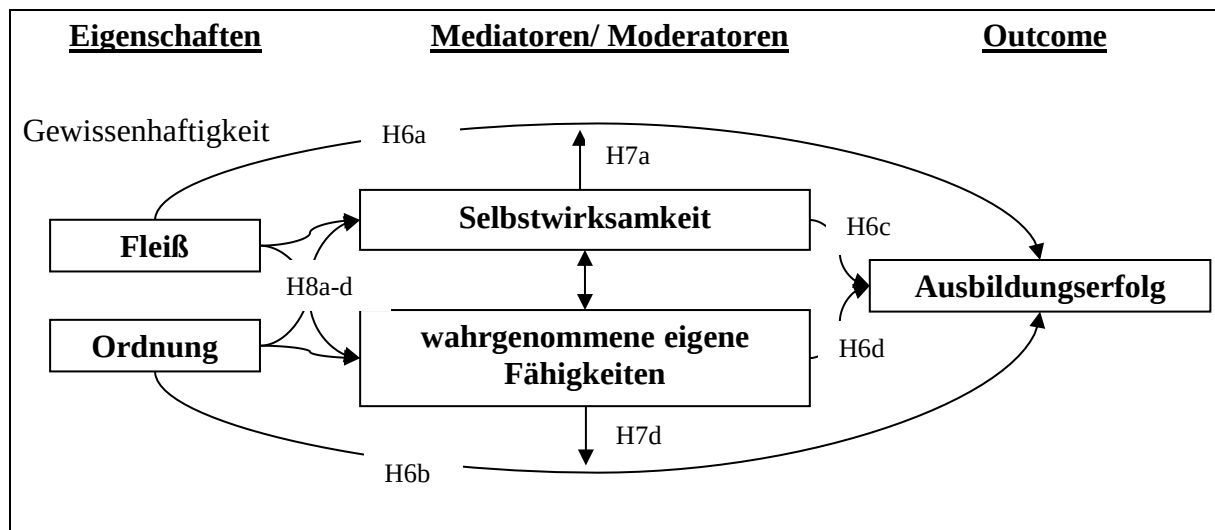


Abbildung 3-1: Konzeptuelles Modell zur Vorhersage von Ausbildungserfolg. (Zur besseren Übersichtlichkeit sind nicht alle Hypothesen eingetragen.)

unter der Hinzunahme von Selbstwirksamkeit bzw. wahrgenommenen eigenen Fähigkeiten als Mediatoren sich die Vorhersagegüte des Fleiß- bzw. Ordnungsaspektes verbessert. Auch Barrick und Mount (2004) besagten, dass

„The fact of the matter is that human behavior at work is complex, and understanding the relation of personality traits and job performance is difficult. Personality traits are enduring, distal forces that influence behavior, but there are both mediating and moderating variables that must be accounted for to adequately explain the effects of personality on human behavior.“ (S. 369).

Entsprechend sollen sowohl moderierende als auch medierende Effekte der beeinflussenden Variablen überprüft werden, um die Zusammenhänge der Variablen möglichst genau überprüfen zu können.

Tabelle 3-2 enthält die inhaltlichen Hypothesen zur empirischen Überprüfung des Vorhersagemodells bzw. zur Vorhersage von Ausbildungsleistung mittels der arbeitsbezogenen Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie Selbstwirksamkeit und wahrgenommenen eigenen Fähigkeiten.

Tabelle 3-2: Inhaltliche Hypothesen zur Überprüfung des Vorhersagemodells

Nr.	Annahmen	Kapitel
H6	Zusammenhänge der UVn und AVn bzw. Vorhersagekraft der untersuchten Variablen.	2.3.1
	H6a: Der Fleißaspekt korreliert mit den Leistungsdaten.	2.3.2
	H6b: Der Ordnungsaspekt korreliert mit den Leistungsdaten.	2.3.3
	H6c: Selbstwirksamkeit korreliert mit den Leistungsdaten.	2.7.2
	H6d: Die wahrgenommenen eigenen Fähigkeiten korrelieren mit den Leistungsdaten.	2.7.3
		6.3.1
H7	Selbstwirksamkeit sowie wahrgenommene eigene Fähigkeiten verändern sich aufgrund der Erfahrungen, die man macht, und beeinflussen somit die Vorhersagegüte des Fleiß- bzw. Ordnungsaspekts.	2.3.2
		2.3.3
		2.8
	H7a: Personen mit hohen Ausprägungen im Fleißaspekt und hohen Ausprägungen im Selbstwirksamkeitskonstrukt sind erfolgreicher in der Ausbildung.	6.3.4
	H7b: Personen mit hohen Ausprägungen im Fleißaspekt, die ihre eigenen Fähigkeiten höher einschätzen, sind erfolgreicher in der Ausbildung.	
	H7c: Personen mit hohen Ausprägungen im Ordnungsaspekt und hohen Ausprägungen im Selbstwirksamkeitskonstrukt sind erfolgreicher in der Ausbildung.	
	H7d: Personen mit hohen Ausprägungen im Ordnungsaspekt, die ihre eigenen Fähigkeiten hoch einschätzen, sind erfolgreicher in der Ausbildung.	
	(Fortsetzung auf nächster Seite)	

Nr.	Annahmen	Kapitel
H8	Unter Hinzunahme von Selbstwirksamkeit bzw. wahrgenommenen eigenen Fähigkeiten klärt Fleiß bzw. Ordnung mehr Varianz im Kriterium auf.	2.3.2 2.3.3
H8a:	Selbstwirksamkeit stellt im postulierten Modell einen Mediator dar, d.h. Fleiß wirkt sich vermittelt über Selbstwirksamkeit auf den Ausbildungserfolg aus.	2.8 6.3.5
H8b:	Selbstwirksamkeit stellt im postulierten Modell einen Mediator dar, d.h. Ordnung wirkt sich vermittelt über Selbstwirksamkeit auf den Ausbildungserfolg aus.	
H8c:	Wahrgenommene eigene Fähigkeiten stellen im postulierten Modell einen Mediator dar, d.h. Fleiß wirkt sich vermittelt über wahrgenommene eigene Fähigkeiten auf den Ausbildungserfolg aus.	
H8d:	Wahrgenommene eigene Fähigkeiten stellen im postulierten Modell einen Mediator dar, d.h. Ordnung wirkt sich vermittelt über wahrgenommene eigene Fähigkeiten auf den Ausbildungserfolg aus.	
Anmerkungen:	Der Spalte „Kapitel“ kann man entnehmen, in welchem Teil des zweiten Kapitels die theoretischen Hintergründe zu den Hypothesen dargestellt sind, die die jeweilige Annahme untermauern. Zudem ist angegeben, in welchem Teil des sechsten Kapitels die zugehörigen statistischen Methoden dargestellt und erläutert sind.	

3.1.3 Gegenüberstellung eines allgemeinen Gewissenhaftigkeitsmaßes mit einer kontextbezogenen Gewissenhaftigkeitsskala

Wie bereits in Abschnitt 2.5 ausführlich dargestellt wurde, erweisen sich kontextualisierte Fragebogenverfahren als validere Instrumente zur Vorhersage beruflicher und schulischer Leistung (Bing et al., 2004; Lievens et al., 2008; Reddock et al., 2010; Schmit et al., 1995). Nur wenn Prädiktor und Kriterium den gleichen Grad an Spezifität aufweisen, kann die Vorhersagegenauigkeit gewährleistet werden. Entsprechend sollten auch die verwendeten diagnostischen Instrumente zur Vorhersage von schulischem und beruflichem Erfolg kontextualisiert sein.

Diese wissenschaftliche Arbeit sieht daher neben der Überprüfung und Evaluation der berufsbezogen formulierten Gewissenhaftigkeitsaspekte Fleiß und Ordnung auch den Vergleich dieser Aspekte mit einem allgemein formulierten Gewissenhaftigkeitsmaß in Bezug auf die prädiktive und inkrementelle Validität vor. Um gegenüber dem gewöhnlich verwendeten, allgemeinen Persönlichkeits- bzw. Gewissenhaftigkeitsmaß einen Vorteil zu haben, müssen die AGS in der Lage sein, die relevanten Kriterien präziser vorherzusagen als das allgemeine, generische Gewissenhaftigkeitsmaß, das sich im Rahmen der praktischen Personalauswahl und -entwicklung sowie in der Forschung bereits etabliert hat. Dies sollte der Fall sein, denn laut Bing et al. (2004) korrelieren kontextualisierte Items stärker mit dem entsprechenden Kriterium, da beide gemeinsame Varianz aufweisen.

Erste Hinweise zur Überlegenheit der prädiktiven Validität der AGS im Vergleich mit einem generischen Maß zur Erhebung der Gewissenhaftigkeitsaspekte *Industriousness* und *Orderliness* des BFAS (DeYoung et al., 2007) sind bereits vorhanden (Struckmeier, 2009). Die Ergebnisse deuten bessere Vorhersagekennwerte der berufsbezogenen Gewissenhaftigkeitsaspekte Fleiß ($r = .10$, n.s.) und Ordnung ($r = -.16$, $p < .05$) im Vergleich mit den generischen Aspekten *Industriousness* ($r = .10$, n.s.) und *Orderliness* ($r = -.08$, n.s.) an.

Die Ergebnisse könnten Implikationen für die zukünftige Entwicklung der eignungsdiagnostischen Praxis und Forschung nach sich ziehen. Es ist zu vermuten, dass die berufsbezogen formulierten Aspekte eine größere Vorhersagekraft besitzen als die allgemeine Skala. Zusätzlich sollten sich unterschiedliche Effekte für die beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung zeigen.

Ein weiterer Punkt bei der Neu- bzw. Weiterentwicklung einer Skala und dem Vergleich ähnlicher Testverfahren ist immer auch der der inkrementellen Validität. Diese bestimmt, inwieweit eine andere bzw. neue Skala zusätzliche Varianz über ein bestehendes, etabliertes In-

strument hinaus aufklärt. Auch aus dem Grund, dass in der Eignungsdiagnostik häufig Verfahren kombiniert werden, ist die Bestimmung der inkrementellen Validität bedeutsam (Hülshager & Maier, 2008). Um die Nützlichkeit des anderen bzw. neuen Verfahrens über das etablierte Maß hinaus nachzuweisen, muss die neu entwickelte Skala bzw. der differenziertere Prädiktor in Bezug auf die Kriteriumsvariable über das andere Konstrukt hinaus zusätzliche Varianz aufklären. Die inkrementelle Validität differenzierter Gewissenhaftigkeitsfacetten über die allgemeine Gewissenhaftigkeitsdimension hinaus wurde von Dudley et al. (2006) umfassend belegt. Den Ergebnissen zufolge klären Gewissenhaftigkeitsfacetten je nach Kriterium zwischen 3,7 % (Berufserfolg allgemein) und 25,9 % (Hingabe zur Arbeit) zusätzliche Varianz auf. Auch für verschiedene Berufsgruppen besitzen die Facetten von Gewissenhaftigkeit von 1,2 % (Kundenberater) bis zu 24 % (angelernte Arbeiter) inkrementelle Validität. Entsprechend wird für die Gewissenhaftigkeitsaspekte Fleiß und Ordnung erwartet, dass diese inkrementelle Validität über das allgemeine Gewissenhaftigkeitsmaß hinaus besitzen.

Zudem sollten die Arbeitsbezogenen Gewissenhaftigkeitsskalen im Vergleich mit dem allgemeinen Gewissenhaftigkeitsmaß eine höhere Reliabilität sowie geringere Streuungen aufweisen, da die Kontextualisierung der Items zu einer Reduktion der intra- und interpersonellen Antwortvarianz führen sollte.

Tabelle 3-3 beinhaltet die inhaltlichen Hypothesen zum Vergleich der AGS bzw. berufsbezogen formulierten Gewissenhaftigkeitsaspekte Fleiß und Ordnung mit einem etablierten allgemeinen Gewissenhaftigkeitsmaß.

Tabelle 3-3: Inhaltliche Hypothesen zum Vergleich der AGS mit einem allgemeinen, generischen Gewissenhaftigkeitsmaß

Nr.	Annahmen	Kapitel
H9	Vergleich der prädiktiven Validität der verschiedenen Gewissenhaftigkeitsmaße:	2.3.1
	H9a:Fleiß – arbeitsbezogen erfasst – sagt Ausbildungserfolg präziser vorher als Gewissenhaftigkeit allgemein.	6.4
	H9b:Ordnung – arbeitsbezogen erfasst – sagt Ausbildungserfolg präziser vorher als Gewissenhaftigkeit allgemein.	
H10	Bestimmung der inkrementellen Validität der AGS:	2.3.1
	H10a:Fleiß besitzt inkrementelle Validität über Gewissenhaftigkeit hinaus.	6.3.5
	H10b:Ordnung besitzt inkrementelle Validität über Gewissenhaftigkeit hinaus.	6.4
	H10c:Fleiß und Ordnung besitzen gemeinsam inkrementelle Validität über Gewissenhaftigkeit allgemein hinaus.	
H11	Vergleich der intra- und interpersonellen Antwortvariabilität:	2.5
	H11a:Die Fleißskala ist reliabler als das allgemeine Gewissenhaftigkeitsmaß.	6.2.1
	H11b:Die Items der Fleißskala weisen eine geringere mittlere Streuung auf als die Items des allgemeinen Gewissenhaftigkeitsmaßes.	6.2.2
	H11c:Die Ordnungsskala ist reliabler als das allgemeine Gewissenhaftigkeitsmaß.	
	H11d:Die Items der Ordnungsskala weisen eine geringere mittlere Streuung auf als die Items des allgemeinen Gewissenhaftigkeitsmaßes.	

Anmerkungen: Der Spalte „Kapitel“ kann man entnehmen, in welchem Teil des zweiten Kapitels die theoretischen Hintergründe zu den Hypothesen dargestellt sind, die die jeweilige Annahme untermauern. Zudem ist angegeben, in welchem Teil des sechsten Kapitels die zugehörigen statistischen Methoden dargestellt und erläutert sind.

3.2 Mögliche konfundierende Variablen

Neben den zentralen Variablen der Untersuchung erfolgte im Rahmen der vorliegenden Studie zudem eine Erhebung weiterer Variablen. Neben allgemeinen demografischen Merkmalen (Abschnitt 3.2.1) sowie beruflichen Informationen (Abschnitt 3.2.2) wurde auch sozial er-

wünschtes Antwortverhalten (Abschnitt 3.2.3) mit erfasst. Diese Variablen sollen als Kontrollvariablen hinsichtlich möglicher konfundierender Effekte untersucht werden.

3.2.1 Demografische Merkmale

Es sollen die demografischen Merkmale *Alter*, *Geschlecht* und *Bildungsstand* der Studienteilnehmer mit erhoben werden, da sie einerseits zur angemessenen Beschreibung der Stichprobe notwendig sind und somit Vergleiche zu anderen Populationen hergestellt werden können. Andererseits ist es auch möglich, dass diese Merkmale als konfundierende Variablen wirken und einen Einfluss auf die Ergebnisse haben könnten. So ist es beispielsweise grundsätzlich möglich, dass persönliche Merkmale von Testpersonen, wie z.B. das Alter oder der Bildungsstand, Auswirkungen auf formale Aufgabenmerkmale, Itemschwierigkeit und -formulierung haben (Bühner, 2011). Dies sollte nicht nur bei der Instruktion sowie Itemformulierung, sondern auch der Analyse der Daten beachtet werden.

3.2.1.2 Alter

Obwohl Gewissenhaftigkeit zu den Traits, also den stabilen Persönlichkeitsmerkmalen zählt, korreliert die Persönlichkeitsdimension in der Normierungsstichprobe des NEO-FFI signifikant mit dem Lebensalter ($r = .17$; Borkenau & Ostendorf, 2008). Die Forscher formulierten entsprechend altersgerechte Normwerte, die die steigende Gewissenhaftigkeit im Alter berücksichtigen. Auch andere Forscher fanden kulturübergreifend Zusammenhänge zwischen Gewissenhaftigkeit und Lebensalter (Fraboni & Cooper, 1989; Srivastava, John, Gosling & Potter, 2003). Den Ergebnissen dieser Studien zufolge tendieren ältere Probanden zu höheren Ausprägungen der Gewissenhaftigkeitsdimension. Bezüglich der Substruktur von Gewissenhaftigkeit konnten sogar noch stärkere Schwankungen in den Merkmalsausprägungen der Gewissenhaftigkeitsfacetten bzw. -aspekte gefunden werden.

Die Fünf-Faktoren-Theorie assoziiert Persönlichkeitsveränderungen mit intrinsischer Reifung, sagt aber nicht, welche Veränderungen konkret beobachtet werden können. Es ist den Annahmen des Modells zufolge möglich, dass Eigenschaftsveränderungen auf evolutionsbiologische Prozesse zurückgeführt werden können. Beispielsweise ist es möglich, dass gewissenhaftere Menschen gesünderen Nachwuchs bekommen und aufziehen, was den Anstieg der Gewissenhaftigkeit im Erwachsenenalter erklären könnte. Der später einsetzende Rückgang der Gewissenhaftigkeit im hohen Alter wäre demzufolge auf eine weniger wichtige Funktion älterer Personen für die Gesellschaft sowie Funktionsstörungen im Gehirn zurückzuführen.

Terracciano, McCrae, Brant und Costa (2005) untersuchten im Rahmen ihrer längsschnittlichen Studie in einer bevölkerungsrepräsentativen Stichprobe unter anderem den Verlauf der Big 5-Persönlichkeitsausprägungen sowie deren Facetten bei den Studienteilnehmern. Sie fanden, dass die untersuchte Kohorte bzw. das Geburtsjahr ein signifikanter Prädiktor von Gewissenhaftigkeit war, in dem Sinne, dass später geborene Menschen höhere Gewissenhaftigkeitsausprägungen aufwiesen. Bezüglich sämtlicher Gewissenhaftigkeitsfacetten zeigten die Studienteilnehmer einen Anstieg ihrer individuellen Eigenschaftsausprägungen bis zu einem Alter von 60 bis 70 Jahren. Danach fielen die Werte wieder moderat ab. Bezüglich der Facette Deliberation zeigte sich der höchste Anstieg. Allgemein wiesen alle Probanden die höchsten Ausprägungen auf der Facette Competence auf.

Insgesamt blieben 85 % der Varianz der Persönlichkeitseigenschaften stabil. Die 15 % Varianz in den Daten führten die Autoren auf Schwankungen aufgrund von Messfehlern sowie normative bzw. nicht-normative gesellschaftliche Veränderungen zurück.

Allerdings ist es möglich, dass die beobachteten Veränderungen auf der Ebene des Traits durch die Veränderungen auf Ebene der Facetten maskiert werden. Aus diesem Grund über-

prüften Jackson und Kollegen (2009) in einer bevölkerungsrepräsentativen Stichprobe ($N_{Bev} = 613$) und in einer kleineren Feldstichprobe ($N_{Feld} = 274$) den Zusammenhang zwischen Alter und den Gewissenhaftigkeitsfacetten Industriousness, Orderliness sowie Impulse Control. Die Probanden der Feldstichprobe waren zwischen 18 und 94 Jahre alt und füllten Selbstbeschreibungsinventare zur Erhebung der interessierenden Persönlichkeitsmerkmale aus. Für die Industriousness-Facette zeigten sich die größten Unterschiede zwischen jungen und mittelalten Erwachsenen ($d = .42$, $p < .05$). Zwischen mittelalten und alten Erwachsenen ergab sich kein signifikanter Unterschied. Zudem konnte ein kurvilinearere Effekt beobachtet werden, demzufolge die Ausprägung der Facette Industriousness bei den Probanden bis zum mittleren Alter hin anstieg, dann konstant blieb und nach dem 80. Lebensjahr wieder abfiel. Bezüglich der Orderliness-Facette wurden keine signifikanten Unterschiede zwischen den verschiedenen Altersgruppen gefunden. Orderliness scheint somit nicht zum allgemeinen Anstieg der Gewissenhaftigkeitsausprägung beizutragen.

In der bevölkerungsrepräsentativen Stichprobe zeigten sich vergleichbare Ergebnisse. Altersunterschiede wurden vor allem zwischen jungen und mittelalten Erwachsenen gefunden. Die mittelalte Probandengruppe wies im Vergleich mit den jungen Studienteilnehmern signifikant höhere Werte in Aspekten der Pflichterfüllung, Entschlossenheit und der Organisation auf. Die Ordnungsfacette unterschied sich über die verschiedenen Gruppen hinweg nicht ($r = .06$, n.s.), während Fleiß leichte Zusammenhänge mit Alter zeigte ($r = .08$, $p < .05$).

Specht, Egloff und Schunkle (2011) berichteten von einem signifikanten Anstieg der Gewissenhaftigkeit bei Personen, die das erste Mal in ihrem Leben in eine geregelte Arbeitstätigkeit eingetreten sind. Sie erklären diesen Effekt mit dem Druck sozialer Normen, der bei den Berufsanfängern dazu führte, sich den allgemein geltenden Konventionen anzupassen, um beruflich erfolgreich zu sein. Die Autoren nehmen an, dass sich die Persönlichkeit einer Person vor allem dann verändert bzw. anders beschrieben wird, wenn diese am Übergang in einen neuen Lebensabschnitt stehe und starken Druck verspüre, sich auf eine bestimmte Art und Weise zu verhalten bzw. merke, dass die vorher gezeigten Verhaltensweisen und Reaktionen nicht erwünscht seien. Auch andere Studien berichten von größeren Gewissenhaftigkeitsveränderungen im jungen Erwachsenenalter (Roberts, Walton & Viechtbauer, 2006).

Allerdings ist der Zusammenhang zwischen Alter und Ausprägung der Gewissenhaftigkeitsfacetten trotz der o.g. Studien noch nicht hinreichend aufgeklärt. So könnten die bisher gefundenen Effekte auch auf individuelle Entwicklungsverläufe oder Kohorteneffekte zurückzuführen sein, da es sich bei den oben beschriebenen Studien lediglich um Querschnittsstudien handelt. Es ist möglich, dass aufgrund sich verändernder gesellschaftlicher Normen ältere Personen mehr Wert auf Gewissenhaftigkeit legen. So galt beispielsweise Pünktlichkeit vor einigen Jahrzehnten noch als „Tugend“, während diese heutzutage im Zuge der Einführung von Gleitzeit und mit Hilfe der modernen Kommunikationsmittel als immer weniger wichtig angesehen wird. Auch eine Tendenz zu sozial erwünschterem Antwortverhalten bei älteren Personen könnte eine Ursache für die gefundenen höheren Zusammenhänge von Gewissenhaftigkeit und Alter sein (Stöber, 2001).

Poropat (2009) konnte in seiner Metaanalyse moderierende Effekte von Alter bezüglich der Big 5-Dimensionen Verträglichkeit, Extraversion und Offenheit finden. Er berichtet jedoch keine Effekte für Neurotizismus und Gewissenhaftigkeit. Sutin et al. (2009) konnten ebenfalls keine moderierenden Effekte von Alter für den Zusammenhang von Persönlichkeit mit beruflichem Prestige und Arbeitszufriedenheit auffinden.

Da bisher keine eindeutigen Befunde über einen möglichen Zusammenhang der Gewissenhaftigkeitsaspekte Fleiß und Ordnung mit Lebensalter vorliegen, soll in der vorliegenden Studie eine mögliche Konfundierung der Ergebnisse mit dem Alter der Probanden berücksichtigt werden. Da es sich jedoch um sehr altershomogene Stichproben handelt, werden keine Alters-

effekte erwartet. Die Überprüfung der Retest-Reliabilität (s. Abschnitt 6.2.2.2) soll zudem Aufschluss über die Stabilität der Eigenschaften geben.

3.2.1.2 Geschlecht

Zur näheren Beschreibung der Stichproben wird das Geschlecht der Probanden mit erhoben. In Bezug auf Gewissenhaftigkeit konnten bisher allerdings keine oder nur sehr kleine signifikanten Unterschiede zwischen Männern und Frauen gefunden werden. So berichten Borkenau und Ostendorf (2008) nur sehr kleine Geschlechtsunterschiede für die Gewissenhaftigkeitsdimension des NEO-FFI ($t = 4,54$, $p < .001$, $d = .09$). Furnham et al. (2003) berichten Korrelationen zwischen Schulleistung und Geschlecht ($r = .26$, $p < .05$), in dem Sinne, dass Schülerinnen besser benotet wurden als Schüler. Allerdings konnten sie keine Unterschiede dahingehend finden, dass ihre weiblichen und männlichen Probanden unterschiedlich gewissenhaft waren. Auch Specht et al. (2011) fanden keine signifikanten Geschlechtsunterschiede bezüglich der Gewissenhaftigkeit bei Frauen und Männern. Sie berichten lediglich von Interaktionseffekten nach dem Tod des Ehepartners. Hier zeigte sich bei Frauen eine leichte Abnahme und bei Männern eine leichte Zunahme der Gewissenhaftigkeit. Insgesamt scheinen sich Männer und Frauen bezüglich ihrer Gewissenhaftigkeit nicht zu unterscheiden, was auch die Ergebnisse der bevölkerungsrepräsentativen Langzeitstudie von Terracciano et al. (2005) zeigten. Es werden somit keine Geschlechtereffekte in dieser Untersuchung erwartet.

3.2.1.3 Schulbildung

Wie in Abschnitt 2.7.2 bereits ausführlich dargestellt wurde, korreliert Gewissenhaftigkeit signifikant mit Schulbildung (Noftle & Robins, 2007; Poropat, 2011). Das heißt, je höher die Schulbildung, desto höher ist auch die Gewissenhaftigkeit einer Person. Dies kann bei der Vorhersage von Berufsschulleistung durch Gewissenhaftigkeit die Aussagekraft der Ergebnisse beeinträchtigen. So ist es im Nachhinein schwer nachzuvollziehen, inwiefern die Leistungen der Schüler auf Unterschiede in den Persönlichkeitseigenschaftsausprägungen, wie z.B. eine höhere oder geringere Gewissenhaftigkeit, oder lediglich auf die in der Schule vorher vermittelten Inhalte zurückzuführen sind. Auch ein sich gegenseitig verstärkender Einfluss der Faktoren ist denkbar in dem Sinne, dass höher gebildete Schüler nicht nur über mehr nützliches Vorwissen, sondern auch über höhere Gewissenhaftigkeitsausprägungen verfügen. Darüber hinaus könnten Gewissenhaftigkeit und Vorbildung auch miteinander interagieren oder von situationalen Umständen abhängig sein. Minbashian et al. (2010) fanden in einer Feldstudie, dass die individuellen Gewissenhaftigkeitsausprägungen von Managern aufgabenabhängig variierten. Es ist denkbar, dass ein vergleichbarer Effekt auch bei Auszubildenden eintritt. Es ist könnte im Rahmen der geplanten Untersuchung somit schwer werden nachzuvollziehen, wie groß der Anteil der schulischen Vorbildung in Bezug auf das Kriterium Ausbildungsleistung ist bzw. welchen Einfluss die Schulbildung der Schüler auf die Ausbildungsleistung hat. Da es möglich ist, dass die in dieser Studie erhobenen Daten nicht unabhängig von der Vorbildung der untersuchten Auszubildenden sind.

Görlich und Schuler (2007) sprechen sich dafür aus, gerade auch bei der Auswahl von Auszubildenden auch die vorherigen Schulleistungen zur Leistungsvorhersage mit heranzuziehen. Die schulische Vorbildung ist jedoch ein eher schlechter Leistungsprädiktor, da sich in den verschiedenen Schulen, Bundesländern und sogar Klassen innerhalb einer Schule die Benotungsregeln teilweise stark unterscheiden (Poropat, 2009). Das heißt konkret, dass die Schulnoten von Ausbildungsbewerbern nur bedingt vergleichbar sind und auf Basis der bisherigen Schulleistungen einer Person nur begrenzt Aussagen über die Leistungsfähigkeit einer Person getroffen werden können.

Zur Überprüfung der Vorhersagekraft vorheriger Schulnoten sowie des Zusammenhangs von Gewissenhaftigkeit mit der schulischen Vorbildung sollen im Rahmen dieser Untersuchung

daher Informationen zur besuchten Schulart sowie Informationen zur letzten durchschnittlichen Schulnote zur Kontrolle mit erhoben werden.

Da gerade bei der Auswahl von Auszubildenden in der Praxis vieler Personalabteilungen vor allem die Schulleistungen als Auswahlkriterium bzw. zur Erfolgs- und Leistungsprognose mit herangezogen werden, kann eine Berechnung der inkrementellen Validität von Eignungstests gegenüber den Informationen der Bewerbungszeugnisse zur Aufklärung der Nützlichkeit dieser Verfahren beitragen.

3.2.2 Berufliche Informationen

Wie auch die Erhebung der demografischen Merkmale der Studienteilnehmer soll die Erhebung beruflicher Informationen einerseits der angemessenen Beschreibung der Stichproben und andererseits der Bestimmung weiterer Einflussfaktoren dienen.

Es sollen Informationen zum Ausbildungsberuf bzw. der Ausbildungsrichtung der untersuchten Studienteilnehmer erhoben werden. Die Informationen zum Ausbildungsberuf sollen der präziseren Beschreibung der Bewerberstichprobe dienen. Die Gruppierung der Bewerber in *kaufmännische* und *technische* Auszubildende erfolgt gemäß des von ihnen ergriffenen Berufes.

Es ist möglich, dass sich die kaufmännischen und technischen Auszubildenden bezüglich ihrer Ausprägungen in den Gewissenhaftigkeitsaspekten Fleiß und Ordnung sowie der Selbstwirksamkeit signifikant voneinander unterscheiden. Dieses Vorgehen wurde auch von Dudley et al. (2006) vorgeschlagen, die in ihrer Metaanalyse fanden, dass man Facetten insbesondere dann verwenden sollte, wenn es darum geht, spezifisches Verhalten vorherzusagen. Da die kaufmännischen Bewerber im Durchschnitt über eine höhere Schulbildung verfügen und vergleichsweise älter sind als die technischen Auszubildenden, ist anzunehmen, dass die Kaufleute sich allgemein höhere Werte in den erfassten Konstrukten zuschreiben. Darüber hinaus ist es möglich, dass die beiden Berufsgruppen differenzierte Effekte hinsichtlich der prädiktiven Validität der Fleiß- und Ordnungsaspekte aufweisen. Da die Anforderungen an eine kaufmännische Ausbildung u.a. das Antizipieren von bspw. Geschäftsprozessen und die Übersicht über allgemeine Abläufe in der Wirtschaftswelt beinhaltet, sollten kaufmännische Auszubildende über höhere Ausprägungen des Fleißaspektes verfügen. Demgegenüber sollte der Ordnungsaspekt für technische Auszubildende eine höhere prognostische Validität besitzen, da hier vor allem das Verstehen und Anwenden von technischen Prozessen, die Einarbeitung in Routineverfahren oder sauberes, detailgetreues Arbeiten im Vordergrund der Ausbildung stehen.

Auch die vorherigen Berufserfahrungen der Studienteilnehmer sollen mit erhoben werden. Dies soll der Überprüfung dienen, inwiefern Personen mit Berufserfahrung berufsbezogene Testverfahren anders beantworten als Personen ohne berufliche Vorerfahrungen.

3.2.3 Sozial Erwünschtes Antwortverhalten

Sozial erwünschtes Antwortverhalten hat zur Folge, dass der Mittelwert einer Bewerbergruppe gegenüber einer neutralen Probandengruppe erhöht ist (Bühner, 2011). Metaanalysen zufolge liegt dieser Effekt zwischen einer halben bis zu einer ganzen Standardabweichung (Birkeland et al., 2006; Viswesvaran & Ones, 1999).

Der theoretische Hintergrund zu sozial erwünschtem Antwortverhalten wurde bereits in Abschnitt 2.4.3 ausführliche behandelt. Es ist bis dato umstritten, ob bzw. in welchem Ausmaß sozial erwünschtes Antwortverhalten die Ergebnisse und Vorhersagekraft von Persönlich-

keitstests, die im Rahmen der Praxis der Personalauswahl eingesetzt werden, beeinflusst (Risavy & Hausdorf, 2011; Ziegler et al., 2010).

Der Hauptkritikpunkt ist, dass sozial erwünschtes Antwortverhalten die Kriteriumsvalidität der verwendeten Persönlichkeitstests verzerren kann. So vermuteten z.B. Sutherland et al. (2007), dass sozial erwünschtes Antwortverhalten möglicherweise einen Einfluss auf den Zusammenhang von Gewissenhaftigkeit und Schulleistung hat. Sie berichten einen deutlichen Zusammenhang zwischen Gewissenhaftigkeit und sozialer Erwünschtheit ($r = .48$, $p < .01$). Insbesondere in der praktischen Vorhersage von Berufsschulleistung mit Hilfe differenzierter sowie kontextbezogener Gewissenhaftigkeitsmaße stehen diesbezüglich allerdings noch entsprechende Überprüfungen aus.

Ziegler et al. (2010) führten in einer Laborstudie eine Überprüfung der Auswirkungen sozial erwünschten Antwortverhaltens auf der Facettenebene der Big 5-Persönlichkeitsdimensionen durch. Die Forschergruppe konnte signifikante bzw. moderate Testwertverschiebungen auf den meisten Big 5-Facetten nachweisen. Allerdings zeigten diese Verschiebungen nicht in allen Facetten in die gleiche Richtung. Einige Facetten zeigten einen stärkeren, andere einen schwächeren Zusammenhang mit sozial erwünschten Antworttendenzen. Auf Ebene der Dimensionen konnten sie keinen signifikanten Einfluss von Faking nachweisen. Zwar beeinflusste das sozial erwünschte Antwortverhalten die Ausprägungen der Selbsteinschätzungen, aber dies hatte keinen Einfluss auf die Kriteriumsvalidität der Persönlichkeitsdimensionen. Auf der Ebene der Facetten waren diese Effekte weniger eindeutig. Ein weiteres Ergebnis dieser Studie war, dass Faking die individuellen Persönlichkeitsunterschiede zwischen den Studienteilnehmern verringerte. Daraus folgerten Ziegler und Kollegen (2010), dass die meisten Teilnehmer ihrer Studie eine ähnliche Vorstellung des „idealen Bewerberprofils“ haben mussten. Eine Überprüfung dieser Effekte in einer realen Bewerbergruppe, um aussagekräftige bzw. generalisierbare Ergebnisse zu erhalten, wurde bis dato jedoch noch nicht vorgenommen.

Die Ergebnisse der Studie von O'Brien und LaHuis (2011) deuten an, dass sozial erwünschtes Antwortverhalten vor allem dann auftritt, wenn allgemein gehaltene Testverfahren verwendet werden. So seien die Effekte von Faking eher auf Unterschiede in den von den Testpersonen in generischen Settings verwendeten Referenzrahmen zurückzuführen. Das Einsetzen kontextualisierter Testverfahren und -items sollte die Faking-Tendenzen ihren Ergebnissen zufolge eher verringern, da die Testpersonen bei der Beantwortung der kontextualisierten Items wüssten, auf welche Situationen sie sich beziehen müssten und nicht die „beste Option“ zur Itembeantwortung auswählen könnte. Doch auch diese Fragestellung wurde bisher noch wenig untersucht. Vor allem aussagekräftige Daten aus dem Feld fehlen zur Überprüfung dieser Annahmen.

Da die Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS) sowie die Selbstwirksamkeitsskala vom Antwort- und Itemformat her so aufgebaut sind wie die meisten verfälschbaren Testverfahren, besteht grundsätzlich die Gefahr, dass auch bei der Bearbeitung dieser Testverfahren die Ergebnisse in Richtung einer erwünschten Antworttendenz verfälscht werden können. Vor allem in Situationen, in denen Personen ein großes Interesse an einer möglichst positiven Selbstdarstellung haben, wie beispielsweise in Personalauswahl- oder -entwicklungsverfahren, ist die Gefahr einer selbstdienlichen Verzerrung in Richtung einer (subjektiv) wünschenswerten Ausprägung groß.

In verschiedenen Vorstudien wurde bereits der mögliche Einfluss sozial erwünschten Antwortverhaltens auf die AGS untersucht. Die Ergebnisse sind widersprüchlich. In einer Stichprobe amerikanischer College-Studenten ($N = 217$), die die AGS unter neutralen Bedingungen bearbeiteten, zeigten sich nach Herausparsialisierung der sozialen Erwünschtheit keine signifikanten Unterschiede in den Ergebnissen (Dreier, 2008). Hingegen zeigen die Ergebnisse

einer Untersuchung deutscher Arbeitnehmer ($N=345$), die die AGS ebenfalls unter neutralen Bedingungen ausfüllten, auf Zusammenhänge des Fleißaspekts ($r=.17, p < .01$) und des Ordnungsaspekts ($r=.28, p < .001$) mit sozial erwünschtem Antwortverhalten hin (Struckmeier, 2009). Allerdings waren die gefundenen Zusammenhänge im Vergleich mit einem generischen Fragebogen zur Erfassung der allgemeinen Gewissenhaftigkeitsaspekte deutlich geringer. Die Aspekte Industriousness ($r=.37, p < .001$) und Orderliness ($r=.25, p < .001$) des BFAS (DeYoung et al., 2007) zeigten deutlich stärkere Zusammenhänge mit sozial erwünschtem Antwortverhalten, was die o.g. Annahmen von O'Brien und LaHuis (2010) bestätigen. Weitere Studien zum Zusammenhang von Faking und den Gewissenhaftigkeitsaspekten Fleiß und Ordnung deuten ebenfalls auf Zusammenhänge zwischen den Konstrukten hin (Ziegeler, 2011; Ziegeler et al., 2012).

Aufgrund des noch unklaren Einflusses sozial erwünschten Antwortverhaltens auf die prädiktive Validität der AGS soll daher der Einfluss möglicher sozial erwünschter Antworttendenzen im Rahmen dieser Studie berücksichtigt werden.

Dies soll einerseits durch den parallelen Einsatz einer Skala zur Erfassung von Fakingtendenzen erfolgen. Da im Rahmen der Bewerbungstestung aus ökonomischen und organisationalen Gründen jedoch keine Skala zur Erfassung von Faking eingesetzt werden konnte, sollen zur Erfassung des Ausmaßes des sozial erwünschten Antwortverhaltens in der Bewerberstichprobe andererseits auch die Werte der Ausbildungsbewerber mit denen der Schülergruppe verglichen werden. Der Vergleich zweier Stichproben ist eine gebräuchliche Methode, um Faking zu untersuchen (Lee et al., 2010). Unterschiedliche Forschungsarbeiten haben zudem zeigen können, dass die Testwerte von Bewerbern signifikant höher ausfallen als die von Stelleninhabern (O'Brien & LaHuis, 2011), weshalb auch die Testwerte innerhalb der Bewerber- bzw. Ausbildungsgruppe miteinander verglichen werden sollen. Dies geschieht im Rahmen der Ermittlung der Retest-Reliabilitäten.

4. Entwicklung der AGS

Das folgende Kapitel beschreibt die Entwicklung der Arbeitsbezogenen Gewissenhaftigkeits-skalen (AGS). Hierzu wird im ersten Abschnitt (4.1) auf die Konstrukte Fleiß und Ordnung eingegangen, die der Skalenkonstruktion zugrunde lagen. In Abschnitt 4.2 wird beschrieben, welche verschiedenen Phasen bisher durchlaufen wurden, um die Items zu generieren, auszuwählen und zu überprüfen. Auch werden die Gründe zur Auswahl des verwendeten Antwortformates sowie der Instruktionen der Skala erläutert. Der nächste Abschnitt (4.3) beinhaltet Informationen über die Art und Ergebnisse der Studien, die bereits zur Weiterentwicklung und Überprüfung der Arbeitsbezogenen Gewissenhaftigkeitsskalen durchgeführt wurden. Abschnitt 4.4 enthält Informationen zur Reanalyse der Items. Der letzte Abschnitt (4.5) beinhaltet Beispiel-Items der AGS.

4.1 Definition der Konstrukte Ordnung und Fleiß

Vor der Entwicklung einer Skala steht immer eine theoriegeleitete Erarbeitung bzw. Definition der zu erfassenden Konstrukte. Eine ausführliche Beschreibung des Ursprungs der Gewissenhaftigkeitsaspekte *Fleiß* und *Ordnung* wurde bereits in Abschnitt 2.3.1 vorgenommen. An dieser Stelle soll nun die Definition der beiden Aspekte vorgenommen werden, die der Test- bzw. Itemkonstruktion zugrunde lag.

Fleiß soll als eher aktiver Aspekt effektives, genaues und zielgerichtetes Arbeitsverhalten sowie Konzentration bei der Arbeit und das Setzen von Prioritäten erfassen. Hohe Werte in diesem Gewissenhaftigkeitsaspekt deuten auf sehr strebsame Personen hin, die hart arbeiten, um ihre Ziele zu erreichen. „Fleißige“ Personen streben nach persönlicher Leistung. Niedrige Werte deuten demgegenüber an, dass eine Person sich nicht für besonders ehrgeizig ansieht und eher durch andere Dinge als dem Willen zur persönlichen Leistung angetrieben wird. In der eignungsdiagnostischen Praxis sollte bezüglich des Fleißaspekts erfragt werden, wie wichtig einer Person persönliche Leistungen sind und inwiefern diese durch Arbeitsziele bestimmt werden bzw. inwiefern eine Person Bestrebungen für den eigenen persönlichen Werdegang hat.

Ordnung beschreibt einen eher passiven Aspekt von Gewissenhaftigkeit. Mit dieser Skala soll vor allem organisiertes, sorgfältiges und planendes Arbeitsverhalten sowie eine Präferenz für Routinen erfasst werden. Niedrige Ausprägungen deuten an, dass solche Personen persönliche Organisiertheit nicht als ihre Stärke ansehen. Hohe Ausprägungen im Gewissenhaftigkeitsaspekt Ordnung beschreiben Personen, für die die persönliche Organisation leichter ist als für die meisten anderen Menschen. Solche Personen gehen grundsätzlich sehr planvoll und systematisch an die Dinge heran. Personen, die sehr ordentlich sind, könnten aber auch als rigide und unflexibel wahrgenommen werden. Sie tendieren dazu, sich überstark an Regeln zu halten und Verantwortlichkeiten zu sehr zu betonen. hohen Ausprägungen soll dieser Aspekt auch eine Neigung zur Pedanterie und Rigidität erkennen lassen können. Rigidität bezieht sich auf eine Tendenz, an alten Vorgehensweisen und Prinzipien festzuhalten, einen Mangel an intellektueller Flexibilität sowie Konservatismus (Van der Zee, Zaal & Piekstra, 2003). Renner (2011) vermutet entsprechend, dass „perfektionistische Selbstdarsteller“ auch in ihrem Beruf dazu tendieren könnten, Fehler und Schwächen zu verbergen oder sich nicht helfen zu lassen, was sich wiederum negativ auf die Produktivität auswirken könne. Ab einem bestimmten Punkt kann eine zu hohe Ausprägung des Ordnungsaspektes unter Umständen also nicht mehr nur hilfreich, sondern auch hinderlich sein. Personen, die Details zu viel Aufmerksamkeit schenken und sich ausschließlich mit diesen beschäftigen, verlieren den Überblick und somit die übergeordneten (Arbeits-) Ziele aus den Augen (Le et al., 2011). Den Autoren zufolge tendieren hoch gewissenhafte bzw. sehr „ordentliche“ Menschen dazu, sich selbst zu betrügen, indem sie zu rigide agieren, was wiederum die Lernfähigkeit und den Wissens-

erwerb der Personen einschränken und somit zu schlechterer Arbeitsleistung führen kann. Wenig ordentliche Personen schätzen sich demgegenüber eher unsystematisch arbeitend ein und empfinden es als schwierig, planvoll vorzugehen.

In der eignungsdiagnostischen Praxis sollten bei der Besprechung der Testwerte allem die Auswirkungen der persönlichen Organisiertheit auf die Effektivität bei der Arbeit in den Fokus gerückt werden. Es könnte auch erfragt werden, welche Strategien sehr ordentliche Personen für ihr Zeitmanagement und die persönliche Organisation verwenden und wie erfolgreich diese sind.

In Abschnitt 2.5 wurden bereits die Vorteile arbeitsbezogener Instrumente in der beruflichen Eignungsdiagnostik beschrieben. Aufgrund der genannten Vorteile einer berufsbezogen formulierten Skala wurde bei der Konstruktion der Skala darauf geachtet, alle Items arbeitsbezogen zu formulieren.

4.2 Testkonstruktion

Die folgenden Abschnitte beschreiben die Vorgehensweisen zur Skalenkonstruktion der AGS. Zuerst werden die Vorüberlegungen geschildert, die der Testkonstruktion (Abschnitt 4.2.1) und Itemauswahl (Abschnitt 4.2.2) vorausgingen. In Abschnitt 4.2.3 ist die Itemgenerierung beschrieben. Abschnitt 4.2.4 erläutert das Vorgehen der Itemselektion und enthält die Ergebnisse der ersten Pilotstudien zu den AGS. Das Antwortformat sowie die Gründe, die zur Auswahl des verwendeten Antwortformats führten, sind in Abschnitt 4.2.5 dargestellt. Der letzte Abschnitt (4.2.6) ist den Testinstruktionen gewidmet.

4.2.1 Allgemeine Vorüberlegungen zur Testkonstruktion

Ziel der Konstruktion der AGS war die Entwicklung eines Testverfahrens, das der Standarddefinition eines Tests entspricht: *„Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung“* (Lienert & Raatz, 1994, S. 1). „Wissenschaftliche Routine“ impliziert eine empirisch abgesicherte Theorie, die auf nachvollziehbaren Grundlagen basiert. Bedeutsam wird ein Messergebnis erst, wenn es dabei hilft, die Vorhersagen einer Theorie zu überprüfen.

Schimmack (2010) kritisierte, dass viele Persönlichkeitstests induktiv entwickelt wurden, d.h. dass diesen keine theoretischen Annahmen bezüglich der Faktorenstruktur zugrunde liegen und die Skalen oftmals durch einfache explorative Faktorenanalysen und -rotationen ermittelt wurden. Im Gegensatz zu diesen liegt den AGS eine eher deduktive bzw. rationale theoretische Basis zugrunde. Das heißt, zwischen der Beantwortung der Items und dem latenten Merkmal Fleiß bzw. Ordnung wird ein deterministischer Zusammenhang angenommen. Die Antwort auf ein Item lässt sich also auf den jeweiligen, a priori definierten Gewissenhaftigkeitsaspekt zurückführen. Da die Itemantworten eine quantitative Persönlichkeitsvariable erfassen, lässt sich der vermutete Zusammenhang zwischen Testantwort und psychischem Merkmal in Form einer Itemfunktion darstellen. Da es sich um einen Persönlichkeitstest handelt, wird von einem monoton steigenden Funktionsverlauf ausgegangen, d.h. mit zunehmender Eigenschaftsausprägung steigt die Wahrscheinlichkeit einer positiven Itemantwort (bei positiv gepolten Items).

Die Testkonstruktion bzw. Itemgenerierung wurde nach dem Modell der Klassischen Testtheorie (KTT) vorgenommen. Die KTT ist gegenwärtig die Konstruktionsgrundlage der meisten psychologischen Testverfahren (Bühner, 2011). Ein großer Vorteil der KTT ist ihre einfache Anwendbarkeit, auch wenn ihre Annahmen nur schwer überprüfbar sind. Dies kann allerdings durch eine inhaltlich begründete und sachgemäße Konstruktion von Items und

Skalen nahezu ausgeglichen werden. Eine Testkonstruktion nach den Prinzipien der Probabilistischen Testtheorie erschien aufgrund der durchzuführenden Konstruktionsprinzipien als zu aufwendig und wurde daher verworfen.

Zu den drei Hauptgütekriterien der KTT gehören die *Objektivität*, die *Reliabilität* sowie die *Validität* eines Testverfahrens. Darüber hinaus gibt es in der KTT noch einige Nebengütekriterien, wie z.B. die *Nützlichkeit*, *Fairness* oder *Ökonomie* eines Testverfahrens.

Sowohl die Reliabilität (Abschnitt 6.2.2) als auch die Validität (Abschnitt 6.3) der AGS sollen im Rahmen der vorliegenden Arbeit überprüft werden. Sie sind daher in Kapitel 6 erläutert. An dieser Stelle soll daher nur auf die Objektivität und einige Nebengütekriterien der AGS eingegangen werden, zu denen aufgrund der bereits vorliegenden Untersuchungen bereits Aussagen getätigt werden können.

Objektivität ist definiert als das Ausmaß, in dem das Untersuchungsergebnis von jeglichen Einflüssen außerhalb der untersuchten Person unabhängig ist (Rammstedt, 2004). Ein Messergebnis soll nur von dem gemessenen Merkmal abhängig sein. Weitere Einflüsse wie beispielsweise Situationsvariablen sollten keinen Einfluss auf das Ergebnis bzw. die Merkmalsausprägung haben. Die Objektivität als Maß, inwieweit eine Standardisierung des diagnostischen Prozesses gelingt, ist Grundvoraussetzung für die Messgenauigkeit (Reliabilität) und damit auch für die Gültigkeit (Validität) eines Verfahrens (Görlich & Schuler, 2007). Schon im Konstruktionsprozess wurde daher auf die Standardisierung sowohl der Verfahrensdurchführung als auch der Datenauswertung geachtet. Allgemein unterscheidet man drei Unterkategorien der Objektivität: *Durchführungs-*, *Auswertungs-* und *Interpretationsobjektivität*.

Durchführungsobjektivität kann man durch ein Konstanthalten der Untersuchungsbedingungen herstellen. Dies kann man durch eine Standardisierung der Bedingungen erreichen. Bei selbst auszufüllenden Fragebogen ist davon auszugehen, dass dies in der Regel gewährleistet ist (Rammstedt, 2004). So sind die Instruktionen zur Bearbeitung der AGS auf der ersten Seite des Fragebogens abgedruckt und auf diese Weise eine einheitliche Bearbeitung möglich.

Die *Auswertungsobjektivität* bezieht sich auf mögliche Fehler, die während der Auswertung des Fragebogens auftreten können. Das Ziel bei der Auswertung eines Fragebogens ist eine möglichst fehlerfreie und eindeutige Transformation der Antworten in Zahlenwerte (Rammstedt, 2004). Im Falle der AGS werden alle Antworten eindeutig in Form einer Auswertungsroutine festen Zahlenwerten zugeordnet, auf deren Basis sich die Skalenkennwerte für die beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung berechnen. Die Itemantworten müssen zusammengezählt und ein Mittelwert errechnet werden. Verschiedene Auswerter kommen mit Hilfe dieser Routine zu den gleichen Skalenkennwerten. Die Auswertung der AGS ist also aufgrund der eindeutigen Vorgaben zur Dateneingabe und -transformation als unabhängig von der auswertenden Person anzusehen.

Die *Interpretationsobjektivität* bezieht sich auf das Ausmaß, in dem die Schlüsse, die aufgrund der numerischen Ergebnisse gezogen wurden, über verschiedene Interpretatoren vergleichbar sind (Rammstedt, 2004). Personen mit gleichen latenten Ausprägungen sollten die gleichen numerischen Werte zugeschrieben werden. Um diese herzustellen, ist es wichtig, entsprechende Normwerte, z.B. in Form von Mittelwerten und Standardabweichungen sowie Konfidenzintervallen, zur Verfügung zu haben, um die jeweiligen Ergebnisse einordnen und vergleichen zu können. Darüber hinaus sollte das Konstrukt, das die jeweilige Skala erfasst, klar definiert sein, um der interpretierenden Person eine Hilfe an die Hand zu geben, wie die erreichten Ergebnisse inhaltlich einzuordnen sind. Zur eindeutigen Einordnung der Testwerte gibt es erste Studien, mit Hilfe derer die Testwerte der AGS verglichen werden können. Eine Normierung der AGS ist derzeit in Vorbereitung.

Aufgrund der vorliegenden Argumentation ist von der Objektivität der AGS auszugehen.

Bezüglich der Nebengütekriterien erfüllen die AGS ebenfalls die Anforderungen der KTT. So postulierte Bühner (2011), dass ein Test dann *ökonomisch* sei, wenn er a) kurz ist bzw. eine angemessene Durchführungszeit beansprucht, b) wenig Material verbraucht, c) einfach zu handhaben ist und d) als Gruppentest durchführbar sowie e) schnell und bequem auszuwerten ist. Weitere Kriterien der Ökonomie sind beispielsweise eine möglichst geringe Zahl an benötigten Testleitern pro Testung, eine Material- bzw. Ressourcen schonende Anwendung sowie Auswertung und eine einfache Interpretation. Insofern können die AGS als ökonomisches Verfahren eingeschätzt werden, denn sie können als Gruppentestung durchgeführt werden, der personelle Aufwand für die Testdurchführung, -auswertung sowie -interpretation ist gering, die Auswertung erfolgt über eine Schablone und lediglich die Testbögen selbst sind Verbrauchsmaterial. Alle diese Anforderungen erfüllen die AGS.

Zudem ist ein Test als *nützlich* anzusehen, wenn er ein Merkmal misst oder vorhersagt, für dessen Untersuchung ein praktisches Bedürfnis besteht. Der Bedarf eines arbeitsbezogenen und differenzierten Gewissenhaftigkeitsmaßes wurde in den vorhergehenden Abschnitten bereits ausführlich dargelegt.

Auch von der *Zumutbarkeit* der AGS ist auszugehen, da die Testpersonen nicht unnötig durch die Testung belastet werden. Verschiedene Untersuchungen belegen, dass die Wahrnehmung eines Auswahlverfahrens durch die Bewerber aufgrund der Länge und subjektiven Einfachheit eines Testverfahrens in dem Sinne beeinflusst wird, dass kürzere und einfachere Verfahren positiver beurteilt werden (Hausknecht, Day & Thomas, 2004; Ryan & Ployhart, 2000). Auch, um negative Emotionen bezüglich der AGS zu antizipieren, bestehen die AGS aus eher wenigen Items pro Skala, die ihrerseits wiederum kurz und prägnant formuliert sind.

Das Nebenkriterium der *Testfairness* erfasst, inwieweit einzelne Gruppen oder Personen in den erfassten Merkmalen unterscheidbar sind und durch den Test bevorzugt bzw. benachteiligt werden (Bühner, 2011). Das Thema der Testfairness ist allerdings sehr komplex und nur schwer zu überprüfen. Es soll daher an dieser Stelle nicht weiter behandelt werden. Erste Hinweise zur Testfairness liefert beispielsweise die Arbeit von Dreier (2008). In der von ihr untersuchten US-Stichprobe ($N=217$) zeigten die Werte der AGS keine Zusammenhänge mit der Kontrollvariable *Race* (Rasse).

4.2.2 Vorüberlegungen zur Itemauswahl

Um eine valide Skala zu entwickeln, galt es für die beiden Konstrukte Fleiß und Ordnung alle wesentlichen Aspekte in Form einzelner Indikatoren bzw. Items zu berücksichtigen. Eine Berücksichtigung aller möglichen Itemkombinationen für die beiden Gewissenhaftigkeitsaspekte erschien aus Gründen der Ökonomie des zu entwickelnden Instruments nicht angemessen und war auch gar nicht möglich. Hierfür können verschiedene Gründe angeführt werden:

1. Sowohl *Fleiß* als auch *Ordnung* sind zwei sehr komplexe Konstrukte. Aufgrund der Komplexität der beiden Gewissenhaftigkeitsaspekte erschien es nicht realisierbar, alle theoretisch möglichen Items zu Fleiß und Ordnung zu formulieren. Wichtig war es, die für die Fragestellung und den Arbeitskontext relevanten, allgemeingültigen Aspekte abzubilden.
2. Ein Bewerber oder Kandidat kann ein berufsbezogen formuliertes Fleiß- oder Ordnungs-Item nur dann bewerten, wenn er zu diesem überhaupt über einen entsprechenden Erfahrungsschatz verfügt. Falls jedoch die notwendige Bewertungsgrundlage für ein Item fehlt, ist eine Aussage nicht möglich und das Item in diesem Falle nicht valide. Daher wurden bei der Itemgenerierung, -selektion und späteren -revidierung ausschließlich solche Items berücksichtigt, die sich auf Grundlage allgemeiner und/ oder zentraler Arbeitsvorgänge oder Tätigkeitsinhalte bewerten lassen. Auf Items, die beispielsweise fachspezifische Tätigkeiten, wie z.B. Präsentationstechniken, abfragen, wurde daher gänzlich verzichtet.

3. Darüber hinaus sind viele Arbeitsaspekte sehr spezifisch und somit für den Kontext der Skalenentwicklung als eher unbedeutend einzustufen. Vor allem, wenn man bedenkt, dass diese spezifischen Arbeitsinhalte keine wesentlichen Auswirkungen auf die Gesamtbeurteilung von Fleiß oder Ordnung haben und dadurch auch die Beeinflussung weiterer Variablen, wie beispielsweise Berufs- oder Schulerfolg, unwahrscheinlich ist. Zwar könnten sich einzelne Arbeitsaspekte in Sonderfällen z.B. auf Seiten des Kandidaten oder Bewerbers als sehr wichtig wahrgenommen werden, gemittelt über die Gesamtheit der Bewerber oder Kandidaten besitzen diese jedoch nur eine geringe Relevanz und sollten daher nicht in die Skala mit aufgenommen werden.
4. Auch aus dem vorangegangenen Grund sollten nur Tätigkeiten oder Aspekte von Berufstätigkeiten mit in den Itempool aufgenommen werden, die sich auf alle Arten von Berufsgruppen und Hierarchieebenen in Unternehmen übertragen lassen, um eine breite Anwendbarkeit der zu entwickelnden Skala zu gewährleisten.

Die angeführten Überlegungen schränken die Art und Anzahl der Items, die im Prozess der Itemgenerierung und -selektion für eine Aufnahme in den Itempool in Frage kamen, deutlich ein. Das Ziel war daher, eine möglichst große Zahl von Items zu finden, die jeweils das Konstrukt Fleiß oder Ordnung hinreichend gut repräsentieren bzw. inhaltlich valide abbilden und dabei auf die verschiedensten Berufsgruppen, Tätigkeiten und Hierarchieebenen anwendbar sind.

Von einem *inhaltlich validen* Testverfahren spricht man, wenn ein Test bzw. jedes Item des Testverfahrens das zu messende Merkmal wirklich und hinreichend präzise erfassen (Bühner, 2011). Präzise bedeutet in diesem Kontext die Abbildung des Konstrukts durch das Item. Das Item muss präzise das relevante Konstrukt messen und darf keinen Überschneidungsbereich mit einem anderen Konstrukt haben. Die Gewährleistung einer hohen Inhaltsvalidität ist einer der wichtigsten Schritte in der Testkonstruktion. Es ist sehr schwierig, die Inhaltsvalidität eines Tests zu bestimmen. Diese ist empirisch nicht überprüfbar. Sie wird in der Regel nicht numerisch, sondern aufgrund logischer und fachlicher Überlegungen bestimmt und daher auch als *Augenscheinvalidität* bezeichnet. Üblicherweise nutzt man die Messwerte eines Tests in einer relevanten Stichprobe, um dessen Inhaltsvalidität bzw. *Kontentvalidität* mittels einer Kriteriums- und Konstruktvalidierung indirekt zu ermitteln. Laut Bühner (2011) bestimmt man mit der Kriteriums- und Konstruktvalidität jedoch nicht die Validität eines Tests im eigentlichen Sinne, sondern die Validität der abgeleiteten Aussagen, welche mithilfe der Testkennwerte getroffen werden (z.B. Verhaltensvorhersage). Daher wird auch oft von validitätsbezogenen Belegen gesprochen.

Wie in Abschnitt 2.5 zuvor beschrieben, weisen arbeitsbezogene Instrumente in der beruflichen Eignungsdiagnostik eine Vielzahl von Vorteilen auf. Daher sollte im vorliegenden Fall eine arbeitsbezogene Skala entwickelt werden. Aus diesem Grund wurde bei der Formulierung aller Items darauf geachtet, dass diese entweder durch die Addition „*bei der Arbeit*“ und/oder einen tätigkeitsbezogenen Inhalt des Items selbst einen Berufsbezug aufwiesen. Da andere Forscher diesbezüglich postuliert haben, dass ein einfaches Hinzufügen der Phrase „*bei der Arbeit*“ nicht genügt, um eine arbeitsbezogene Skala zu entwickeln und bei vielen Items auch keinen Sinn macht (Lievens et al., 2008; Schmit et al., 1995), sollte eine neue Skala mit kontextualisierten Items entwickelt werden.

Laut Schuler und Kollegen (2007) wird ein ökonomisch durchzuführendes und einfach zu verstehendes Verfahren als besonders praktikabel bewertet, was letztlich für die Verwendung eines solchen Verfahrens spricht. Bei der Konstruktion der Skala sowie der zugehörigen Instruktionen wurde daher vor allem auch darauf geachtet, diese möglichst einfach verständlich zu formulieren. Vor allem auch im Hinblick auf die zu testenden Personen wurde auf eine leichte Verständlichkeit der Items sowie der Instruktionen geachtet.

Die beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung sollten darüber hinaus nicht als hinreichende Facetten zur Bildung eines Gesamtmittelwertes „Gewissenhaftigkeit“ dienen, da sonst interpersonelle Varianz verloren gehen würde, die für die Vorhersage von z.B. Leistungsdaten benötigt wird (vgl. Abschnitte 2.8.2 und 3.1.2). Sowohl der Ordnungs- als auch der Fleißaspekt sind als voneinander getrennte, wenn auch interkorrelierte, Subdimensionen zu betrachten. Höhere Zusammenhänge zwischen den beiden Skalen sind auch aufgrund des Arbeitsbezugs der Skala jedoch möglich, da dieser als „Bias“ ebenfalls gemeinsame Varianz aufklären könnte. Bei der Itemgenerierung sollte daher darauf geachtet werden, dass die neu zu formulierenden Items jeweils nur einer der beiden Skalen zuzuordnen sind.

4.2.3 Itemgenerierung

Wie in Abschnitt 2.6 dargestellt, liegen bereits einige Fragebogenverfahren zur Erfassung von Gewissenhaftigkeit bzw. Fleiß und Ordnung vor, einige davon auch arbeitsbezogen formuliert. Da die angeführten Instrumente jedoch nicht den angeführten Ansprüchen an eine ökonomische, arbeitsbezogene und differenzierte Skala zur Erfassung der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung genügten, wurde die Neuentwicklung einer solchen Skala angeregt.

Fünf Experten aus den Bereichen der beruflichen Eignungsdiagnostik und differentiellen Psychologie, die sich darüber hinaus umfangreich mit den Ergebnissen der Studie von DeYoung und Kollegen (2007) beschäftigt hatten, generierten in Deutschland und den USA unabhängig voneinander Items. Es handelte sich um eine erfahrungsgeleitet-intuitive Top-Down-Technik, da sich die Experten an den in den Abschnitt 4.1 beschriebenen theoretischen Definitionen der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie den in Abschnitt 4.2.1 angeführten Vorüberlegungen zur Itemauswahl orientierten und in mehreren Workshops austauschten. Zur Generierung der Items wurden die beiden Konstrukte in ein nomologisches Netzwerk eingebettet und eine Sammlung von Definitionsmerkmalen zusammengestellt. Das Ziel war trotz der in Abschnitt 4.2.2 geschilderten Grenzen eine möglichst repräsentative und trennscharfe Erfassung der Konstrukte durch die Items. Items, die nicht mit den Formulierungen der Konstrukte stimmig waren, wurden nicht in den Itempool aufgenommen.

Da sich in den meisten Persönlichkeitsfragebogen die „Aussageform“ als Itemform durchgesetzt hat, wurden auch die Items der AGS in Aussageform generiert.

Die Itemgenerierung erfolgte sowohl in englischer als auch in deutscher Sprache. Die Items wurden unter besonderer Berücksichtigung der Sprache generiert, da Sprache das Medium von Fragebögen ist. Sowohl das Verständnis der Items als auch der geschriebenen bzw. vortragenen Instruktionen ist essentielle Vorbedingung für die korrekte Anwendung und Interpretation der Ergebnisse eines Instruments. Neben der formalen Struktur eines solchen Fragebogens ist es wichtig, dass jede Testperson unabhängig von ihrem Bildungsniveau den Inhalt und die Bedeutung eines Items erfassen und verstehen kann. Ist dies nicht der Fall, unterscheiden sich die von den Testautoren erdachte Bedeutung und Antwort der Befragten, was zur Invalidität des Instruments führt (DuVernet et al., 2010).

Es kann in diesem Falle sogar dazu kommen, dass sich die Antworten einer Testperson bei mehrfacher Testanwendung so unterscheiden, so dass diese Antworten gar nicht mehr miteinander vergleichbar sind und es zur Unreliabilität im Sinne einer geringen Retest-Reliabilität des Fragebogens kommen kann.

Um möglichst einfach verständliche Items zu verwenden, wurde bei der Itemgenerierung daher darauf geachtet, möglichst kurze und prägnante Formulierungen zu wählen. Unangemessene oder schwierige Wörter, die leicht missverstanden werden können, wurden nicht verwendet. Auch Items, die nur von bestimmten Altersgruppen verstanden werden können, ent-

weder aufgrund veralteter Ausdrücke oder aufgrund der Verwendung von Jugendsprache, wurden nicht in den Itempool aufgenommen. Weder (regionale) Sprichwörter noch Redewendungen oder Bildsprache wurden verwendet. Ebenso wurde die Verwendung der als problematisch erachteten „und-/ oder-Verbindungen“ bzw. die Verwendung von zwei oder mehr Konjunktionen vermieden, damit der Antwortende nicht unsicher ist, wenn er nur einem Teil der Aussage zustimmen kann oder möchte. Auch abstrakte Ausdrücke oder zweideutige Items mit mehreren Interpretationsmöglichkeiten wurden nicht in den Itempool mit aufgenommen.

Auf diese Weise wurde darüber hinaus versucht, die Items möglichst unabhängig und neutral von sprachlichen bzw. kulturellen Besonderheiten zu erzeugen. Denn durch die Unabhängigkeit von Sprache erhofft man sich bei objektiven Persönlichkeitstests, dass diese leichter auf kulturell-sprachliche und soziale bzw. sozioökonomische Zielgruppen in anderen Kultur- und/ oder Sprachkreisen übertragbar sind (Ortner et al., 2007). Bei der Generierung der Items wurde mit Blick auf die zunehmende Globalisierung der Arbeitswelt daher auch darauf geachtet, dass die Skala in Zukunft leicht übersetzbar sein würde. Dadurch wurde auch der Kritikpunkt aufgenommen, dass in Bezug auf Persönlichkeitsfragebögen teilweise erhebliche Übersetzungsproblematiken vorliegen würden.

Ein mögliches Problem bei der Bearbeitung der AGS – insbesondere im Kontext der beruflichen Eignungsdiagnostik – betrifft die Möglichkeit verzerrter Antworttendenzen. Tests sollten durch die getesteten Personen nicht willentlich oder unwillentlich in eine gewünschte Richtung verfälscht werden können (Bühner, 2011). Eine solche Forderung ist jedoch unrealistisch, da Personen immer in der Lage sind, Testergebnisse willentlich zu beeinflussen. Dies gilt v.a. für Fragebogenverfahren. In Abschnitt 2.4.3 wurde hierzu bereits eingehend das Thema der Möglichkeit sozial erwünschten Antwortverhaltens erläutert.

Als Gegenmaßnahmen zu Fakingtendenzen empfiehlt Kanning (2011), die Kandidaten bereits vor der Testung in der Instruktion darauf hinzuweisen, dass man nicht immer wissen könne, welche Antwort in einem Fragebogen „erwünscht“ sei und welche nicht. Daher bestehe die Möglichkeit, dass sich Kandidaten bei unwahrheitsgemäßen bzw. manipulierten Antworten schlimmstenfalls selbst schädigen könnten, wenn sie am Ende aufgrund ihrer falschen Angaben einen Job erhalten, der ihnen am Ende gar keine Freude bereitet, da er nicht den persönlichen Neigungen einer Person entspreche. Darüber hinaus soll man dem Autor zufolge auch auf den möglichen Einsatz von Kontrollskalen („Lügenskalen“) hinweisen und solche Skalen ggf. auch einsetzen. Bühner (2011) merkt hierzu allerdings an, dass Aufforderungen zu korrekter Testbearbeitung im Rahmen der Testinstruktion zu keinen besseren Ergebnissen führen als die Zusicherung von absoluter Anonymität der Ergebnisse. Auch Andeutungen oder Warnungen, dass das Verfälschen von Antworten erkannt werde und mit negativen Konsequenzen verbunden sei, ist dem Autor zufolge bei der Bekämpfung sozial erwünschter Antwortstile nicht effektiv.

Auch eine Eliminierung von Items, die besonders anfällig für sozial erwünschte Antworttendenzen sind, wurde von ihm als Gegenmaßnahme vorgeschlagen.

Vor allem aber empfiehlt Kanning (2011) eine Normierung der Testverfahren anhand von Bewerberstichproben, um darüber die Effekte der verzerrten Selbstdarstellung zu neutralisieren bzw. „mitzunormieren“.

Allerdings ist es auch nicht in allen Berufen sinnvoll, Personen mit einer Tendenz zur positiv verzerrten Selbstdarstellung auszusortieren. So kann sozial erwünschtes Antwortverhalten auch ein Indikator für z.B. soziale Kompetenz sein. Aus diesem Grund wurde beispielsweise auch im NEO-FFI (Borkenau & Ostendorf, 1993, 2008) bewusst auf die Integration einer Skala zur Erfassung sozialer Erwünschtheit verzichtet. Die Autoren begründen diesen Schritt damit, dass es bisher keine überzeugenden Belege gebe, die eine geringere Validität sozial erwünscht beantworteter Fragebögen im Vergleich mit „normal“ beantworteten Testverfahren

aufgezeigt haben. Der Einbezug einer Faking-Skala in normale Persönlichkeitsfragebogen birgt Borkenau und Ostendorf (1991) zufolge zwei Nachteile. Zum einen besteht die Gefahr, dass Testanwender die Ergebnisse von Probanden mit hohen Erwünschtheitswerten nicht interpretierten, was diagnostisch und ethisch problematisch sei. Zum anderen würde der Einbezug einer Faking-Skala einen Fragebogen unnatürlich aufblähen. Zudem könnten die existierenden Maße zur Erfassung von Faking entsprechende Probanden nicht eindeutig identifizieren. Entsprechend gibt es Forschungsergebnisse, die zeigen, dass diese Fragebögen substantielle Unterschiede in anderen Persönlichkeitsvariablen wie z.B. Neurotizismus, Verträglichkeit und Gewissenhaftigkeit abbilden. Es ist daher möglich, dass diese Fragebögen mehr messen als nur soziale Erwünschtheit (Bühner, 2011). Eine Korrelation dieser Maße mit einem Persönlichkeitsfragebogen sagt daher nichts darüber aus, ob eine Person die Werte verfälscht hat oder nicht. Pauls und Crost (2004) konnten zudem zeigen, dass auch diese Methode gegenüber der Verfälschung, die sie eigentlich aufdecken soll, anfällig ist.

Dem Problem der Akquieszenz lässt sich Borkenau und Ostendorf (1991) zufolge am besten durch eine ausgewogene Verschlüsselung der Items begegnen. Da negativ gepolte Items tendenziell jedoch schlechtere Itemkennwerte aufweisen, bringt auch diese Methode Nachteile mit sich. So entsteht bei negativ formulierten Items die Gefahr, dass die Beantwortung dieser Items von den Getesteten ein zu hohes Maß an kognitiver Beteiligung erfordert, da hierfür aufgrund der doppelten Verneinung eine gedankliche Umpolung des betreffenden Items notwendig wird. Dies kann wiederum zu falschen Antworten und somit einer erhöhten Messfehlerrate führen.

Aufgrund der uneindeutigen Befundlage zum Einfluss sozial erwünschten Antwortverhaltens sowie dessen Antizipation, wurde auf die Integration einer „Lügenskala“ in die AGS, vor allem auch aufgrund ökonomischer Überlegungen, verzichtet. Es wurden aber negativ formulierte Items generiert, deren Beantwortung keine zu starke kognitive Beanspruchung erforderte, um die Akquieszenz-Problematik etwas zu verringern. Hierzu wurde auf doppelte Verneinungen verzichtet und sich darum bemüht, inhaltlich gegensätzliche Formulierungen zu den Konstrukten zu formulieren. Auch dient die Integration der negativ gepolten Items der Sicherstellung, dass alle Personen die Items gewissenhaft gelesen und verstanden haben. Mit insgesamt nur vier negativ gepolten Items sollten die statistischen Probleme, die mit der Integration dieser Items einhergehen, bei den AGS eher gering ausfallen.

4.2.4 Itemauswahl

Basierend auf den unabhängigen Itemformulierungen der fünf Experten wurde im Rahmen einer Expertensitzung über die einzelnen Items beraten. Aus dem Itempool wurden anhand verschiedener Auswahlkriterien wie beispielsweise *Prägnanz*, *Einfachheit* oder *interkultureller Unabhängigkeit* 72 Items ausgewählt, die im Rahmen einer Pilotstudie hinsichtlich verschiedener statistischer Kennwerte auf ihre Güte und faktorenanalytische Skaleneigenschaften überprüft wurden. Hierzu zählten die interne Konsistenz, Itemschwierigkeit, Trennschärfen sowie Faktorladungen der Items.

Die Items wurden im deutschsprachigen Raum an Arbeitnehmern verschiedener Berufsgruppen ($N=50$) überprüft (Struckmeier, 2009). Im englischsprachigen Raum wurden die Items Doktoranden einer amerikanischen Universität ($N=59$) vorgelegt, die diese bewerteten (Dreier, 2008). Basierend auf den Ergebnissen der zwei Pilotstudien wurden für die Fleißskala zwölf Items ausgewählt, die diesen Aspekt reliabel abbildeten ($\alpha = .90$ bzw. $.86$) und insgesamt 42 % der Varianz aufklärten. Für den Ordnungsaspekt wurden elf Items identifiziert, die zusammen 31 % der Varianz aufklärten und eine gute interne Konsistenz aufwiesen ($\alpha = .82$ bzw. $.84$). Unter den ausgewählten Items waren insgesamt nur vier negativ gepolte Aussagen. Die negativ formulierten Items verfügten in der Regel über geringere Zusammenhänge mit

der jeweiligen Skala und wurden daher in den meisten Fällen aufgrund ihrer schlechten Itemkennwerte nicht in den Fragebogen übernommen. Alle vier negativ formulierten Items waren der Skala Ordnung zugeordnet. Insgesamt bildeten somit 23 Items die Testversion der AGS.

4.2.5 Antwortformat

Das Antwortformat betreffend galt es zunächst, die Anzahl der Antwortstufen (4.2.4.1) und danach die Benennung der Pole bzw. Stufen (4.2.4.2) festzulegen.

4.2.5.1 Anzahl der Stufen

Bei Items in Aussageform liegt das Antwortformat in der Regel in Form einer gebundenen, mehrstufigen Rating- bzw. Likertskala vor, auf der die Probanden ihre Antworten markieren können. Diese ermöglichen eine quantitative Beurteilung der Eigenschaftsausprägungen einer Person. Der Differenzierungsgrad hängt davon ab, wie genau oder differenziert die Probanden das jeweilige Item beantworten können bzw. sollen. Je differenzierter das Antwortformat, desto mehr Möglichkeiten ergeben sich, zwischen Personen zu unterscheiden. Ebenso steigen die Validität und Reliabilität eines Verfahrens, je mehr Antwortkategorien verwendet werden (Bühner, 2011). Ein Antwortformat mit ungerader Anzahl an Antwortkategorien birgt die Möglichkeit einer „neutralen“ Antwortkategorie, was für die meisten Probanden einfacher zu handhaben ist. Unter Gesichtspunkten der Reliabilität und Validität ist eine fünf- oder siebenstufige Likertskala als Antwortformat am besten geeignet. Bei mehr Antwortmöglichkeiten steigt zwar die Reliabilität, aber die Modellpassung nimmt ab, v.a. bei längeren Tests. Zu viele Antwortmöglichkeiten können sich zudem negativ auf die Itemeigenschaften auswirken, da Testpersonen mit zu vielen Antwortmöglichkeiten auch überfordert sein könnten. Die Entscheidung, wie viele Antwortstufen zur Verfügung gestellt werden, sollte sich somit vor allem nach der Differenzierungsfähigkeit der getesteten Personen richten. Werden zu wenige oder zu viele Stufen verwendet, können Messfehler entstehen. Bei zu wenigen Stufen besteht die Gefahr, dass der wahre Wert nicht in den Stufen enthalten ist, sondern zwischen diesen liegt. Im Falle einer zu differenzierten Skala besteht darüber hinaus die Gefahr, dass ein zu hohes Maß an Genauigkeit vorgetäuscht wird. Ein weiteres Problem mehrstufiger Antwortformate stellen sogenannte „Antwortstile“ dar. So tendieren manche Personen dazu, lediglich mittlere Werte oder ausschließlich Extremwerte anzukreuzen, was die Aussagekraft der Ergebnisse des Tests verringert.

In den empirischen oder Sozialwissenschaften kommen vor allem fünf- oder siebenfach gestufte Ratingskalen zum Einsatz. Man erhält ausreichend differenzierte Antworten über die individuellen Merkmalsausprägungen einer Person und die Auswertung dieser Antwortformate ist zudem ökonomisch (Bühner, 2011).

Da die Items der AGS eher einfach verständlich und wenig komplex formuliert sind, erschien ein fünfstufiges Antwortformat mit bipolarer Antwortskala für die AGS angemessen. Auch aus praktischen Gesichtspunkten sprachen einige Gründe für die Verwendung einer fünfstufigen Likert-Skala. Da viele Persönlichkeitsfragebögen ebenfalls ein fünfstufiges Antwortformat verwenden (z.B. NEO-FFI; Ostendorf & Borkenau, 1993, 2008), wird durch die Verwendung desselben Antwortformats eine Vergleichbarkeit der Testwerte hergestellt. So können beispielsweise die Ausprägungen der Aspekte Fleiß und Ordnung direkt mit dem NEO-FFI-Wert der Gewissenhaftigkeit verglichen werden, ohne die Ergebnisse vorher aufwendigen Transformationsprozessen unterwerfen zu müssen. Auch können die AGS auf diese Weise leicht zur Ergänzung weiterer Testverfahren eingesetzt werden. Insbesondere im Hinblick auf die angestrebte Nutzergruppe der AGS, zu der auch Nicht-Psychologen zählen könnten, ist dies ein nicht zu vernachlässigender Punkt die Interpretation der Ergebnisse betreffend. Aber auch die Testkandidaten profitieren von einheitlichen Antwortformaten, da diese sich dann

nicht immer wieder umgewöhnen müssen. Verständnisprobleme und somit Fehlerwahrscheinlichkeiten bei der Beantwortung der Items werden auf diese Weise verringert.

4.2.5.2 Benennung der Pole/ Stufen

Bei Aussageitems wird die Antwort meist als Maß der Zustimmung bzw. Ablehnung formuliert. Dies kann als eine Art Standard angesehen werden. Aus diesem sowie den oben angeführten Gründen der Vergleichbarkeit und Praktikabilität wurde auch die Benennung der Pole bzw. Antwortstufen der AGS-Items an die Nomenklatur des NEO-FFI (Borkenau & Ostendorf, 1993, 2008) angeglichen. Die Ausprägungen wurden daher mit „0= *starke Ablehnung (SA)*“, „1= *Ablehnung (A)*“, „2= *neutral (N)*“, „3= *Zustimmung (Z)*“ und „4= *starke Zustimmung (SZ)*“ benannt.

4.2.6 Instruktionen

Besonderes Augenmerk wurde auch auf die Formulierung der Instruktionen der AGS gelegt, um im Vorwege der Testung alle (möglichen) Fragen der Testpersonen zu antizipieren bzw. beantworten und Fehler bei der Beantwortung zu vermeiden. Fehlende Informationen in den Instruktionen können zu einer geringen Transparenz eines Testverfahrens führen. Dies kann seitens der Testpersonen zu Überlegungen und falschen Schlüssen führen, warum der betreffende Test zu bearbeiten sei und was eigentlich gemessen werden solle (Ortner et al., 2007). Eine Darlegung des Zwecks bzw. der Ziele eines Testverfahrens soll demzufolge zu weniger gedanklicher Beschäftigung mit den Items selbst führen und eine unverfälschtere, d.h. „wah-rere“ Beantwortung der Items nach sich ziehen.

Auch ist es möglich, dass es im Falle fehlender oder nur unzureichend vorhandener Instruktionen zu Beginn eines Tests bei den Getesteten zu unterschiedlichen Annahmen über die abgefragten Ansichten oder Verhaltensweisen kommen kann. Die Getesteten entwickeln demzufolge unterschiedliche Bezugsrahmen, die zur Beantwortung der Items herangezogen werden, was letztlich zu verringerten Reliabilitäten und Validitäten führen kann. Die möglichen Auswirkungen unterschiedlicher Referenzrahmen wurden in Abschnitt 2.5 bereits ausführlich erläutert. Da das Ziel des Fragebogens die Abbildung der latenten Konstrukte Fleiß und Ordnung für den Schul- bzw. Arbeitskontext war, wurde nicht nur durch die arbeitsbezogene Formulierung der Items auf die Aktivierung dieses Kontextes geachtet, sondern auch schon bei der Formulierung der Instruktionen. Alle Testpersonen sollten nach der Testinstruktion und vor der Bearbeitung der Items dieselben Erwartungen an das Testverfahren haben.

Allgemein orientierte sich die Formulierung der Instruktionen des NEO-FFI (Borkenau & Ostendorf, 1993, 2008), da diese klar, prägnant und kurz sind und somit stark den Ansprüchen an eine ökonomische und leicht verständliche Skala gerecht werden. In den Instruktionen wurde darüber hinaus der Arbeitsbezug der Skala deutlich gemacht. Die Testpersonen werden darauf hingewiesen, dass sie die Items der AGS beantworten sollen, indem sie sich in ihren persönlichen Schul- oder Arbeitsalltag hineinversetzen.

4.3 Voruntersuchungen

Die folgenden Abschnitte stellen die Forschungsarbeiten vor, die bisher zur Überprüfung und Weiterentwicklung der AGS durchgeführt wurden. Es ist wichtig, im Rahmen der Entwicklung einer Skala verschiedene Vortests mit Personen aus relevanten Zielgruppen durchzuführen (Bühner, 2011). Diese – möglichst realistischen – Vortests dienen der Überprüfung und Antizipation von z.B. technischen Problemen (z.B. unausgefüllte Seiten bei mehrseitigem

Druck), Verständlichkeitsproblemen (z.B. Verwendung von Fremdwörtern), Akzeptanz (z.B. durch Rücklaufquoten) oder Antworttendenzen (z.B. Deckeneffekte).

Abschnitt 4.3.1 beschreibt die Ergebnisse der Skalenüberprüfung der 23-Item-Vorversionen der AGS in Deutschland und den USA. Abschnitt 4.3.2 gibt Auskunft über bisher gefundene Zusammenhänge der AGS mit anderen Persönlichkeitseigenschaften bzw. psychologischen Konstrukten sowie schulischen und beruflichen Leistungskriterien.

4.3.1 Skalenüberprüfung

Die Skalenüberprüfung der Vorversion der AGS wurde parallel in Deutschland für die deutschsprachige Testversion und den USA für die englischsprachige Testversion vorgenommen. Die Studien bzw. deren Ergebnisse sind in den folgenden Abschnitten 4.3.1.1 und 4.3.1.2 dargestellt.

4.3.1.1 Überprüfung der deutschsprachigen AGS-Vorversion

Die deutschsprachige Vorversion der AGS wurde im Rahmen einer Personalentwicklungsmaßnahme von Arbeitnehmern verschiedener Unternehmen ($N = 340$) bearbeitet, die über mindestens ein Jahr Berufserfahrung verfügten (Struckmeier, 2009). Neben der AGS wurden weitere Fragebogenverfahren eingesetzt, um mögliche Zusammenhänge mit weiteren psychologischen Kriterien bzw. Konstrukten zu erforschen (s. Abschnitt 4.3.2). So wurden beispielsweise u.a. auch die Gewissenhaftigkeitsitems der BFAS (DeYoung et al., 2007) mit abgefragt, um die Konstruktvalidität der AGS zu überprüfen.

In der Studie konnte mittels konfirmatorischer Faktorenanalysen die Zweifaktorenstruktur sowohl in den AGS als auch im BFAS nachgewiesen bzw. repliziert werden. Die Skalen wiesen darüber hinaus befriedigende bis gute interne Konsistenzen auf. Für den Fleißaspekt konnten interne Konsistenzen von $\alpha = .73$ (BFAS) bis $\alpha = .84$ (AGS) ermittelt werden. Der Ordnungsaspekt wies mit $\alpha = .77$ (BFAS) bzw. $\alpha = .79$ (AGS) etwas niedrigere interne Konsistenzen auf. Tabelle 4-1 gibt einen Überblick über die Reliabilitäten sowie Skaleneigenschaften und Interkorrelationen der vier Gewissenhaftigkeitsaspekte.

Tabelle 4-1: Ergebnisse der Skalenüberprüfung der AGS-Vorversion in Deutschland (nach Struckmeier, 2009)

Variable	M^a	SD	1	2	3	4
AGS-Fleiß	3,91	.50	(.84)			
AGS-Ordnung	3,85	.56	.33***	(.79)		
BFAS-Fleiß	3,76	.49	.41***	.41***	(.73)	
BFAS-Ordnung	3,67	.57	.16**	.67***	.47***	(.77)

Anmerkungen: M = Mittelwert; SD = Standardabweichung; ** $p < .01$; *** $p < .001$ (zweiseitig). ^a Das Skalenniveau war mit „1= starke Ablehnung (SA)“, „2= Ablehnung (A)“, „3= neutral (N)“, „4= Zustimmung (Z)“ und „5= starke Zustimmung (SZ)“ definiert.

Insgesamt verfügten die AGS über etwas höhere Reliabilitätskennwerte als die Gewissenhaftigkeitsaspekte der BFAS. Die Interkorrelation der AGS war mit $r = .33$ ($p < .001$) vergleichsweise niedrig. Der mäßige Zusammenhang sprach jedoch dafür, dass die die AGS zwei differenzierte Aspekte von Gewissenhaftigkeit erfassen. Die Interkorrelation der Gewissenhaftigkeitsaspekte der BFAS betrugen $r = .47$ ($p < .001$) und waren somit etwas höher. Die Fleißaspekte korrelierten untereinander mit $r = .41$ ($p < .001$), während die Ordnungsaspekte untereinander mit $r = .67$ ($p < .001$) den höchsten Zusammenhang aufwiesen, was für eine hohe Konstruktähnlichkeit sprach.

4.3.1.2 Überprüfung der englischsprachigen AGS-Vorversion

In einer Untersuchung an amerikanischen Collegestudenten ($N=217$) wurde die englischsprachige Vorversion der AGS überprüft (Dreier, 2008). Auch in dieser Untersuchung wurden neben den AGS u.a. die Gewissenhaftigkeitsitems der BFAS (DeYoung et al., 2007) mit erhoben, um die Übereinstimmung der Skalen zu ermitteln.

Die Überprüfung des theoretischen Modells mittels konfirmatorischer Faktorenanalysen ergab auch in dieser Stichprobe eine Überlegenheit des zweifaktoriellen Modells gegenüber einer einfaktoriellen Lösung. Für die Skala Fleiß ergaben sich sowohl für die Items der AGS ($\alpha = .84$) als auch die der BFAS ($\alpha = .81$) gute Reliabilitätskennwerte. Die internen Konsistenzen der Ordnungsskala waren für die Items der AGS-Skala befriedigend ($\alpha = .78$), bei denen der BFAS-Skala gut ($\alpha = .82$). Tabelle 4-2 gibt einen Überblick über die Reliabilitäten und Interkorrelationen der verschiedenen Gewissenhaftigkeitsaspekte.

Tabelle 4-2: Ergebnisse der Skalenüberprüfung der englischen Vorversion der AGS (nach Dreier, 2008)

Variable	<i>M</i>	<i>SD</i>	1	2	3	4
AGS-Fleiß	2,86	.46	(.83)			
AGS-Ordnung	2,63	.55	.55**	(.78)		
BFAS-Fleiß	2,32	.57	.45**	.49**	(.81)	
BFAS-Ordnung	2,42	.64	.30**	.70**	.49**	(.82)

Anmerkungen: *M*= Mittelwert; *SD*= Standardabweichung; ** $p < .01$; * $p < .05$ (zweiseitig).

Die Interkorrelationen der Skalen ergaben einen guten Zusammenhang für die Skalen der AGS ($r = .55$, $p < .01$) und eine mittlere Interkorrelation der beiden Gewissenhaftigkeitsaspekte der BFAS ($r = .49$, $p < .01$). Die Fleißaspekte korrelierten mit $r = .45$ ($p < .01$) ebenfalls mäßig miteinander, wohingegen die Ordnungsaspekte mit $r = .70$ ($p < .01$) skalenübergreifend einen hohen Zusammenhang aufwiesen. Dieser Trend spricht für eine hohe inhaltliche Übereinstimmung der Ordnungsskalen. Mit Zusammenhängen von $r = .16$ bis $r = .41$ ($p < .01$) wiesen die heterogene Aspekte skalenübergreifend erwartungsgemäß die geringsten Zusammenhänge auf.

4.3.2 Zusammenhänge mit weiteren psychologischen Konstrukten

Neben den Studien zur Überprüfung der Skaleneigenschaften wurden die AGS auch auf Zusammenhänge mit anderen Variablen hin überprüft, welche im Folgenden dargestellt werden. Neben dem Zusammenhang mit Persönlichkeitseigenschaften wie sozialer Erwünschtheit (Abschnitt 4.3.2.1) oder Selbstwirksamkeit und wahrgenommenen eigenen Fähigkeiten (Abschnitt 4.3.2.2) liegen auch bereits erste Ergebnisse zu leistungsnahen Konstrukten wie Karriereinteressen (Abschnitt 4.3.2.3), Copingstrategien (Abschnitt 4.3.2.4) und Selbstführungsstrategien (Abschnitt 4.3.2.5) sowie schulischen (Abschnitt 4.3.2.6) und beruflichen Leistungskriterien (Abschnitt 4.3.2.7) in Kombination mit den AGS vor.

4.3.2.1 Soziale Erwünschtheit

Die Möglichkeit sozial erwünschten Antwortverhaltens als Einflussvariable bei der Beantwortung von Persönlichkeitsinventaren, insbesondere im eignungsdiagnostischen Kontext, wurde bereits in den Abschnitten 2.4.3 sowie 3.2.3 ausführlich dargestellt. Bisherige Forschungsarbeiten gehen von keinen bzw. minimalen Einflüssen sozialer Erwünschtheit auf die Ergebnisse bzw. Zusammenhänge der AGS aus.

So untersuchte Dreier (2008) im Rahmen ihrer Befragung von Collegestudenten ($N=217$) auch den möglicherweise konfundierenden Einfluss des Konzeptes der sozialen Erwünschtheit mit den Testwerten der AGS sowie die Zusammenhänge mit anderen Persönlichkeits- und Leistungsvariablen. Zur Erfassung der Sozialen Erwünschtheit wurde die CMSDS

(Crowne & Marlowe, 1960) eingesetzt. Es ergaben sich mäßige Korrelationen des Konzeptes mit Fleiß ($r = .16$, $p < .05$) sowie Ordnung ($r = .16$, $p < .05$). Ein Vergleich der untersuchten Zusammenhänge zwischen unabhängigen und abhängigen Variablen mit und ohne Auspartialisierung der sozialen Erwünschtheit ergab keine signifikanten Unterschiede in den Ergebnissen. Die Autorin folgerte daher, dass das Konzept keinen bedeutsamen Einfluss auf die AGS und die Zusammenhänge der Ordnungs- und Fleißskalen mit anderen psychologischen Persönlichkeits- und Leistungsvariablen hat.

Struckmeier (2009) fand in ihrer Befragung von Mitarbeitern verschiedener norddeutscher Unternehmen ($N = 340$), dass sozial erwünschtes Antwortverhalten die Ergebnisse ihrer Studie bzw. die Zusammenhänge zwischen den unabhängigen und abhängigen Variablen kaum beeinflussten, obwohl soziale Erwünschtheit sowohl mit der Fleiß- ($r = .17$, $p < .01$) als auch der Ordnungsskala ($r = .28$, $p < .001$) korrelierte. Die Autorin empfahl, in Fragebogenverfahren nicht zu sehr die möglichen Effekte sozialer Erwünschtheit zu kontrollieren, da auf diese Weise auch Anteile der wahren Varianz entfernt werden könnten, was wiederum die gefundenen Zusammenhänge künstlich verringern und somit verfälschen könnte.

Auch Praxisuntersuchungen zum Einfluss sozial erwünschter Antworttendenzen bezüglich der AGS existieren bereits (Ziegeler, 2011; Ziegeler et al., 2012). Die Untersuchung beinhaltete u.a. die Analyse der Daten einer Feldstichprobe aus Bewerbern um einen Ausbildungsplatz ($N = 65$) bei einem großen deutschen Hausgerätehersteller, die längsschnittlich untersucht wurden. Während des Auswahlverfahrens im Rahmen eines Testtages wurden u.a. die AGS eingesetzt. Zu einem zweiten Testzeitpunkt ein Jahr später wurde zusätzlich die Soziale-Erwünschtheits-Skala-17 (SES-17; Stöber, 1999) eingesetzt. Die Daten wurden mit einer neutralen Kontrollgruppe gleichaltrigen Schulabsolventen ($N = 97$) verglichen. Die Ergebnisse bestätigen Unterschiede in den Stichproben bezüglich des Ausmaßes des Fakings in Persönlichkeitsinventaren dahingehend, dass die Bewerber sich signifikant höhere Eigenschaftsausprägungen zuschrieben als die Probanden der Kontrollgruppe. Darüber hinaus konnte aufgezeigt werden, dass neben den Persönlichkeitseigenschaften auch die Schulart der Befragten das Ausmaß sozial erwünschten Antwortverhaltens beeinflusste.

4.3.2.2 Selbstwirksamkeit und wahrgenommene eigene Fähigkeiten

Die Konstrukte Selbstwirksamkeit sowie wahrgenommene eigene Fähigkeiten wurden bereits in den Abschnitten 2.3.2 und 2.3.4 ausführlich erläutert.

Der Zusammenhang von Fleiß und Ordnung mit den beiden Selbsteinschätzungsstrukturen wurde mit der Vorversion der AGS sowie den beiden Gewissenhaftigkeitsaspekten der BFAS schon einmal in einer Stichprobe amerikanischer Collegestudenten ($N = 217$) untersucht und verglichen (Dreier, 2008). Hierzu wurden neben der Vorversion der englischsprachigen AGS auch die Gewissenhaftigkeitsaspekte der BFAS (DeYoung et al., 2007), allgemeine Selbstwirksamkeit (GSE= General Self-Efficacy; Schwarzer & Jerusalem, 1995) sowie acht Items zur Erfassung wahrgenommener eigener Fähigkeiten (PA= Perceived Abilities, Nofle & Robins, 2007) erhoben und analysiert. Tabelle 4-3 gibt eine Übersicht über die Zusammenhänge der aufgeführten Variablen.

Tabelle 4-3: Zusammenhänge der AGS mit Selbstwirksamkeit und wahrgenommenen eigenen Fähigkeiten (nach Dreier, 2008)

Variable	r AGS-Fleiß	r AGS-Ordnung	r BFAS-Fleiß	r BFAS-Ordnung
GSE	.44**	.22**	.48**	.16*
PA	.35**	.25**	.40**	.25**

Anmerkungen: r = Korrelation; GSE= General Self-Efficacy/ Allgemeine Selbstwirksamkeit; PA= Perceived Abilities/ wahrgenommene eigene Fähigkeiten; * $p < .05$; ** $p < .01$ (zweiseitig).

Insgesamt zeigten sich signifikante Zusammenhänge für alle Variablenkombinationen. Fleiß und Selbstwirksamkeit korrelierten mit $r_{AGS} = .44$ ($p < .01$) bzw. $r_{BFAS} = .48$ ($p < .01$) vergleichsweise hoch miteinander. Die Ordnungsaspekte korrelierten hingegen nur mäßig mit Selbstwirksamkeit ($r_{AGS} = .22$, $p < .01$ bzw. $r_{BFAS} = .16$, $p < .05$). Bezüglich der wahrgenommenen eigenen Fähigkeiten zeigte sich eine ähnliche Tendenz. Auch hier zeigten die Fleißaspekte höhere Zusammenhänge mit dem Konstrukt ($r_{AGS} = .35$, $p < .01$ bzw. $r_{BFAS} = .40$, $p < .01$) als die Ordnungsskalen ($r = .25$, $p < .01$). Inhaltlich sprachen die hohen Zusammenhänge zwischen den Skalen für eine inhaltlich-konzeptuelle Verwandtschaft der Konstrukte. Eine Überprüfung der Skalenstrukturen mittels konfirmatorischer Faktorenanalysen sprach jedoch für die Eigenständigkeit der verschiedenen Faktoren (Dreier, 2008).

Neben den Zusammenhängen der Variablen wurde auch ein möglicher mediiender Effekt der beiden Selbsteinschätzungskonstrukte für die Vorhersage von Schulleistung und Karriereinteressen (s. Abschnitt 4.3.2.3) durch die Gewissenhaftigkeitsaspekte Fleiß und Ordnung überprüft. Die Ergebnisse der Mediationsanalyse zeigten einen deutlichen mediienden Effekt sowohl von Selbstwirksamkeit als auch wahrgenommener eigener Fähigkeiten bei der Vorhersage von Schulleistung durch die Gewissenhaftigkeitsaspekte Fleiß und Ordnung.

Bezüglich der Vorhersage der Karriereinteressen (Schein, 1998; s. nächster Abschnitt 4.3.2.3) durch Fleiß und Ordnung ergaben sich unterschiedliche Ergebnisse, sodass von einer partiellen Mediation ausgegangen wurde. Sowohl für Selbstwirksamkeit als auch für die wahrgenommenen eigenen Fähigkeiten konnte ein mediiender Einfluss bei der Vorhersage des Karriereankers *Technisch-funktionale Kompetenz* für beide Gewissenhaftigkeitsaspekte der AGS gefunden werden. Die beiden Selbsteinschätzungskonstrukte mediierten ebenfalls die Vorhersage des Karriere Ankers *General Management*, allerdings nur für den Gewissenhaftigkeitsaspekt Fleiß. Für *Selbstständigkeit/ Unabhängigkeit* sowie *Sicherheit und Beständigkeit* konnten keine mediienden Effekte nachgewiesen werden. Bezüglich der vier weiteren Karriereanker wurden aus Gründen der Ökonomie keine Berechnungen durchgeführt.

4.3.2.3 Karriereinteressen

Das Konzept der Karriereinteressen wurde von Dreier (2008) mittels des *Karriereanker* von Schein (1998) operationalisiert. Die verschiedenen Karriereanker beschreiben die Bereiche einer Person, auf die man aufgrund besonderer Fähigkeiten, persönlicher Beweggründe, Motive oder Werthaltungen besonderen Wert legt oder die man keinesfalls aufgeben möchte. Sie sind als Abbild der Persönlichkeit einer Person zu verstehen bzw. Vorstellung, die man von sich selber hat, und die sich mit der Zeit der Berufstätigkeit herauskristallisieren (Schein, 1998). Bisherige Untersuchungen sprechen dem Autor zufolge für eine zeitliche Stabilität der Karriereanker. Auch wenn eine Person nicht in der Lage sei, den persönlichen Karriereanker zu verwirklichen, so bliebe dieser über die Zeit stabil.

Ursprünglich zum besseren Verständnis von Managerkarrieren entwickelt, fand der Autor acht verschiedene Karriere-Orientierungen, die als festes Element im Selbstkonzept jedes Menschen mehr oder weniger stark verankert sind. Der folgende Abschnitt erläutert jeden der acht Karriereanker kurz und gibt einen Hinweis darauf, wie dieser jeweils konzeptuell mit einem der beiden Gewissenhaftigkeitsaspekte zusammenhängt.

1. *Technische/ funktionale Kompetenz* (TF): Dieser Anker beschreibt den Wunsch einer Person, auf einem Gebiet eine Expertenfunktion zu erlangen bzw. sich auf einem Tätigkeitsgebiet zu spezialisieren. In ihrem Arbeits- oder Interessenfeld streben diese Menschen nach einer möglichst hohen Fachkompetenz und möchten bei der Ausübung ihrer Tätigkeit möglichst freie Hand haben. Aufgrund des Strebens nach Eigenständigkeit und der fachlichen Zielorientierung ist dieser Karriereanker eher mit dem Gewissenhaftigkeits-

aspekt Fleiß in Verbindung zu bringen. Mit dem Ordnungsaspekt gibt es eher wenig inhaltliche Übereinstimmung.

2. *Befähigung zum General Management (GM)*: Dieser Karriereanker beschreibt das Interesse am Führen und Gestalten als solchem. Menschen mit diesem Karriereanker streben in der Regel Management-Positionen an und sind sehr ehrgeizig, möglichst schnell in der Hierarchie eines Unternehmens aufzusteigen, unabhängig von der Branche oder ihrer eigenen Spezialisierung. Sie streben allgemein nach Verantwortung und Eigenständigkeit. Aufgrund dieser Eigenschaften ist ein Zusammenhang dieses Ankers mit dem Gewissenhaftigkeitsaspekt Fleiß konzeptuell sinnvoll. Mit dem Ordnungsaspekt gibt es eher keine inhaltlichen Zusammenhänge.
3. *Selbstständigkeit/ Unabhängigkeit (SU)*: Menschen, denen Selbstständigkeit und Unabhängigkeit wichtig sind, streben danach, ihre Arbeit nach den eigenen Bedürfnissen und Rhythmen zu erledigen. Ihnen fällt es daher schwer, beispielsweise feste Arbeitszeiten und Vorschriften einzuhalten. Sie streben eher danach, ihre eigenen Normen als Bewertungsmaßstab anzulegen und bevorzugen daher eher Berufe, in denen sie unabhängig sind und ihre eigenen Vorschriften machen können. Menschen mit einer hohen Ausprägung in diesem Karriereanker werden daher eher einen hohen Zusammenhang mit dem Fleißaspekt aufweisen. Da sie feste Regeln ablehnen, wird ein Zusammenhang mit der Ordnungsskala negativ ausfallen.
4. *Sicherheit/ Beständigkeit (SB)*: Menschen mit einem Fokus auf Sicherheit und Beständigkeit ist es wichtig, abgesichert zu sein und eine klare Perspektive zu haben. Sie möchten sich nicht um ihre Zukunft sorgen müssen. Festen Regeln und Vorschriften stehen sie positiv gegenüber, da diese ihre Welt strukturieren und überschaubar machen. Werte und Normen sind wichtige Orientierung im Leben und Beruf dieser Personen, weshalb ein starker Zusammenhang mit dem Ordnungsaspekt vorhanden ist. Der Fleißaspekt charakterisiert diese Menschen kaum bis gar nicht, weshalb hier keine Überschneidungen erwartet werden.
5. *Unternehmerische Kreativität (UK)*: Dieser Karriereanker drückt das Bedürfnis aus, etwas Eigenes zu schaffen. Menschen mit diesem beruflichen Fokus haben den Drang, ein eigenes Unternehmen, Produkt oder Dienstleistung zu schaffen. Sie gehen hierbei sehr zielstrebig und eigeninitiiert vor, weshalb hier vor allem Zusammenhänge mit dem Fleißaspekt vorliegen sollten. Inhaltliche Übereinstimmungen mit dem Ordnungskonzept liegen nicht vor.
6. *Dienst oder Hingabe für eine Idee oder Sache (DH)*: Menschen mit diesem Karriereanker ist es wichtig, dass ihre Arbeit die Welt auf irgendeine Art verbessert. Ihnen sind übergeordnete moralische Werte wichtiger als das Einhalten strenger Vorschriften oder Regeln. In diesem Bestreben, die Welt zu verbessern, gehen sie eher pragmatisch und zielorientiert vor, weshalb auch bei diesem Konzept eher konzeptuelle Ähnlichkeiten mit dem Fleißaspekt zu finden sind als mit dem Ordnungsaspekt.
7. *Totale Herausforderung (TH)*: Personen mit diesem Karriereanker ist es vor allem wichtig, immer wieder neue Herausforderungen zu haben und diese bewältigen zu können. Für sie ist das Leben und/ oder der Beruf ein täglicher Kampf bzw. Wettkampf, in dem es nur darum geht zu gewinnen (oder zu verlieren). Da diese Menschen daher ausschließlich an ihren Zielen orientiert sind, ist bezüglich der Gewissenhaftigkeitsaspekte vor allem ein hoher Zusammenhang mit den Items der Fleißskala zu erwarten. Feste Regeln und Vorschriften bzw. Ordnung sind für Menschen mit einem Karriereanker in diesem Bereich nicht so relevant.
8. *Lebensstilintegration (LS)*: Der achte Karriereanker ist bei solchen Menschen hoch ausgeprägt, denen ihre berufliche Karriere nicht besonders wichtig ist oder nicht viel bedeutet. Der Beruf dient diesen Menschen vor allem als Broterwerb und darf mit den persönlichen Interessen der Menschen nicht kollidieren. Ein ausgeglichenes Verhältnis von Arbeit und

Privatleben ist diesen Menschen wichtig. Daher ist ihnen im Beruf weder zielstrebiges noch besonders ordentliches Arbeiten wichtig.

Der Zusammenhang der beschriebenen acht Karriereanker und der Gewissenhaftigkeitsaspekte Fleiß und Ordnung wurde im Rahmen verschiedener Studien bzw. Veröffentlichungen dargestellt und diskutiert. So stellten Dreier et al. (2010) die korrelationsanalytischen Zusammenhänge der AGS mit dem Karriereanker vor, die auch in Tabelle 4-4 dargestellt sind.

Tabelle 4-4: Zusammenhänge der AGS mit den Karriereankern (nach Dreier, Moldzio & Kasper, 2010)

Karriereanker	<i>r</i> AGS-Fleiß	<i>r</i> AGS-Ordnung
Technisch-funktionale Kompetenz	.24**	.05
Befähigung zum General Management	.16*	.06
Selbstständigkeit/ Unabhängigkeit	-.06	-.32**
Sicherheit/ Beständigkeit	.05	.24**
Unternehmerische Kreativität	.14*	-.06
Dienst/ Hingabe für eine Idee oder Sache	.19**	.11
Totale Herausforderung	.30**	-.02
Lebensstilintegration	.05	.05

Anmerkungen: *r* = Korrelationskoeffizient; **p* < .05; ** *p* < .01 (zweiseitig).

Es zeigte sich, dass die Gewissenhaftigkeitsaspekte Fleiß und Ordnung unterschiedlich stark mit den verschiedenen Karriereankern korrelieren. Die Zusammenhänge waren zudem aussagekräftiger als die Zusammenhänge der Skala mit einem allgemeinen Gewissenhaftigkeitsmaß (Dreier & Moldzio, 2010). Trotz der gefundenen Zusammenhänge und der inhaltlichen Verwandtschaft der Gewissenhaftigkeitsaspekte mit einigen Karriereankern handelt es sich bei den Verfahren aber um voneinander unabhängige bzw. verschiedene Skalen, wie Dreier (2008) in ihrer Studie nachweisen konnte. Eine Vorhersage der Karriereinteressen durch die beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung wurde teilweise durch Selbstwirksamkeitserwartungen und/ oder wahrgenommene eigene Fähigkeiten mediiert (siehe vorheriger Abschnitt 4.3.2.2). Dieser Effekt zeigte sich vor allem für den Anker *Technisch/ funktionale Kompetenz*. Allerdings wurden in der Studie von Dreier (2008) aus ökonomischen Gründen nur vier der acht Anker untersucht, weshalb zu dieser Fragestellung nur teilweise Aussagen gemacht werden können.

4.3.2.4 Copingstrategien

Coping ist ein Konzept, das auf das transaktionale Stressmodell von Lazarus (1974) zurückgeht. Mit dem Begriff *Coping* werden kognitive Ansätze und Verhaltensweisen zusammengefasst, die darauf abzielen, spezifische internale und/ oder externale Anforderungen zu bewältigen, die die natürlichen oder allgemeinen Ressourcen einer Person erschöpfen oder sogar übersteigen (Lazarus & Folkman, 1984). Jeder Mensch verwendet unterschiedliche Strategien, um mit Stress umzugehen, abhängig von der wahrgenommenen Bedrohlichkeit einer Situation. Die Bewertung einer Situation wird Lazarus (1974) zufolge in drei Schritten vorgenommen. Zuerst wird eine allgemeine Klassifizierung der Situation als herausfordernd, bedrohlich oder potenziell schädigend bzw. verlustbringend vorgenommen (= *Primary Appraisal*). Danach erfolgt eine Sekundärbewertung (= *Secondary Appraisal*), in der überprüft wird, ob eine Situation mit den verfügbaren Ressourcen zu bewältigen ist und eine Bewältigungsstrategie entworfen, die neben den Situationsparametern auch von individuellen Persönlichkeitseigenschaften einer Person sowie deren kognitiven Ressourcen abhängig ist. Die Bewältigungsstrategie wird als *Copingstrategie* bezeichnet. Auch hier kann man nach Lazarus (1974) drei Arten von Copingstrategien unterscheiden: *problem-*, *emotions-* und *bewertungsorientiertes Coping*.

In ihrer Metaanalyse fanden Connor-Smith und Flachsbart (2007) einen starken Zusammenhang zwischen Gewissenhaftigkeit und Copingstrategien, wie beispielsweise Problemlöseverhalten ($r = .30$) oder kognitiver Restrukturierung ($r = .20$). Für soziale oder emotionsorientierte Bewältigungsstrategien hingegen konnten die Forscher jedoch keine Zusammenhänge nachweisen.

Korinteli (2009) untersuchte den Zusammenhang der beiden Gewissenhaftigkeitsskalen der AGS mit aktiven Copingstrategien in einer Stichprobe von Arbeitnehmern in unterschiedlichen Unternehmen ($N = 340$). Unter aktiven Copingstrategien wurden in dieser Untersuchung solche Bewältigungsmechanismen zusammengefasst, die potenziell belastende oder gefährliche Situationen in der Zukunft antizipieren. Hierzu wurden proaktive, reflektive und strategisch-planende Bewältigungsmechanismen in einem Fragebogen erhoben und zu einem Gesamtwert verrechnet. Die Autorin fand einen starken korrelativen Zusammenhang zwischen aktivem Coping und dem Fleißaspekt ($r = .52, p < .01$) sowie einen moderaten Zusammenhang mit dem Ordnungsaspekt ($r = .32, p < .01$). Diese Tendenz spiegelte sich auch in den Gewissenhaftigkeitsaspekten der BFAS wider ($r_{Ord.} = .35, p < .01$ bzw. $r_{Fleiß} = .52, p < .01$). Allerdings konnte weder ein mediiender noch ein moderierender Effekt aktiver Copingstrategien bei der Vorhersage von Berufserfolg durch Gewissenhaftigkeit gefunden werden.

4.3.2.5 Selbstführungsstrategien

Selbstführung ist im Rahmen der Selbstmanagement-Theorie als Selbstbeeinflussungsprozess definiert, durch den man die notwendige Selbstmotivation erreicht, um ein bestimmtes Ziel zu erreichen bzw. sich in einer erwünschten Art und Weise zu verhalten (Struckmeier, 2009). Es gibt drei Klassen von Selbstführungsstrategien: *verhaltensorientierte Strategien* (behavior-focused strategies), *konstruktive Gedankenmuster* (constructive thought pattern strategies) und *natürliche Belohnungsstrategien* (natural reward strategies) (Andreßen & Konradt, 2007). Verschiedene Forscher postulierten, dass Gewissenhaftigkeit und Selbstführung konzeptuell ähnlich seien, da im Selbstführungskonstrukt auch Verhaltensweisen wie beispielsweise Zielstrebigkeit oder Verantwortungsübernahme erfasst würden. Struckmeier (2009) überprüfte die Zusammenhänge der beiden Gewissenhaftigkeitsaspekte mit dem Konzept der Selbstführung. Die Ergebnisse der Korrelationsanalysen der Studie sind in Tabelle 4-5 zusammengefasst.

Tabelle 4-5: Korrelationen der AGS mit Selbstführungsstrategien (nach Struckmeier, 2009)

Variable	Behavior-focused Strategies	Constructive Thought Pattern Strategies	Natural Reward Strategies	Self-Leadership gesamt
AGS-Fleiß	.35***	.28***	.28***	.39***
AGS-Ordnung	.19**	.18**	-.05	.19***

Anmerkungen: ** $p < .01$; *** $p < .001$ (zweiseitig).

Die Fleißskala wies höhere Zusammenhänge mit den Selbstführungsstrategien auf als die Ordnungsskala. Dies entspricht der konzeptionellen Definition der beiden Konstrukte (vgl. Abschnitt 4.1), die den Fleißaspekt vor allem zielorientierte und eigeninitiierte Arbeitsweisen erfassen lässt. Eine Korrelation der Konzepte aufgrund konzeptueller Gemeinsamkeiten konnte mittels konfirmatorischer Faktorenanalysen ausgeschlossen werden.

4.3.2.6 Schulische Leistungskriterien

Die Zusammenhänge zwischen schulischen Leistungskriterien und Gewissenhaftigkeit wurden bereits in Abschnitt 2.7.2 ausführlich dargestellt. Korinteli (2009) konnte in ihrer Untersuchung einige Zusammenhänge zwischen dem Ordnungsaspekt und dem Bildungsniveau der Studienteilnehmer ($N = 430$) auffinden. So fand die Autorin jeweils negative Zusammenhänge des Ordnungsaspekts mit dem Schulbildungs- ($r = -.22, p < .01$) und Ausbildungsniveau ($r = -$

.20, $p < .01$) der befragten Mitarbeiter. Die Ergebnisse weisen darauf hin, dass je weniger gut ausgebildet ein Mitarbeiter ist, desto weniger ordentlich ist diese Person.

Der Zusammenhang der Gewissenhaftigkeitsaspekte Fleiß und Ordnung mit Schulleistung konnte bisher allerdings noch nicht nachgewiesen werden. So fand Dreier (2008) in ihrer Untersuchung in einer Stichprobe amerikanischer Collegestudenten ($N = 217$) keine signifikanten Korrelationen der AGS mit dem aktuellen Notendurchschnitt (GPA). Lediglich die Fleiß-Skala der BFAS korrelierte leicht mit dem Schulleistungskriterium ($r = .15$, $p < .05$).

4.3.2.7 Berufliche Leistungskriterien

Die Zusammenhänge zwischen beruflichen Leistungskriterien und Gewissenhaftigkeit wurden bereits in Abschnitt 2.7.1 ausführlich dargestellt. Im Rahmen der Untersuchung von Struckmeier (2009) wurde aufgrund der theoretischen Befunde zur prädiktiven Validität des Gewissenhaftigkeitskonstruktes bzw. der Gewissenhaftigkeitsfacetten der Zusammenhang der AGS mit Arbeitsleistung erfasst. Arbeitsleistung wurde in dieser Untersuchung durch vier Items erfasst, die das sogenannte *in-role behavior*, d.h. die allgemeinen Aufgaben eines Mitarbeiters, maßen. Außerdem wurde das allgemeine Arbeitsniveau jedes Mitarbeiters erfragt. Alle fünf Einschätzungen wurden durch die jeweiligen Vorgesetzten der Mitarbeiter ($N = 174$) vorgenommen und am Ende in einen Gesamtscore verrechnet, der als Index für die Arbeitsleistung herangezogen wurde.

Mittels hierarchischer Regressionsanalysen wurde die Vorhersagekraft der beiden Gewissenhaftigkeitsaspekte in Bezug auf die Arbeitsleistung ermittelt. Die Ergebnisse zeigten einen positiven Effekt von Fleiß auf Arbeitsleistung ($\beta = .22$, $p < .01$). Für Ordnung ergab sich ein negativer Effekt ($\beta = -.22$, $p < .01$). Diese Effekte wurden in Ansätzen auch von den Gewissenhaftigkeitsaspekten der BFAS gezeigt. Insgesamt konnten die AGS 9 % der Varianz aufklären.

Die Ergebnisse sind als erste Hinweise dafür zu werten, dass die AGS bzw. die damit erfassten Gewissenhaftigkeitsaspekte Fleiß und Ordnung signifikant Arbeitsleistung vorhersagen können. Im Vergleich mit einer nicht-arbeitsbezogen formulierten Skala erwiesen sich die AGS hierfür als das besser geeignete Instrument.

Korinteli (2009) konnte in ihrer Studie diese Effekte nicht replizieren. Zwar konnte sie ebenfalls einen moderaten negativen Zusammenhang zwischen dem Ordnungsaspekt und Arbeitsleistung nachweisen ($r = -.16$, $p < .05$), jedoch fand sie keinen bedeutsamen Zusammenhang zwischen Fleiß und Arbeitsleistung ($r = .10$, $p < .10$). Auch zwischen den Gewissenhaftigkeitsaspekten und Hierarchieebene, Dauer der Beschäftigung sowie Arbeitserfahrung in Jahren konnte die Autorin keine signifikanten Zusammenhänge auffinden.

4.4 Itemreanalyse

Basierend auf den Erkenntnissen der beiden Studien zur Skalenüberprüfung der AGS sowie den gefundenen Zusammenhängen mit weiteren Persönlichkeits- und Leistungskonstrukten wurden die AGS als differenziertes und nützliches Instrument zur Ergänzung der bestehenden eignungsdiagnostischen Instrumente im Bereich der beruflichen Eignungsdiagnostik eingestuft (Dreier, 2008; Dreier et al., 2010; Dreier & Moldzio, 2010; Korinteli, 2009; Struckmeier, 2009). Die Itemkennwerte waren vielversprechend, jedoch noch verbesserungswürdig. In der deutschen Stichprobe hatte sich insbesondere ein Item als problematisch erwiesen, weshalb im Sommer 2009 für die deutschsprachige Version der AGS eine Itemreanalyse durchgeführt wurde. Das Ziel der Itemreanalyse war eine Verbesserung der Gütekriterien des Fragebogens.

Die Entwicklung der englischsprachigen Testversion wurde aufgrund ökonomischer und durchführungstechnischer Probleme ab diesem Zeitpunkt nicht weiter verfolgt. Aus diesem

Grund beziehen sich alle folgenden Informationen von nun an ausschließlich auf die deutschsprachige Version der AGS.

Die Itemreanalyse lief in drei Schritten ab. Zuerst wurden von drei Experten wiederum Items generiert bzw. ausgewählt. Als Grundlage hierzu dienten sowohl die alten Items der AGS und ein Pool an neu generierten Items. Außerdem wurden in den Itempool die schon bestehenden Items des Beruflichen Selbstbeschreibungsinventars (BSI; Glug, 2004) mit aufgenommen, da diese auf Basis vergleichbarer, theoretischer Überlegungen konstruiert wurden wie die Items der AGS. Insgesamt ergab sich hieraus ein Itempool von 52 Aussagen, die in die im Rahmen der Itemreanalyse evaluiert wurden.

Der zweite Schritt der Itemreanalyse war die Bewertung der Items durch eine Stichprobe von Arbeitnehmern verschiedenster Berufszweige ($N=49$). Hierzu wurden die Items im Rahmen einer Online-Studie beantwortet und evaluiert. Die Testpersonen erhielten per E-Mail eine Einladung zur Teilnahme an der Evaluationsstudie. Die Teilnehmer schätzten sich selbst bzw. ihr Arbeitsverhalten anhand der vorgegebenen Items ein. Neben der Selbsteinschätzung war in der Evaluationsstudie auch eine Kommentarfunktion enthalten, in der die Testpersonen ihre Eindrücke und Anmerkungen zu den Items notieren konnten. Die Experten erhofften sich hierüber Aufschluss über die Itemverständlichkeit oder -eignung sowie ggf. weitere Formulierungsvorschläge. Allerdings ergaben sich aus der Kommentarfunktion keine nennenswerten Informationen oder Vorschläge, die die Skalenkonstruktion beeinflusst hätten, weshalb hierauf nicht detaillierter eingegangen wird.

Der dritte Schritt der Itemreanalyse war die Analyse der Daten. Unabhängig voneinander führten zwei Experten Faktoren- und Itemanalysen durch. Basierend auf der Reliabilitätsanalyse bzw. Itemkennwerten wie der internen Konsistenz, Trennschärfen und Itemschwierigkeiten wurden insgesamt 30 Items aus dem Itempool zur Bildung der AGS ausgewählt. Die Items der Endversion der AGS bestanden aus den 23 Items der Vorversion der AGS, einem Item des BSI (Glug, 2004) sowie sechs neu entwickelten Items. Die Items waren auf jeden der beiden Gewissenhaftigkeitsaspekte gleichverteilt, d.h. sowohl Fleiß als auch Ordnung wurden jeweils mit 15 Items erfasst. Diese wurden durch Übereinstimmung der Experten ausgewählt. Bei unterschiedlichen Meinungen wurden die Items diskutiert, bis Einigkeit über das betreffende Item vorlag.

Mittels explorativer Faktorenanalyse wurden zwei Faktoren ermittelt, die die Daten hinreichend gut abbildeten, womit die Zweidimensionalität der AGS bestätigt wurde.

Die Skalenzugehörigkeit jedes Items wurde mittels Faktorladungen überprüft. Basierend auf diesen faktorenanalytischen Analysen wurden die Itemkennwerte für die Aspekte Fleiß und Ordnung ermittelt.

Die Fleißskala bestand aus insgesamt 25 Items, wovon acht negativ gepolt waren. Die interne Konsistenz der 25 Fleiß-Items war gut ($\alpha=.86$). Aus diesen wurden 15 Items zur Bildung des Fleiß-Aspektes der AGS ausgewählt. Keines der Items war negativ gepolt. Das Löschen der nicht-verwendeten Items führte zu einer guten internen Konsistenz von $\alpha=.89$ der verbleibenden 15 Items.

Die Ordnungsskala bestand insgesamt aus 27 Items, wovon sieben negativ gepolt waren. Die interne Konsistenz der 27 Ordnungs-Items war mit $\alpha=.88$ gut ausgeprägt. Auch für diese Skala wurden 15 Items ausgewählt, die die Ordnungs-Skala der AGS bildeten. Hiervon waren vier negativ gepolt. Neben den elf alten AGS-Items wurden drei neu generierte Items sowie ein Item des BSI (Glug, 2004) in die neue Ordnungs-Skala übernommen. Die 15 Ordnungs-Items verfügten nach Löschung der nicht-verwendeten Items über eine interne Konsistenz von $\alpha=.87$, was ebenfalls als gut zu bezeichnen ist.

Eine detaillierte Darstellung der Itemreanalyse sowie der Ergebnisse der zugehörigen Faktoren- und Reliabilitätsanalysen ist bei Kasper (in Vorbereitung) nachzulesen.

Die weitere Überprüfung der Skalengüte der AGS ist u.a. ein Ziel dieser Untersuchung.

4.5 Beispielitems

In dieser Studie wurde die 30-Item-Version der AGS verwendet, in der die Items der beiden Gewissenhaftigkeitsaspekte abwechselnd abgefragt werden.

Einige Beispiel-Items für die Fleißskala lauten:

- a. *„Herausforderungen bei der Arbeit spornen mich an.“*
- b. *„Mein Ziel ist es, möglichst viel bei der Arbeit zu leisten.“*
- c. *„Ich bin produktiver als andere, die einen ähnlichen Job haben.“*
- d. *„Ich stecke viel Energie in meine Arbeit.“*
- e. *„Neue Aufgaben motivieren mich immer wieder.“*

Einige Beispiel-Items für die Ordnungsskala lauten:

- a. *„Ich mag geregelte Arbeitszeiten.“*
- b. *„Das Einhalten von Regeln ist mir wichtig.“*
- c. *„Ich arbeite genau.“*
- d. *„Ich gestalte meinen Arbeitsplatz übersichtlich.“*
- e. *„Ich bin kein ordnungsliebender Mensch.“* (negativ gepoltes Item)

Aus urheberrechtlichen Gründen können im Rahmen dieser Forschungsarbeit nicht alle Items der AGS zitiert werden.

5. Untersuchungsplanung/ Operationalisierung

In diesem Kapitel wird die Planung der Untersuchungen bzw. Bewerbertestungen zur Überprüfung der in Abschnitt 3 abgeleiteten Fragestellungen erläutert. Im ersten Abschnitt dieses Kapitels (5.1) wird dargestellt, wie die unabhängigen Variablen auf Prädiktorebene operationalisiert wurden. Der zweite Abschnitt beinhaltet die Operationalisierung der Kriteriumsvariablen (5.2). Die Erhebung der demografischen Merkmale der Untersuchungsteilnehmer (5.3.1), möglicher konfundierender Einflüsse (5.3.2) und weiterer Variablen (5.3.3) ist im dritten Abschnitt dargestellt. Eine Beschreibung des Versuchsplans befindet sich im letzten Teil (Abschnitt 5.4).

5.1 Operationalisierung der Variablen auf Prädiktorebene

Ein Ziel dieser Arbeit war es, die zugrunde liegende Faktorenstruktur sowie die prädiktive, inkrementelle und Konstruktvalidität der Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS) anhand einer Feldstichprobe zu überprüfen. Entsprechend wurden neben den AGS weitere Persönlichkeitstests zur Erfassung der Gewissenhaftigkeitsdimension (Abschnitt 5.1.1) verwendet. Diese sowie einige Selbsteinschätzungskonstrukte (Abschnitt 5.1.2 und 5.1.3) wurden auch für die Überprüfung des theoretischen Vorhersagemodells herangezogen. Zur Bestimmung der inkrementellen Validität wurde die Vorhersagekraft der AGS darüber hinaus zusätzlich mit der prädiktiven Validität verschiedener Leistungstests (Abschnitt 5.1.4) verglichen.

Die Auswahl der Test- und Fragebogenmaße wurde hauptsächlich durch die praktischen Ansprüche der Kunden an das zugrunde liegende Auswahlverfahren bestimmt. Aus diesem Grund sind alle in dieser Untersuchung verwendeten Persönlichkeitstests Selbstauskünfte. Es handelt sich jedoch bei allen in der Untersuchung verwendeten Testverfahren um etablierte Maße, deren prädiktive Validität für berufseignungsdiagnostische Zwecke ausreichend belegt ist.

5.1.1 Gewissenhaftigkeit

Das Merkmal Gewissenhaftigkeit wurde mit Hilfe von zwei Persönlichkeitsinventaren erfasst. Die Probanden beantworteten jeweils die AGS (s. Kapitel 4) zur Erfassung der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie das NEO-FFI (Borkenau & Ostendorf, 1993, 2008) als Maß für allgemeine Gewissenhaftigkeit. Der NEO-FFI diene außerdem zur Erfassung der weiteren relevanten Persönlichkeitseigenschaften des Fünf-Faktoren-Modells: Neurotizismus bzw. Ausgeglichenheit, Extraversion, Offenheit für Erfahrungen sowie Verträglichkeit. Einige dieser Merkmale sollen auch zur Konstruktvalidierung der AGS herangezogen werden (siehe Abschnitt 6.6).

AGS. Da die AGS im vorherigen Kapitel 4 bereits detailliert beschrieben wurden, wird im Folgenden nur das NEO-FFI beschrieben.

NEO-FFI. Beim NEO-FFI handelt es sich um die deutschsprachige Übersetzung und Adaptation des NEO Five-Factor Inventory (NEO-FFI) von Costa und McCrae (1992). Das NEO-FFI ist ein faktorenanalytisch konstruiertes, multidimensionales Fragebogenverfahren, welches der Erfassung individueller Merkmalsausprägungen dient. Es erfasst anhand von fünf Skalen, denen jeweils zwölf Items zugeordnet sind, die Persönlichkeitsdimensionen des Fünf-Faktoren-Modells: Neurotizismus ($\alpha = .87$), Extraversion ($\alpha = .81$), Offenheit für Erfahrungen ($\alpha = .75$), Verträglichkeit ($\alpha = .72$) sowie Gewissenhaftigkeit ($\alpha = .84$). Die Retest-Reliabilitäten nach fünf Jahren sind ähnlich hoch ($r = .71$ bis $r = .82$). Faktorenanalysen der Items des NEO-FFI sowie gemeinsame Faktorenanalysen der Skalen des NEO-FFI und anderer Persönlich-

keitsinventare zeigen die Konstruktvalidität des Verfahrens. Untermuert wird diese durch substanzielle Korrelationen zwischen den Skalen des NEO-FFI und Selbsteinschätzungen auf Adjektivskalen ($r = .54$ bis $r = .80$) sowie Korrelationen zwischen Selbst- und Bekanntenbeurteilungen mittels NEO-FFI ($r = .49$ bis $r = .61$) (Borkenau & Ostendorf, 2008).

Aufgrund der einfachen Instruierbarkeit, des geringen Durchführungsaufwandes und der schnellen Auswertung des Persönlichkeitsinventars sowie dessen prädiktiver Validität für die berufliche Eignungsdiagnostik, wurde das NEO-FFI zur Erfassung der Persönlichkeitsmerkmale in der Auszubildendenauswahl eingesetzt.

Zwar besitzt das NEO-FFI keine Vergleichswerte für Stichproben, in denen auch Probanden unter 16 Jahren getestet werden, den Autoren zufolge ist ein sinnvoller Einsatz des NEO-FFI aber auch bei 14-16-jährigen möglich. Es bleibt allerdings fraglich, wie stabil die Persönlichkeitsausprägungen bei so jungen Testteilnehmern bereits sind. Der Einsatz der Skala in jüngeren Stichproben wird, vor allem auch für Forschungsarbeiten, jedoch empfohlen. Eine Verwendung der Skala im Rahmen dieser Arbeit ist somit gerechtfertigt.

Tabelle 5-1 enthält jeweils ein Beispiel-Item für jedes der fünf Merkmale. Als Antwortformat steht eine fünffach gestufte Likert-Skala von 0 = „starke Ablehnung“ über 1 = „Ablehnung“, 2 = „neutral“ und 3 = „Zustimmung“ bis 4 = „starke Zustimmung“ zur Verfügung.

Tabelle 5-1: Beispiel-Items für die Skalen des NEO-FFI (nach Borkenau & Ostendorf, 1993, 2008)

Dimension	Beispiel-Item
Neurotizismus	„Wenn ich unter starkem Stress stehe, fühle ich mich manchmal, als ob ich zusammenbräche.“
Extraversion	„Ich habe gerne viele Leute um mich herum.“
Offenheit für Erfahrungen	„Ich probiere oft neue und fremde Speisen aus.“
Verträglichkeit	„Ich versuche zu jedem, dem ich begegne, freundlich zu sein.“
Gewissenhaftigkeit	„Wenn ich eine Verpflichtung eingehe, so kann man sich auf mich bestimmt verlassen.“

5.1.2 Selbstwirksamkeit

Die Selbstwirksamkeitserwartungen der Probanden wurden mit der Beruflichen Selbstwirksamkeitsskala (BSW) von Schyns und von Collani (2002) erfasst. Es handelt sich hierbei um eine eindimensionale Skala, die das Selbstwirksamkeitskonzept mit 19 Items auf einer 5-stufigen Likert-Skala von 0 = „starke Ablehnung“ über 1 = „Ablehnung“, 2 = „neutral“ und 3 = „Zustimmung“ bis 4 = „starke Zustimmung“ erfasst. Die Ergebnisse von Schyns und von Collani (2002) deuten darauf hin, dass die Skala gut für den Einsatz in der angewandten Forschung sowie in der Organisationspsychologie geeignet ist. Das Instrument ist berufsbezogen formuliert, aber die Items wurden dennoch so allgemein bzw. aufgabenübergreifend gehalten, dass die BSW gut in der allgemeinen Organisationspsychologie bzw. in verschiedenen Unternehmen in unterschiedlichen Hierarchieebenen verwendet werden kann. Die interne Konsistenz der Selbstwirksamkeitsskala beträgt $\alpha = .92$ (Cronbachs Alpha) und ist somit als sehr gut zu bezeichnen. Die Split-Half-Reliabilität ist ebenfalls gut ($\alpha = .89$). Den Autoren zufolge zeigt die Berufliche Selbstwirksamkeitsskala Zusammenhänge beispielsweise mit Aufgabenanforderungen ($r = .22$, $p < .01$), Leader-Member-Exchange ($r = .17$, $p < .01$), Arbeitszufriedenheit in Bezug auf den Vorgesetzten ($r = .16$, $p < .01$) sowie in Bezug auf die Aufgabenanforderungen ($r = .31$, $p < .01$). Auch Zusammenhänge mit Commitment sowie Persönlich-

keitsvariablen wurden gefunden. Ein Beispiel-Item lautet *„Beruflichen Schwierigkeiten sehe ich gelassen entgegen, weil ich mich immer auf meine Fähigkeiten verlassen kann.“*.

Die Items der Selbstwirksamkeitsskala wurden in dieser Studie zur Maskierung mit den Items des NEO-FFI vermischt. Die Instruktion der Skala erfolgte somit gleichzeitig mit dem NEO-FFI.

5.1.3 Wahrgenommene eigene Fähigkeiten

Die wahrgenommenen eigenen Fähigkeiten die Ausbildung bzw. Schule betreffend wurden im Rahmen der Nachtestung der Azubi-Stichprobe bzw. im Rahmen der Testung der Schüler mit Hilfe von acht Items erhoben, die aus der Studie von Nofle und Robins (2007) stammten und von drei unabhängigen Experten aus dem Englischen ins Deutsche übersetzt sowie für die beschriebene Untersuchung adaptiert wurden. Später wurden die Items noch einmal von einem Muttersprachler rückübersetzt, sodass von einer angemessenen inhaltlichen Übertragbarkeit der Items ausgegangen werden kann.

Die wahrgenommenen eigenen Fähigkeiten können formal in zwei Bereiche unterteilt werden: die wahrgenommenen eigenen (schulischen) Fähigkeiten und den wahrgenommenen akademischen Erfolg bzw. die wahrgenommene eigene Leistung.

Die wahrgenommenen akademischen Fähigkeiten wurden mit Hilfe von sechs Items erfasst, die wiederum in zwei Subfacetten unterteilt werden können. Zum einen erfassen drei Items direkte Selbsteinschätzungen bzw. Angaben über die eigenen Fähigkeiten. Ein Item lautete z.B. *„Wie schätzen Sie Ihre schulischen Fähigkeiten im Vergleich zum durchschnittlichen XY-Schüler ein?“*. Das Antwortformat zu diesem Item war eine fünfstufige Skala mit den Antwortmöglichkeiten 0= *„schwach“*, 1= *„unterdurchschnittlich“*, 2= *„durchschnittlich“*, 3= *„überdurchschnittlich“* und 4= *„hervorragend“*. Die beiden anderen Items lauteten *„Ich bin zuversichtlich, dass ich in der Ausbildung erfolgreich sein werde.“* sowie *„Ich denke, ich habe die Fähigkeiten, mich in der Ausbildung zu behaupten.“*. Diese zwei Items wurden ans Ende der Beruflichen Selbstwirksamkeitsskala gesetzt, da sie dasselbe 5-stufige Antwortformat hatten wie die BSW-Skala (s. Abschnitt 5.1.2).

Die anderen drei Items können als erwartete Leistungen in der Schule bzw. Ausbildung zusammengefasst werden. Sie lauteten *„Welches wäre die beste Durchschnittsnote, die Sie fähig wären (zum Ende der Ausbildung) zu erreichen?“*, *„Realistisch betrachtet, welche Durchschnittsnote werden Sie Ihrer Meinung nach (zum Ende der Ausbildung) wahrscheinlich erreichen?“* sowie *„Was wäre der niedrigste Notendurchschnitt, mit dem Sie für sich zufrieden wären?“*. Das dazugehörige Antwortformat war jeweils frei, das heißt die befragten Auszubildenden und Schüler konnten jeweils frei ihren Notendurchschnitt eintragen.

Die Bemühungen in der Schule wurden mit zwei Items erfasst, die das Verhalten und den eigenen Einsatz zur Erreichung von Schul- bzw. Ausbildungszielen messen. Bei einem Item handelte es sich um eine Einschätzung des eigenen Einsatzes (*„Wie sehr haben Sie sich im letzten Schuljahr angestrengt, um die Noten zu erreichen, die Sie sich zum Ziel gesetzt haben?“*), das auf einer 5-stufigen Skala mit den Ausprägungen 0= *„gar nicht“*, 1= *„unterdurchschnittlich“*, 2= *„durchschnittlich“*, 3= *„überdurchschnittlich“* und 4= *„sehr stark“* bewertet wurde. Beim zweiten Item (*„Wie viele Stunden pro Woche haben Sie durchschnittlich im letzten Schuljahr außerhalb des Unterrichts für die Schule aufgebracht (z.B. für Klausurvorbereitung, Hausaufgaben)?“*) war jeder der Befragten aufgefordert, die Anzahl der Stunden anzugeben, die er pro Woche außerhalb der Schule bzw. Ausbildung in diese investiert hatte.

Nofle und Robins (2007) berichten für die Items der wahrgenommenen akademischen Fähigkeiten eine gute bis sehr gute interne Konsistenz (Cronbachs Alpha .84 bis .91). Dreier (2008)

berichtete über eine zufriedenstellende interne Konsistenz der Items (.72), allerdings wurden in ihrer Untersuchung – wie auch in der vorliegenden Studie – die Items zur wahrgenommen verbalen Intelligenz nicht mit erhoben.

Die aufgeführten Items sind eine kurze und effektive Methode, die wahrgenommenen eigenen Fähigkeiten zuverlässig zu erfassen.

5.1.4 Allgemeine kognitive Fähigkeiten

Neben den Persönlichkeitsvariablen wurden im Rahmen der Azubi-Auswahl unter anderem auch einige Untertests bzw. -skalen aus Leistungstests zur Vorhersage des Berufserfolges und Unterstützung der Auswahlentscheidung durchgeführt. Die Leistungsdaten sind somit nur für die Stichprobe der Azubi-Bewerber vorhanden. Je nach Ausbildungsrichtung wurden neben Rechtschreibkenntnissen und technischem Verständnis vor allem die Rechenfertigkeiten und das logisch-schlussfolgernde Denken als Indikatoren für die allgemeine kognitive Leistungsfähigkeit der Auszubildenden erfasst. Bei der Auswahl der Testverfahren wurde jeweils darauf geachtet, dass diese sowohl für die Anwendung in der Berufseignungsdiagnostik sowie Anwendung in jüngeren Stichproben geeignet bzw. normiert sind. Im Folgenden sind nur die Testverfahren beschrieben, die in dieser Studie durch die Mittelung der vier zugehörigen *t*-Werte der Untertests in die Analysen als allgemeiner Intelligenzindikator eingegangen sind. Dies sind die Untertests 3, 4 und 8 des Leistungsprüfsystems (LPS; Horn, 1983) sowie die Skala 4 des IST 2000 (Intelligenz-Struktur-Test-2000; Amthauer, Brocke, Liepmann & Beauducel, 1999), die in den folgenden Abschnitten (5.1.4.1 und 5.1.4.2) beschrieben werden.

5.1.4.1 Leistungsprüfsystem

Das Leistungsprüfsystem (LPS; Horn, 1983) ist ein Intelligenztest, der auf den Veröffentlichungen von Thurstone basiert und dessen Primärfaktoren der Begabung mit jeweils zwei Untertests erfasst. Allgemein erfasst der Test die „flüssige“ Intelligenz, d.h. die allgemeine Fähigkeit, wichtige Beziehungen in komplexen Situationen zu erkennen. Das LPS wurde vor allem mit dem Ziel entwickelt, die zugrundeliegende Begabungsstruktur von Schülern zu erfassen, um bei der Wahl des günstigsten Ausbildungs- und Berufsweges Orientierung zu bieten, da solche Verfahren zur Zeit der ersten Entwicklung des LPS noch nicht verfügbar waren (Horn, 1983). Die langjährige Anwendung des LPS u.a. bei Bewerbern und Schülern bestätigt den LPS als wertvolle und nützliche Entscheidungshilfe in Auswahl-situationen. Es bestehen hochsignifikante Korrelationen mit Schulnoten sowohl mit der Gesamtleistung als auch einzelnen Untertests des LPS. Auch mit anderen Intelligenzmaßen (z.B. dem IST, s. Abschnitt 5.1.4.2) bestehen signifikante Zusammenhänge ($r = .74$), sodass von einem inhaltsvaliden Verfahren ausgegangen werden kann. Allerdings gibt es im LPS auch Skalen, für die es beispielsweise im IST keine Entsprechungen gibt, was den Einsatz ausgewählter Subskalen des LPS im Rahmen der Auszubildendenauswahl begründet. Der LPS wurde an über $N = 10.000$ Personen geeicht (Horn, 1983). Als Vergleichsmaßstab stehen daher repräsentative Stichproben aus verschiedenen Altersstufen zur Verfügung, in dem insbesondere auch Augenmerk darauf gelegt wurde, auch leistungsschwache Personen miteinzubeziehen. Daraus resultieren altersgestaffelte Normtabellen für alle Untertests für 10- bis 50-Jährige. Insgesamt handelt es sich um ein sehr zuverlässiges Verfahren. Die Retest-Reliabilität für das Gesamtergebnis des LPS beträgt $r = .95$. Die korrigierten Halbierungszuverlässigkeiten der Untertests sind ebenfalls sehr gut ($r = .90$ bis $.99$). Es gibt zwei Testversionen A und B, welche sich in der Schwierigkeitsabstufung und Gültigkeit entsprechen. In dieser Untersuchung wurden die Untertests der Testform A verwendet. Die objektive Instruktion sowie Auswertung und Interpretation des LPS ist durch festgelegte Anleitungen und Schablonen gewährleistet. Durch Auszählen der richtigen Antworten je Untertest und Aufsummierung dieser zu einem Summen- bzw. Rohwert je Skala wird die Lösungsqualität ermittelt. Es gibt jeweils maximal

40 Punkte pro Untertest zu erreichen. Der Testwert wird anhand der jeweiligen Altersnorm ausgewertet.

Gemeinsam sind die Untertests 3 und 4 ein guter Indikator für die allgemeine Denkfähigkeit und Intelligenz einer Person. Eine hohe Leistung wird durch das schnelle Erfassen von Abständen und Richtungen in Buchstaben- und Zahlenfolgen sowie ein Merken – über Störungen durch eingeschobene Denkinhalte hinweg – erleichtert. Weist ein Proband ein sehr gutes Ergebnis in diesen Skalen auf, darf auf eine potenziell hohe mathematische Begabung geschlossen werden. Horn (1983) zufolge zeigten Studenten eines polytechnischen Studiums signifikant bessere Ergebnisse in den Untertests 3 und 4 als Psychologiestudenten. Da sowohl kaufmännische als auch technische Auszubildende über gute mathematische Fähigkeiten im logisch-schlussfolgernden Denken verfügen sollten, wurden die Untertests 3 und 4 sowohl in der kaufmännischen als auch in der technischen Testbatterie verwendet. Die Untertests werden im Folgenden erläutert.

Untertest 3 erfasst die Denkfähigkeit bzw. das Erkennen von Regeln sowie logisch-schlussfolgerndes Denken anhand einer figuralen Aufgabenstellung. Die Probanden müssen in einer Reihe von neun Figuren die einfachste Gesetzmäßigkeit herausfinden und das Zeichen identifizieren bzw. markieren, das nicht in diese Reihe passt. In diesem Test werden Horn (1983) zufolge visuell begabte Personen bevorzugt. Zudem sei der Untertest weitgehend unabhängig von schulischer Vorbildung. Die Retestreliabilität beträgt $r = .66$ und die Korrelation des Testergebnisses mit der Gesamtleistung $r = .81$. Die Durchführung des Tests dauert fünf Minuten.

Untertest 4 ist ähnlich aufgebaut wie Untertest 3. Aus einer Reihe von neun Gliedern (Buchstaben oder Zahlen) soll der Proband das Zeichen finden, das nicht der Gesetzmäßigkeit entspricht und somit nicht in die Zeile hineinpasst. Es werden wiederum Denkfähigkeit und das Erkennen von Regelmäßigkeiten bzw. logisch-schlussfolgerndes Denken erfasst. Die Retestreliabilität ist zufriedenstellend ($r = .77$). Der Untertest korreliert zu $.84$ mit dem Gesamtergebnis. Die Durchführungsdauer beträgt acht Minuten. Das erfolgreiche Bearbeiten der Aufgabe setzt die Kenntnis von Zahlen und Buchstaben voraus.

Technische Begabung ist eine Kombination von Leistungen, die am Zustandekommen technischer Produkte einen wesentlichen Anteil haben. Bei einer Zusammenfassung der Untertests 7 bis 10 ist eine gute und schnelle Erfassung dieser Begabung möglich – unabhängig davon, ob jemand Erfahrungen auf diesem Gebiet besitzt. Für Architekten, Maschinenbauer und ähnliche Berufe ist eine hohe Leistung insbesondere in den Untertests 8, 9 und 10 von großer Bedeutung. Zukünftige Betriebsingenieure wiesen auf diesen Skalen zudem sehr hohe Werte auf (Horn, 1983). Solche Testpersonen, die einen gehobenen technischen Beruf ergreifen wollen (z.B. Maschinenbau), sollten dem Testautor zufolge besonders in den Aufgabenreihen 7 bis 11 gute Leistungen zeigen – und zwar auch relativ zu den übrigen Aufgabengruppen. Obwohl die in diesen Skalen gemessenen Eigenschaften also sehr wichtig für technische Berufe sind, werden diese Fähigkeiten dem Testautor zufolge nicht von Lehrern erkannt. Der LPS erfüllt hier daher an dieser Stelle eine wichtige Ergänzungsfunktion. Die Zuverlässigkeit der vier Untertests ist mit $r = .83$ (Retestreliabilität) als gut zu bewerten. Aus diesen Gründen wurden die Untertests 7 bis 10 in das technische Testpaket mit übernommen. Da aber auch in kaufmännischen Berufen, vor allem in technischen Betrieben, wie dies bei den Auftraggebern der Ausbildungstestungen der Fall war, technisches Grundverständnis wichtig ist, wurde darüber hinaus Untertest 8 aufgrund seiner einfachen Verständlichkeit auch in die kaufmännische Testbatterie integriert.

Untertest 8 erfasst das räumliche Vorstellungsvermögen. Es muss ein zweidimensionaler Grundriss im Geiste gefaltet werden, um auf diesem abgedruckte Buchstaben auf einer dreidimensionalen Figur wiederzufinden. Die Testpersonen haben vier Minuten Zeit zur Lösung der verschiedenen Aufgaben. Das Ergebnis des Untertests korreliert zu $r = .63$ mit dem Ge-

samtergebnis. Die Retestreliabilität ist befriedigend ($r = .70$). Dem Testautor zufolge sollen zukünftige Techniker und Architekten in diesem Untertest gute Ergebnisse erzielen, da das Vorstellen von Figuren und der Symbolvergleich von großer Bedeutung sind, wenn es um die Erfassung potenziellen technischen Talentes geht (Horn, 1983).

5.1.4.2 Intelligenz-Struktur-Test

Der Intelligenz-Struktur-Test-2000 (IST-2000) von Amthauer, Brocke, Liepmann und Beauducel (1999) ist ein Intelligenztest, der fünf der sieben Primärfaktoren nach Thurstone erfasst (*verbale, numerische und figurale Intelligenz, Merkfähigkeit* sowie *schlussfolgerndes Denken*). Das Grundmodul enthält neun Aufgabengruppen zu den inhaltsbezogenen Fähigkeiten (verbal, numerisch, figural) sowie zwei Aufgabengruppen (verbal, figural) zur Erfassung der Merkfähigkeit. Von den neun Aufgaben der Grundmodul-Kurzform werden jeweils drei Aufgabengruppen zu einer Skala zusammengefasst, woraus die drei Skalen zur Erfassung der verbalen, numerischen und figuralen Intelligenz resultieren. Insgesamt ergeben sich gute bis sehr gute Reliabilitätsschätzungen für die Skalen, insbesondere für die numerischen Fähigkeiten (Cronbachs Alpha .93 bzw. Split-Half-Reliabilität .95). Die Testinstruktion erfolgt mittels standardisierter Instruktionsanleitung und die Auswertung der Ergebnisse mit Schablonen, womit die Objektivität des Verfahrens garantiert werden soll. Die Anzahl der richtigen Aufgabebearbeitungen ergibt sich durch Addieren der einzelnen richtigen Lösungen. Dabei geht jede richtige Lösung mit dem gleichen Gewicht in den Summenwert ein. Die Summenwerte der Aufgabengruppen gehen in die Skalenwerte ein. Es gibt zudem altersspezifische Normen, die jeweils nach Gymnasiasten und Nicht-Gymnasiasten unterteilt sind. Gemäß der Zuordnung zu den einzelnen Normgruppen erfolgt eine Umsetzung der Rohwertsummen in Standardwerte bzw. (in diesem Falle) t -Werte für die jeweilige Aufgabengruppe. Es gibt zwei Testversionen A und B. In Form B wurden gegenüber Form A allerdings nur die Aufgabenfolge und teilweise die Reihenfolge der Distraktoren vertauscht. Es handelt sich somit um Pseudoparallelförmigkeiten mit der gleichen Aufgabensubstanz. Für beide Testformen können daher die gleichen Normwerttabellen verwendet werden. In der vorliegenden Untersuchung wurden jeweils die Testversionen der Form A verwendet.

Den Autoren zufolge kann zu Forschungszwecken bei bestimmten Fragestellungen der Einsatz einzelner Aufgabengruppen sinnvoll sein, z.B. bei der Erfassung von kognitiven Fähigkeiten, für die Standardinventare u.a. noch nicht verfügbar sind. Im Rahmen der Bewerbertestungen war v.a. die Erfassung der mathematisch-rechnerischen Fähigkeiten von besonderem Interesse. Aus diesen sowie Gründen der Ökonomie wurden lediglich die Aufgaben zu numerischen Intelligenz (Rechenaufgaben, Zahlenreihen sowie Rechenzeichen) in die Testbatterien mit aufgenommen. Die numerische Intelligenz erfasst zahlengebundene Intelligenzfähigkeiten sowie die Fähigkeit, logische Beziehungen zwischen Zahlen herzustellen. Um den sprachlichen Anteil bei den Anforderungen möglichst gering zu halten, werden die Aufgaben zur numerischen Intelligenz nicht verbal präsentiert. Im Rahmen der kaufmännischen Azubi-Testungen wurden die drei Aufgabengruppen zur Erfassung der numerischen Intelligenz in die Testbatterie aufgenommen. Da bei den technischen Bewerbern bereits der Untertest *Grundrisse berechnen* (AZUBI-TH; Görlich & Schuler, 2007) die Rechenfertigkeiten der Bewerber überprüft, wurde in die technische Testbatterie lediglich die *Aufgabengruppe Rechenaufgaben* mit übernommen. Aus diesem Grund geht zur besseren Vergleichbarkeit lediglich die Skala 4 (Rechenaufgaben, RA) des IST-2000 mit in die Berechnung des Intelligenzindikators mit ein.

Rechenaufgaben (RA). Diese Aufgabengruppe verlangt Rechenoperationen im Bereich der reellen Zahlen, mit denen die Grundrechenfertigkeiten einer Person erfasst werden können. Die Autoren ließen zur Berechnung der Skaleneigenschaften die erste Aufgabe weg, da bei dieser ein „Warming-up-Effekt“ vorliegt. Unter Ausschluss dieser Aufgabe weist die Skala

gute Reliabilitäten (Cronbachs Alpha .83 bzw. Split-Half .84) für die Gesamtnormgruppe auf. Die Probanden haben für die Lösung der 20 Aufgaben zehn Minuten Zeit.

5.2 Operationalisierung der Kriteriumsvariablen

Als zentrale Erfolgsgröße dienten in dieser Untersuchung die Schul- bzw. Berufsschulleistungen der getesteten Schüler bzw. Auszubildenden sowie eine Vorgesetztenbeurteilung der Ausbildungsleiter des ausbildenden Unternehmens, das als innerbetriebliches Erfolgsmaß angesehen werden kann. Die Erhebung der Schul- sowie Berufsschul- bzw. IHK-Noten sowie der Vorgesetztenbeurteilung ist in den folgenden Abschnitten (5.2.1 bis 5.2.3) nachzulesen.

5.2.1 Zeugnisnoten

Als Kriteriumsvariablen der Kontrollgruppe dienten die Abschlussnoten der Schüler. Diese lagen für die Gruppe der Gymnasiasten als durchschnittliche Abschlussnote vor. Für die Gruppe der Real- und Hauptschüler waren die Abschlusszeugnisse der Schüler inklusive der einzelnen Fachnoten verfügbar, aus denen eine durchschnittliche Abschlussnote ausgerechnet wurde. Hierbei gingen – im Gegensatz zum Abitur – alle Fächernoten zu gleichen Teilen in die Abschlussnote mit ein. Die Noten entsprechen der gängigen Bewertung in Deutschland mit 1 = „sehr gut“, 2 = „gut“, 3 = „befriedigend“, 4 = „ausreichend“ sowie 5 = „ungenügend“ und 6 = „mangelhaft“. Schüler, deren Abschlussnote schlechter als 4,00 war, haben die Abschlussprüfung nicht bestanden.

5.2.2 Berufsschul- und IHK-Noten

Die Berufsschulnoten der eingestellten Auszubildenden wurden im Dezember 2011 auf Nachfrage von dem ausbildenden Unternehmen zur Verfügung gestellt. Es handelte sich hierbei um jeweils eine Durchschnittsnote sowie den zugehörigen Einzelnoten aus verschiedenen Fächern. Für die Berechnungen in dieser Arbeit wurden die Noten der verschiedenen Einzelächer nicht verwendet, da diesbezüglich keine ausreichenden Zellbesetzungen erreicht werden konnten. Die Berufsschulnoten stellen somit immer eine Mittelung sämtlicher Berufsschulfächer dar.

Zudem wurden im Mai 2012 die Ergebnisse der IHK-Zwischenprüfungen der technischen Auszubildenden durch das ausbildende Unternehmen an Moldzio & Partner – Institut für Personalauswahl weitergeleitet. Die Noten der Zwischenprüfung beinhalteten die Ergebnisse der Prüfungen nach zwei bzw. eineinhalb Jahren Lehrzeit, die grundlegende Kenntnisse, Fertigkeiten und Fähigkeiten des Ausbildungsberufes schriftlich sowie praktisch überprüft. Die Ergebnisse der Zwischenprüfung bilden 40 % des Gesamtergebnisses der Abschlussprüfung und sind somit als guter Indikator der allgemeinen Ausbildungsleistung anzusehen.

Für die Bewertung der Prüfungsleistungen gilt in den deutschen Industrie- und Handelskammern der in Tabelle 5-2 erläuterte Bewertungsschlüssel. Die Zwischenprüfungsnoten lagen sowohl als Punktwerte als auch in Notenform vor. Für die Analysen wurden allerdings lediglich die Punktwerte verwendet, da diese eine höhere Variabilität aufwiesen als die gerundeten Noten.

Aufgrund der begrenzten Zeit dieser Studie lagen noch keine IHK-Abschlussnoten vor. Auch die Zwischenprüfungen im kaufmännischen Bereich wurden bis zum Abschluss der Datenerhebung noch nicht durchgeführt.

Tabelle 5-2: Notenspiegel/ Bewertungsschlüssel der IHK (Quelle: IHK-Ostwestfalen, 2012)

Punktzahl	Note	Erläuterung
100 – 92	1 = sehr gut	...eine den Anforderungen in besonderem Maße entsprechende Leistung.
unter 92 – 81	2 = gut	...eine den Anforderungen voll entsprechende Leistung.
unter 81 – 67	3 = befriedigend	...eine den Anforderungen im Allgemeinen entsprechende Leistung.
unter 67 – 50	4 = ausreichend	...eine Leistung, die zwar Mängel aufweist, aber im Ganzen den Anforderungen noch entspricht.
unter 50 – 30	5 = mangelhaft	...eine Leistung, die den Anforderungen nicht entspricht, jedoch erkennen lässt, dass die notwendigen Grundkenntnisse vorhanden sind.
unter 30 – 0	6 = ungenügend	...eine Leistung, die den Anforderungen nicht entspricht und bei der selbst die Grundkenntnisse lückenhaft sind.

5.2.3 Vorgesetztenbeurteilung

Zusätzlich zu den Leistungsdaten aus der Berufsschule wurde eine Vorgesetztenbeurteilung des allgemeinen Ausbildungserfolges der eingestellten kaufmännischen und technischen Auszubildenden eingeholt. Die Beurteilung der Auszubildenden erfolgte durch die Ausbildungsleiter des Unternehmens und orientierte sich an der in Tabelle 5-3 erläuterten Eingruppierung.

Tabelle 5-3: Bewertungsschlüssel der Auszubildenden

Ein-gruppierung	Erläuterung
A	Auszubildende in dieser Gruppe zeigen ausnahmslos sehr gute Leistungen, zeichnen sich durch eine außergewöhnlich rasche Auffassungsgabe, eine zielorientierte und strukturierte sowie eigenständige Arbeitsweise aus und erfüllen stets die an sie gestellten Anforderungen.
A-	Auszubildende in dieser Gruppe zeigen in der Regel sehr gute Leistungen, zeichnen sich durch eine stets rasche Auffassungsgabe, eine zielorientierte und strukturierte sowie vor allem eigenständige Arbeitsweise aus und erfüllen die an sie gestellten Anforderungen meist voll. Nur selten benötigen sie Hilfe oder liefern fehlerhafte Arbeitsergebnisse.
B+	Auszubildende in dieser Gruppe zeigen in der Regel gute bis sehr gute Leistungen. Sie zeichnen sich durch eine recht zielorientierte und strukturierte Arbeitsweise aus und erfüllen die an sie gestellten Anforderungen mit nur wenig Hilfestellung. Die benötigten Hilfen nehmen sie konstruktiv und dankbar an. Sie liefern solide und brauchbare Arbeitsergebnisse.
B	Auszubildende in dieser Gruppe zeigen in der Regel gute bis befriedigende Leistungen und zeichnen sich durch eine eher strukturierte Arbeitsweise aus. Die an sie gestellten Anforderungen erfüllen sie, benötigen allerdings verhältnismäßig oft Hilfestellungen, liefern danach jedoch solide Arbeitsergebnisse.
B-	Auszubildende in dieser Gruppe zeigen in der Regel ausreichende Leistungen. Sie haben jedoch manchmal Schwierigkeiten, die an sie gestellten Anforderungen umzusetzen oder zu verstehen. Routinearbeiten erledigen sie nach einer Zeit der Eingewöhnung eigenständig, haben aber Schwierigkeiten, sich an Neuerungen oder Ausnahmen zu gewöhnen. Die Arbeitsergebnisse müssen regelmäßig überprüft und Hilfestellung geleistet werden.
C	Auszubildende in dieser Gruppe zeigen in der Regel nur mangelhafte Leistungen und zeichnen sich durch eine langsame Auffassungsgabe sowie eine eher chaotische Arbeitsweise aus. Sie erfüllen die an sie gestellten Anforderungen nur selten.

5.3 Erhebung weiterer Variablen

Im Rahmen der beiden Untersuchungen galt es, über die unabhängigen und abhängigen Variablen hinaus einige Kontrollvariablen zu erheben, um deren potenziellen Einfluss auf die Ergebnisse der Studie erfassen zu können. Zum einen wurden einige demografische Merkmale der Bewerber bzw. Schüler (Abschnitt 5.3.1) erhoben, wie beispielsweise Alter, Geschlecht und Schulbildung. Zum anderen sollte der mögliche Einfluss potenziell konfundierender Variablen (Abschnitt 5.3.2) wie beispielsweise sozial erwünschten Antwortverhaltens erfasst werden.

5.3.1 Demografische Merkmale der Untersuchungsteilnehmer

Es wurden in beiden Stichproben einige demografische Merkmale der Teilnehmer erfasst bzw. abgefragt. Dies geschah einerseits, um eventuelle Gruppenunterschiede und moderierende Einflüsse aufzudecken und andererseits, um die Stichproben adäquat beschreiben zu können und somit auch eine Vergleichbarkeit der Daten mit beispielsweise anderen Studienergebnissen oder Probandengruppen herzustellen.

Zu den erhobenen demografischen Merkmalen gehörten *Alter* (in Jahren zum Testzeitpunkt), *Geschlecht* und *Schulbildung*.

Im Falle der Ausbildungsbewerber wurden bei T1 das Alter sowie Geschlecht der Testpersonen anhand von Informationen des Auftraggebers ermittelt. Die Art der Schulbildung konnte jeweils dem Abschluss- bzw. aktuellen Zwischenzeugnis, welches der Testbatterie vom Auftraggeber jeweils beigelegt wurde, entnommen werden. Es wurde in der Bewerberstichprobe zwischen *Haupt-*, *Real-* und *Fachhochschülern* sowie *Gymnasiasten* unterschieden. Im Rahmen der Nachbefragung (T2) wurden die Auszubildenden noch einmal gebeten, Angaben zu ihrem bisher höchsten Bildungsabschluss (*Haupt-*, *Real-*, *Fachhochschulabschluss*, *Abitur* oder ggf. einem *zuvor erlernten Beruf*) zu machen.

In der Kontrollgruppe gaben die Testpersonen ihr Alter bzw. Geburtsdatum sowie Geschlecht jeweils am Tag der Testung selbst an. Die Art der besuchten Schulart war vorab bekannt, da jeweils klassenweise getestet wurde und diese Variable durch die Versuchsleiter einfach ergänzt werden konnte. Es gab demzufolge *Haupt-* und *Realschüler* sowie *Gymnasiasten* in der Kontrollgruppe.

5.3.2 Kontrollvariablen

Im Rahmen der Untersuchung galt es, zusätzlich einige Kontrollvariablen zu erheben, die einen potenziellen Einfluss auf die Ergebnisse haben könnten. Hierzu wurde zum einen das sozial erwünschte Antwortverhalten untersucht (Abschnitt 5.3.2.1). Zum anderen wurde erfragt, inwiefern die Teilnehmer der Studie schon über berufliche Vorerfahrungen verfügten (Abschnitt 5.3.2.2). Abschließend wird kurz auf weitere, potenzielle Einflussvariablen eingegangen (Abschnitt 5.3.2.3).

5.3.2.1 Sozial erwünschtes Antwortverhalten

Sozial erwünschtes Antwortverhalten wurde mit Hilfe der deutschen Version der Sozialen-Erwünschtheits-Skala-17 (SES-17, Stöber, 1999) erfasst. Diese ist im Vergleich zu anderen Skalen ein ökonomisches und valides Instrument zur Erfassung von Tendenzen zur positiven Selbstdarstellung. So ist die interne Konsistenz der Skala zufriedenstellend (Cronbachs $\alpha = .72$ bis $.75$) und die Retest-Stabilität über einen Zeitraum von vier Wochen beträgt $r = .82$. Mit anderen Skalen zur Erfassung sozialer Erwünschtheit weist die SES-17 zufriedenstellende Zusammenhänge auf ($r = .74$ bis $.76$). Darüber hinaus hat die SES-17 eine höhere Augen-

scheinvalidität und soziale Akzeptanz als bisherige, veraltete Verfahren, da bei der Skalenkonstruktion insbesondere auf eine zeitgemäße Formulierung der Items geachtet wurde. Stöber (1999) empfiehlt daher den Einsatz der SES-17 insbesondere für den Einsatz in jüngeren Stichproben.

Die SES-17 besteht aus 17 Items, die häufiges, aber sozial unerwünschtes Verhalten sowie seltenes, aber sozial erwünschtes Verhalten beschreiben. Sieben Items sind negativ gepolt. Ein Beispielitem lautet: „*Im Straßenverkehr nehme ich stets Rücksicht auf die anderen Verkehrsteilnehmer.*“. Als Antwortformat dient eine bipolare Skala mit den Polen 1= „*richtig*“ sowie 0= „*falsch*“. Item 4 wurde, wie von Stöber (2001) empfohlen, aufgrund schlechter Itemkennwerte von den Analysen ausgeschlossen, so dass nur 16 Items der SES-17 in der Untersuchung verwendet wurden.

5.3.2.2 Berufliche Vorerfahrungen

Um eventuelle Einflüsse beruflicher Vorerfahrung auf die Beantwortung der Items identifizieren und statistisch kontrollieren zu können, galt es, einige weitere Kontrollvariablen zu erheben – sofern dies mit den Rahmenbedingungen der Untersuchungen vereinbar war.

Im Rahmen der Nachbefragung der Ausbildungsbewerber bzw. der Kontrollgruppentestung wurden die Probanden zu ihren *beruflichen Vorerfahrungen* befragt. Dies geschah jeweils mit zwei vergleichbaren Items. Die Probanden wurden gefragt, ob sie neben der Schule bzw. vor der Ausbildung schon einmal berufliche Erfahrungen sammeln konnten, beispielsweise in Form von Ferienjobs, Praktika oder Aushilfsarbeiten. Zu dieser Frage gab es ein bipolares Antwortformat mit 0= „*nein*“ und 1= „*ja*“. Im Falle einer positiven Antwort wurden die Probanden gebeten, ihre beruflichen Vorerfahrungen inhaltlich zu spezifizieren („*ja, und zwar als...*“). Dieses Vorgehen wurde zuvor auch schon von anderen Autoren als sinnvoll erachtet und vorgenommen (z.B. Lievens et al., 2008).

Zudem wurde die Kontrollgruppe nach ihren Ideen zu ihrem späteren Berufsleben befragt („*Haben Sie schon eine Idee, was Sie nach der Schule später einmal beruflich machen möchten?*“). Auch hier gab es wiederum ein bipolares Antwortformat (0= „*nein*“ und 1= „*ja*“). Im Falle einer positiven Antwort wurden die Testpersonen zudem gebeten, ihren Berufswunsch inhaltlich zu spezifizieren (freies Antwortformat).

Die Auszubildenden wurden an dieser Stelle gebeten, ihren *Ausbildungsberuf* zu benennen (freies Antwortformat). Diese Angabe wurde zudem durch eine Information seitens des ausbildenden Betriebes verifiziert bzw. im Falle fehlender Angaben ergänzt.

5.3.2.3 Weitere Einflüsse

Neben den oben genannten potenziellen Einflüssen wurden alle Fragebogenpakete zusätzlich per Augenschein auf *stereotype Antworttendenzen* hin überprüft.

Zudem sollte eine Unterteilung der Auszubildenden in ihre *Ausbildungsrichtungen* (*technisch* vs. *kaufmännisch*) aufklären, inwiefern Unterschiede in den Daten bezüglich der Ausbildungsberufe der Testpersonen vorliegen.

Mögliche *Reihenfolgeeffekte bei der Testgabe* konnten aufgrund der praktischen Relevanz der Testergebnisse für die ausbildenden Betriebe nicht kontrolliert werden. Die Reihenfolge der Tests in den Testbatterien war je nach Ausbildungsrichtung vorgegeben und nicht variabel (s. Abschnitt 5.4.1). Generell ist es möglich, dass vorausgehende Items einen Einfluss auf die Beantwortung von später präsentierten Items haben. Dies lässt sich dadurch erklären, dass sich die Testperson implizit oder explizit noch mit bestimmten Aspekten der zuvor abgefragten Inhalte beschäftigt und dies die Beantwortung der späteren Items beeinflusst (Sudman,

Bradburn & Schwarz, 1996). Die Reihenfolgen in den Testbatterien der Nachbefragungen bzw. Kontrollgruppenuntersuchungen wurden ebenfalls nicht variiert. Vergleichbare Voruntersuchungen haben allerdings keine Reihenfolgeeffekte bezüglich der Testgabe feststellen können (Dreier, 2008; Korinteli, 2009; Struckmeier, 2009), sodass auf eine Variation der Testreihenfolgen in diesem Fall verzichtet wurde.

6. Statistische Auswertung der Daten

In diesem Kapitel wird dargestellt, welche statistischen Analysen zur Untersuchung der Fragestellungen herangezogen werden und welche Voraussetzungen bzw. Gütekriterien diesen Verfahren zugrunde liegen. Zuvor werden jeweils die verschiedenen methodischen Schritte dargestellt, die zur Aufbereitung des Datensatzes und Analyse der Messgüte dienen.

Der erste Teil (Abschnitt 6.1) behandelt die Analyse und Behandlung fehlender Werte. Es wird geschildert, was die Ursachen fehlender Werte sein können, ab wann man von einer hohen Fehlendrate sprechen kann und welche Methoden zur Ersetzung entsprechender Missings angewendet werden.

Im darauf folgenden Abschnitt (6.2) werden die verwendeten Methoden zur Überprüfung der Skaleneigenschaften der Arbeitsbezogenen Gewissenhaftigkeitsskalen beschrieben, die notwendig sind, um zu überprüfen, inwiefern es sich bei dem neu entwickelten Verfahren um ein zuverlässiges und valides Verfahren zur Erfassung der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung handelt.

Abschnitt 6.3 beinhaltet die Erläuterung der Methoden, die zur Überprüfung der Vorhersage der Kriteriumsvariablen angewendet werden.

Der letzte Abschnitt (6.4) widmet sich dem Vergleich der prädiktiven Validität der Gewissenhaftigkeitsaspekte Fleiß und Ordnung in Bezug zur Gewissenhaftigkeitsdimension des NEO-FFI (Borkenau & Ostendorf, 1993, 2008). Hierzu soll die inkrementelle Validität der AGS ermittelt werden.

6.1 Analyse und Behandlung fehlender Werte

Wie in fast jeder Untersuchung haben auch in der vorliegenden Studie nicht alle Probanden die Testverfahren bzw. Selbstbeschreibungsinventare (z.B. AGS, NEO-FFI, Selbstwirksamkeitsskala) vollständig bearbeitet, weshalb es zu fehlenden Werten gekommen ist. Es besteht die Gefahr, dass es bei mangelnder Berücksichtigung fehlender Werte v.a. auf Ebene der Prädiktoren in den Datenanalysen zu Fehlschlüssen bzw. -entscheidungen kommen kann. Da in dieser Untersuchung darüber hinaus ein besonderes Augenmerk auf der Überprüfung der Arbeitsbezogenen Gewissenhaftigkeitsskalen als Prädiktor liegt, wurde dieser Problematik vor der Durchführung der Analysen besondere Aufmerksamkeit geschenkt.

Es gibt zwei verschiedene Arten von fehlenden Werten, die in den vorliegenden Daten aufzufinden sind. Vor allem handelt es sich um „*Item-Nonresponse*“-Fehler, welche unvollständige Datensätze aufgrund fehlender Beantwortung verschiedener Items verursachen. Darüber hinaus haben nicht alle befragten Auszubildenden in der Nachbefragung die zur korrekten Zuordnung notwendigen persönlichen bzw. demografischen Informationen mitgeteilt, sodass deren Testergebnisse aufgrund der fehlenden Angaben nicht mit den Daten der ersten Erhebungswelle gepaart werden konnten. Diese Datensätze bzw. Personen werden entsprechend als *Drop-Out*-Daten bezeichnet, was einen Abbruch der Mitarbeit in einer längsschnittlichen Untersuchung beschreibt und auch als *Panelmortalität* bezeichnet wird (Göthlich, 2009). Für die Kontrollgruppe der Schüler ist dies der Fall, wenn für die befragten Testpersonen kein Abschlusszeugnis verfügbar war. Im Folgenden wird dargestellt, wie in dieser Studie methodisch mit den fehlenden Werten umgegangen wird.

6.1.1 Analyse der Fehlendmechanismen

Die Gründe für das Fehlen von Daten bzw. Antworten auf Items können verschiedene Ursachen haben. Wichtiger als die fehlenden Daten selbst sind jedoch zuerst mögliche Mechanismen oder Muster zu identifizieren, die den nicht-beantworteten Items zugrunde liegen können

(Tabachnick & Fidell, 2001). Man kann drei bzw. vier verschiedene Arten von Fehlendmechanismen unterscheiden.

Missing Completely At Random (MCAR) beschreibt einen zufälligen Datenausfall, zu dem es keine Informationen gibt, warum Werte fehlen (Göthlich, 2009). Eine Regelmäßigkeit im Sinne eines Zusammenhangs zwischen den in der Untersuchung erhobenen Variablen und dem Muster der fehlenden Werte kann nicht gefunden werden. Da die Werte zufällig fehlen, wird dieser Mechanismus auch als „ignorierbar“ bezeichnet und beschreibt den günstigsten Fall fehlender Daten (Tabachnik & Fidell, 2001).

Missing At Random (MAR) bezeichnet ebenfalls ein zufälliges und zu ignorierendes Fehlen von Werten. Dieser unterscheidet sich vom MCAR-Mechanismus in der Form, dass zwar Informationen darüber vorliegen, warum Items nicht beantwortet wurden, diese aber trotzdem unabhängig vom Ergebnis sind (Göthlich, 2009). Es handelt sich somit um einen bedingt zufälligen Ausfall von Daten, auf den man sich auch beziehen kann, wenn in der Analyse der MCAR-Test zwar signifikant wird, ein Zusammenhang zu den erhobenen Variablen aber dennoch ausgeschlossen werden kann (Tabachnik & Fidell, 2001).

Missing Not At Random (MNAR) beschreibt einen nicht zufälligen Datenausfall, der nachweislich das Ergebnis der Untersuchung beeinflusst bzw. verzerrt. In diesem Fall hängt der Fehlendmechanismus mit den untersuchten Variablen zusammen. Das zugrundeliegende Antwort- bzw. Nicht-Antwortmuster sollte erfasst und in die Analysen als beeinflussende Variable mit einbezogen werden (Göthlich, 2009).

Observed At Random (OAR) wird teilweise in der Literatur als weiterer Fehlendmechanismus genannt. In diesem Fall ist die Beantwortung eines Items unabhängig von der Ausprägung einer anderen Variablen (Göthlich, 2009). Gemeinsam mit dem MAR-Mechanismus ergibt dieses Fehlermuster den MCAR-Mechanismus.

Mit Hilfe des MCAR-Tests in der SPSS-Software können Fälle mit Fehlendmustern identifiziert werden (Tabachnick & Fidell, 2001). In dieser Untersuchung werden alle Selbsteinschätzungsdaten dieser Analyse unterzogen.

6.1.2 Elimination von Fällen mit zu hohem Anteil fehlender Werte

Ein zu hoher Anteil fehlender Werte in den Daten kann ein Hinweis auf fehlendes Itemverständnis oder mangelnde Teilnahmebereitschaft sein. Datensätze mit einem hohen Anteil fehlender Werte können demzufolge stark durch Messfehler verzerrt sein, weshalb diese nicht mit in die Analysen eingehen sollten.

In der Literatur gibt es bisher noch keine eindeutige Regel, ab wie viel Prozent man von einem „hohen“ Anteil fehlender Werte sprechen kann. Als Faustregeln wird oftmals eine Grenze von 5 % für große Stichproben angeführt (Tabachnik & Fidell, 2001). Ein Ersatz fehlender Werte hat den Autorinnen zufolge bis zu diesem Prozentsatz keinen oder nur einen geringen Einfluss auf die Ergebnisse einer Studie, egal welche Methode man hierzu verwendet. Göthlich (2009) zufolge seien 1 % bis 10 % fehlende Werte (ohne den Einbezug von Drop-Out-Werten) normal. SPSS verwendet wiederum die 5 %-Grenze als Basis für die Analyse fehlender Werte. Aus diesen Gründen wird in der vorliegenden Untersuchung die Obergrenze für fehlende Werte auf 5 % festgesetzt. Hat eine Person mehr als 5 % der Items nicht beantwortet, wird sie aus dem Datensatz gelöscht.

6.1.3 Mittelwertbasierte Ersetzung fehlender Werte

Eine einfache Methode, Datensätze mit fehlenden Werten zu behandeln, stellte eine komplette Löschung der entsprechenden Fälle dar. Da ein solches Vorgehen allerdings den Anteil der in

der Stichprobe verbleibenden, kompletten Datensätze zu stark verringern würde und zusätzlich zu diesem starken Informationsverlust auch eventuell zu Ergebnisverzerrungen führen könnte, wird in dieser Untersuchung von diesem Vorgehen Abstand genommen.

Stattdessen werden die fehlenden Werte fallweise durch den Mittelwert des jeweiligen Indikators ersetzt. Diese Methode stellt ein konventionelles bzw. Ad-hoc-Imputationsverfahren dar (Bühner, 2011) und ist für die Zwecke dieser Untersuchung das effektivste Vorgehen zur Ergänzung fehlender Werte. Auf den Einsatz aufwendiger multipler Imputationsverfahren wurde verzichtet, da die Überlegenheit dieser Methoden gegenüber den herkömmlichen Imputationsverfahren noch nicht ausreichend belegt ist und hiermit unter Umständen eine Genauigkeit suggeriert wird, die es in den Daten gar nicht gibt und eventuell andere Probleme in den Daten übersehen werden könnten (Göthlich, 2009). Eine dadurch entstehende, leichte Varianzeinschränkung wird hierbei in Kauf genommen. Die Mittelwertimputation wird bei denjenigen Fällen angewendet, deren Anteil fehlender Werte sich unterhalb der 5 %-Grenze befindet.

6.2 Überprüfung der Skaleneigenschaften der AGS

Bevor die Zusammenhänge der Arbeitsbezogenen Gewissenhaftigkeitsskalen Fleiß und Ordnung mit den Leistungsvariablen sowie die Überprüfung der Modellannahmen und der Vergleich mit dem NEO-FFI (Borkenau & Ostendorf, 2008) vorgenommen werden können, soll zuerst die Messgüte der AGS bestimmt werden, um eine Interpretierbarkeit der Ergebnisse bzw. Zusammenhänge sicher zu stellen. Hierzu werden verschiedene Analysen zur Überprüfung der Itemgüte (6.2.1), der Skalenhomogenität (6.2.2) sowie der Konstruktvalidität der AGS (6.2.3) vorgenommen.

6.2.1 Analysen auf Itemebene

Zu Beginn der Überprüfung der AGS werden die Eigenschaften der einzelnen Items dargestellt. Zuerst sollen allgemeine deskriptiven Daten (6.2.1.1) dargestellt werden. Danach soll überprüft werden, inwiefern die AGS nach Kriterien der Itemschwierigkeit (6.2.1.2) und der Trennschärfe (6.2.1.3) die zwei Faktoren Fleiß und Ordnung abbilden.

6.2.1.1 Deskriptiva

Zur objektiven Beschreibung und als Basis für die Einschätzung einer Skala gehören neben den Skalenmittelwerten und zugehörigen Standardabweichungen auch die entsprechenden Itemkennwerte, d.h. *Itemmittelwerte* und *-standardabweichungen* sowie *Minimum* und *Maximum* der Testwerte.

Da viele statistische Verfahren eine Normalverteilung der Testwerte voraussetzen, soll an dieser Stelle auch die univariate Normalverteilung der erhobenen Gewissenhaftigkeitsindikatoren überprüft werden. Eine Veranschaulichung der Antworthäufigkeiten und Indikator für die Überprüfung der Normalverteilung der Daten bietet die Darstellung eines Histogramms mit Normalverteilungskurve. Darüber hinaus soll die Verteilung der Werte mit dem *Kolmogorov-Smirnov-Test* (s. Abschnitt 6.5.3) überprüft werden, auch wenn es sich bei diesem Verfahren um ein wenig trennscharfes Verfahren zur Überprüfung der eindimensionalen Normalverteilung handelt (Rasch & Kubinger, 2006).

Darüber hinaus kann die *Schiefe* der Rohwertverteilung (v) Hinweise zur Verteilung der Kennwerte in den Stichproben liefern. Die Schiefe ist eine Bezeichnung für die Neigung einer statistischen Verteilung. Sie ist also ein Maß für die Asymmetrie einer Verteilung. Da die Gauß'sche Normalverteilung symmetrisch ist ($v = 0$), ist die Schiefe ein Indikator, um die Normalverteilung der Daten zu überprüfen. Positive Werte ($v > 0$) geben eine rechtsschiefe

Verteilung an, während negative Werte ($v < 0$) auf eine linksschiefe Verschiebung hindeuten. Lienert und Raatz (1994) zufolge kann bei großem N (über 400) der Bereich zwischen $v = -0,50$ und $v = +0,50$ als praktisch insignifikant gelten und die Rohwerteverteilung dennoch als normal angesehen werden.

Auch die *Kurtosis* bzw. *Wölbung* der Verteilung kann Hinweise auf die Normalverteilung der Daten liefern. Die Kurtosis ist ein Maß für die Steilheit einer eingipfligen Häufigkeitsverteilung. Werte in Verteilungen mit geringer Wölbung streuen gleichmäßig, während bei Verteilungen mit hoher Kurtosis die Streuung aus extremen Ereignissen resultiert. Zur Bestimmung der Normalverteilung berechnet man den *Exzess* (E), der die Differenz der Wölbung zur Kurtosis der Dichtefunktion einer normalverteilten Zufallsgröße angibt. Der Exzess einer Normalverteilung beträgt $E = +3,0$. Bei großen Stichproben können vom Vorliegen einer Normalverteilung der Daten ausgehen, wenn der Exzess zwischen $+2,00$ und $+4,00$ liegt (Lienert & Raatz, 1994).

6.2.1.2 Itemschwierigkeiten

Die Itemschwierigkeit wird auch als *psychometrische* oder *statistische Itemschwierigkeit* bzw. *Popularität* (P) bezeichnet. Die empirische Definition von Schwierigkeit widerspricht dem Alltagsverständnis einer hohen Schwierigkeit und ist definiert durch die prozentuale Häufigkeit, mit der ein Item von einer Stichprobe richtig beantwortet wird. Diese Definition lässt sich auch auf Fragebogenmaße übertragen, bei denen es zwar keine "richtigen" oder "falschen" Antworten, sondern nur Zustimmung und Ablehnung gibt, sogenannte *Schlüsselantworten* (Lienert & Raatz, 1994). Demzufolge gilt für Persönlichkeitstests ein Item, dem viele Personen zugestimmt haben bzw. dieses in Schlüsselrichtung beantwortet haben, als psychometrisch leichtes Item. Ein Item, dem nur wenige Personen zugestimmt haben hingegen als psychometrisch schwieriges Item.

Die Werte der Itemschwierigkeit P gehen von 0 bis 1. Von einer niedrigen Itemschwierigkeit spricht man bei Werten über $.80$, von einer hohen bei Werten unter $.20$ (Bühner, 2011; Lienert & Raatz, 1994). Ziel bei einer Fragebogenkonstruktion ist eine mittlere Itemschwierigkeit, d.h. ein Wert um $.50$, da dies die beste Voraussetzung für eine hohe Trennschärfe (s. Abschnitt 6.1.2.3) ist. Lienert und Raatz (1994) räumen allerdings ein, dass ein einzelnes Item unter Umständen auch einmal sehr leicht, d.h. von vielen beantwortet, oder sehr schwer, d.h. von wenigen beantwortet, sein kann, sofern der Test selbst insgesamt eine mittlere Schwierigkeit aufweist. Darüber hinaus sollte eine eher breite Streuung der Schwierigkeitsindizes angestrebt werden, um in allen Bereichen des gemessenen Eigenschaftsbereichs Personen differenzieren zu können (Bühner, 2011).

Zur Bestimmung der Itemschwierigkeit wird häufig der Itemmittelwert als Indikator herangezogen. Um eine Vergleichbarkeit der Items, auch mit anderen Fragebogenmaßen, herzustellen, kann man jedoch auch die Itemschwierigkeit P_i eines Items berechnen. Diese kann man zwar nicht direkt mit SPSS ermitteln, sie lässt sich allerdings relativ leicht „per Hand“ mit folgender Formel berechnen:

$$P_i = \frac{M_i}{\max(i)}$$

Zur Ermittlung der Itemschwierigkeit P_i muss der Itemmittelwert M_i durch den maximal erreichbaren Wert in Item i $\max(i)$ dividiert werden. Die Codierung der Antwortstufen der AGS geht jeweils von 0 bis 4, weshalb die Itemmittelwerte jeweils durch $\max(i) = 4$ geteilt werden.

6.2.1.3 Trennschärfen

Die Trennschärfe eines Items ist ein Kennwert dafür, wie viel ein Item zum Gesamtskalenwert beiträgt, d.h. wie gut ein Testergebnis aufgrund der Beantwortung eines Items vorherge-

sagt werden kann. Eine hohe Trennschärfe drückt aus, dass ein Item die Eigenschaften der getesteten Personen exakt erfasst, d.h. in diesem Falle beispielsweise Testpersonen mit hohen Werten in den AGS sich auch als ordentlicher bzw. fleißiger als Personen mit niedrigen Werten beschreiben.

Die Trennschärfe wird durch den *Trennschärfekoeffizienten* r_{it} erfasst. Es wird die Korrelation der Messwerte eines Items mit dem Summenwert der restlichen Items einer Skala ermittelt. Man kann hierzu entweder die Korrelation zwischen jedem Item und dem Summenwert der restlichen Items der Skala berechnen oder aus der Korrelation des Items mit dem gesamten Summenwert das jeweilige zu korrelierende Item herausrechnen. Letzteres stellt die Formel zur Berechnung der part-whole-korrigierten Trennschärfe eines Items dar, die berechnet wird, wenn es sich um einen intervallskalierten und normalverteilten Testwert handelt (Bühner, 2011).

Der Wertebereich der korrigierten Trennschärfe liegt zwischen 0 und 1. Man spricht von einer niedrigen Trennschärfe, wenn $r_{it} < .30$ ist. In diesem Falle wird das Item sowohl von Probanden mit niedrigen als auch mit hohen Merkmalsausprägungen in ähnlicher Weise beantwortet. Solche Items gelten allgemein als unbrauchbar. Von einer hohen Trennschärfe spricht man, wenn $r_{it} > .50$ ist. Demzufolge kennzeichnen Werte zwischen $r_{it} = .30$ bis $r_{it} = .50$ eine mittlere Trennschärfe.

Die Trennschärfe der Testitems ist Lienert und Raatz (1994) zufolge allerdings abhängig von der jeweiligen Itemschwierigkeit (s. 6.2.1.3), der Homogenität bzw. Dimensionalität der Skala (s. 6.2.2.3), der Stellung eines Items innerhalb des Testverfahrens sowie der Reliabilität (s. 6.2.2.1). Auch die Itemverteilungen können sich auf die Höhe der Itemtrennschärfen auswirken. Unterscheidet sich die Schiefe eines Items deutlich von der Schiefe des korrigierten Skalenwertes, ist die Chance, dass dieses Item sowohl mit Items anderer Schiefe als auch mit dem korrigierten Skalenwert korreliert, geringer als unter optimalen Bedingungen bei Normalverteilung oder gleicher Schiefe. Es ist allerdings aus Gründen der Skalenkonstruktion notwendig, dass eine Skala Items mit unterschiedlich schiefen Verteilungen enthält. Daher ist bei Itemselektionsvorgängen darauf zu achten, dass Items nicht nur aufgrund ihrer geringen Trennschärfe aus einem Test entfernt werden, sondern auch die Itemschwierigkeit, -standardabweichung und insbesondere die Verteilung der Items berücksichtigt wird. Darüber hinaus sind in homogenen Stichproben keine so hohen Trennschärfen zu erwarten wie in heterogenen.

6.2.2 Messgüte der Gewissenhaftigkeitsaspekte Fleiß und Ordnung

Nachdem die Eigenschaften der einzelnen Items der AGS ermittelt wurden, soll überprüft werden, ob es sich bei den beiden Gewissenhaftigkeitsaspekten Fleiß und Ordnung um homogene und stabile Skalen handelt. Hierzu sollen die Reliabilität (6.2.2.1) sowie die Retest-Reliabilität (6.2.2.2) berechnet werden. Außerdem soll mittels konfirmatorischer Faktorenanalyse (6.2.2.3) überprüft werden, ob das theoretisch formulierte Faktorenmodell in den beiden Datensätzen wiedergefunden werden kann. Die hierzu notwendigen Analysen werden in den folgenden drei Abschnitten erläutert.

6.2.2.1 Reliabilitätsschätzungen der Subskalen

Die Messgenauigkeit oder auch *Reliabilität* ist ein wichtiges Gütekriterium einer Messung bzw. eines Tests. Unter Reliabilität versteht man den „Grad der Genauigkeit, mit dem ein Test ein bestimmtes Merkmal misst, unabhängig davon, was er zu messen beansprucht“ (Bühner, 2011, S. 60). Man kann auch von der *Replizierbarkeit* eines Messergebnisses sprechen (Rammstedt, 2004).

Um eine Aussage über die Reliabilität zu erhalten, die unabhängig von der Einheit der Skala ist, wird ein Varianzverhältnis gebildet. Die Reliabilität eines Messwerts in der Population (ρ_{tt}) entspricht dem Verhältnis der Varianz der wahren Werte (σ^2_T) zur Varianz der beobachteten Werte (σ^2_X). Diese wird auch als *Determinationskoeffizient* bzw. gemeinsame Varianz der wahren und beobachteten Werte in der Population bezeichnet (ρ^2_{XT}).

$$\rho_{tt} = \frac{\sigma^2_T}{\sigma^2_X} = \rho^2_{XT}$$

Da man die Varianz der wahren Werte allerdings nicht kennt, muss man zur Bestimmung der Reliabilität eines Tests bzw. einer Skala verschiedene Schätzmethoden verwenden. Insbesondere für die Persönlichkeitsforschung bieten sich hierzu Konsistenzmethoden an, die auf Korrelationen oder Kovarianzen der Items basieren. Man spricht dann von *interner* oder *innerer Konsistenz* (Bühner, 2011). Zur Bestimmung der internen Konsistenz wird ein Test einer Stichprobe einmalig dargeboten. Methodisch wird er dann in so viele „Untertests“ zerlegt, wie er Items besitzt. Es erfolgt eine statistische Überprüfung des inneren Zusammenhangs der Items unter Berücksichtigung der Testlänge. Man interpretiert den Konsistenzkoeffizienten also als die durchschnittliche Korrelation zwischen allen möglichen Testhälften. Konsistenzkoeffizienten liefern eine genaue Schätzung der Reliabilität. Als Standard und relativ universell anwendbare Methode zur Schätzung der inneren Konsistenz gilt *Cronbachs Alpha* (α ; Cronbach, 1951), welches in dieser Studie verwendet werden soll.

Die Reliabilität ist immer ein Wert, der zwischen 0 und 1 liegt. Ein Wert von 1 bedeutet, dass eine Messung perfekt genau ist, ein Wert von null bedeutet, dass eine Messung der zugrunde liegenden Variablen nicht vorliegt. Eine Reliabilität von .50 bedeutet, dass systematische Varianz und Messfehlervarianz die Unterschiede zwischen Personen in diesem Messwert zu gleichen Anteilen bestimmen. D.h. die Varianz eines Testwerts setzt sich zu 50 % aus Messfehlern oder unsystematischen Einflüssen und zu 50 % aus durch den Prädiktor erklärter Varianz zusammen (Bühner, 2011). Eine zuverlässige Interpretation der Ergebnisse bzw. Testwerte ist in diesem Falle nur bedingt möglich, da keine „Messung“ im eigentlichen Sinne vorliegt. Im Idealfall ist die Verteilung der intraindividuellen Messwerte sehr schmal, d.h. nur wenige Werte weichen vom wahren Wert einer Person ab. Die Reliabilität geht dann gegen 1.

Hohe Reliabilitäten sind für die Individualdiagnostik besonders wichtig, da aus einem individuellen Ergebnis die wahre Merkmalsausprägung abzuschätzen versucht wird (Görlich & Schuler, 2007). Als Richtwerte zur Bewertung der Güte eines Tests bezeichnet Bühner (2011) Werte von $\alpha < .80$ als niedrige, $.80 < \alpha < .90$ als mittlere und $\alpha > .90$ als hohe Reliabilität. Die Orientierungswerte der DIN 33430 geben vor, dass eine Reliabilität mindestens zwischen $\alpha = .70$ und $.85$ liegen sollte (DIN, 2002).

6.2.2.2 Retest-Reliabilität

Neben der internen Konsistenz interessierte auch die Stabilität der Merkmalsausprägungen der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie – zu Vergleichszwecken – der Selbstwirksamkeitsskala, weshalb im Rahmen der Nachbefragung die eingestellten Ausbildungsbewerber die AGS und die Selbstwirksamkeitsskala erneut bearbeiteten, um hieraus die sog. Retest-Reliabilität zu errechnen. Bei der Retest-Methode handelt es sich um die wiederholte Darbietung ein und desselben Tests an ein und derselben Stichprobe in einem angemessenen Zeitabstand (Bühner, 2011, S. 159). Durch die Retest-Methode werden der Messfehler sowie der Einfluss weiterer Zufallsvariablen auf die Merkmalsausprägung geschätzt, die sich zwischen zwei Messungen über die Zeit ergeben. Die Schätzung der Retest-Reliabilität erfolgt durch die Korrelation der Rohwertpaare zwischen der ersten Testung (T1) und der Testwiederholung mit demselben Test zu einem zweiten Testzeitpunkt (T2).

$$Rel_{T1T2} = r(X_{T1}, X_{T2})$$

Vorausgesetzt ein Merkmal hat sich während der Zeit zwischen T1 und T2 nicht verändert, sollte die Beantwortung der Testitems nach dem Zeitintervall in derselben Weise erfolgen. Lienert und Raatz (1994) zufolge sind Persönlichkeitsitems, die stark aktualitätsbedingtes Verhalten erfassen, weniger reliabel als solche, die allgemeine Verhaltenstendenzen erfassen. Schwankungen zwischen den Werten schreiben die Autoren Erinnerungsmodifikationen zu, die zu einer Reduzierung der Reliabilität führen können. Daher sollte für die Testwiederholung ein Zeitpunkt gewählt werden, zu dem einerseits die Erinnerungs- und/ oder Übungseffekte möglichst keinen Einfluss (mehr) haben, andererseits aber die Merkmalskonstanz noch als gegeben angenommen werden kann. Zu einer Scheinreliabilität kommt es, wenn das Zeitintervall zu kurz gewählt wird, beispielsweise durch Memorieren der Antworten durch die Testperson.

Fällt das Zeitintervall zwischen den Testungen allerdings zu groß aus, kann die Retest-Reliabilität den Zusammenhang zwischen T1 und T2 jedoch auch unterschätzen, z.B. wenn unterschiedliche personenspezifische Veränderungen eintreten, die die Rangreihe unabhängig vom gemessenen Merkmal verändern.

In diesem Sinne macht die Retest-Reliabilität Aussagen über die *zeitliche Stabilität* eines untersuchten Merkmals und wird auch als *Stabilitätskoeffizient* bezeichnet. Bei einer hohen Retest-Reliabilität kann man daher von einer zeitlichen Stabilität bzw. Merkmalskonstanz der untersuchten Eigenschaft ausgehen. Der umgekehrte Schluss ist so jedoch nicht gültig. Eine niedrige Retestreliabilität kann sich trotz einer hohen zeitlichen Stabilität auch dann ergeben, wenn der Test in sich nicht konsistent ist (Lienert & Raatz, 1994).

Ein hoher Retest-Koeffizient mit einem langen Zeitintervall zwischen T1 und T2 spricht also stets für eine hohe *Merkmalskonstanz*. Gibt eine kurzfristige Wiederholung einen hohen, eine lange einen niedrigen Reliabilitätskoeffizienten, so besteht entweder eine geringe Merkmalskonsistenz oder der Test besitzt eine Eigenschaft, die man als *Funktionsfluktuation* bezeichnet (Lienert & Raatz, 1994). Damit meinen die Autoren, der Test prüfe nach einer gewissen Zeit etwas anderes als vorher oder bei unterschiedlichen Bedingungen jeweils einen anderen Aspekt des zu untersuchenden Merkmals. Zum Teil geben Retest-Koeffizienten, die auf einem kurzen Zeitintervall beruhen, im Wesentlichen darüber Auskunft, ob ein Test *bedingungskonstant* ist. In sehr homogenen Stichproben kann es auch aufgrund der geringen Fehlervarianz zu geringeren Retest-Reliabilitäten kommen. In diesem Falle ist die Retest-Methode zur Bestimmung der Reliabilität ungeeignet.

Bezieht man die Retest-Reliabilität auf die intraindividuelle Werteverteilung einer Person, liegt eine perfekte Reliabilität ($\alpha = 1.00$) dann vor, wenn in aufeinander folgenden Messungen immer derselbe Wert erzielt wird. Da dies aus den o.g. Gründen jedoch unwahrscheinlich ist, geht man in der Regel bei der Retestreliabilität im Vergleich zur internen Konsistenz von etwas geringeren Werten aus.

Darüber hinaus kann ein Streudiagramm der individuellen Testwerte zu T1 und T2 Informationen über die *absolute* sowie *relative Stabilität* der Testwerte geben (Bühner, 2011, S. 160). Die *relative Stabilität* gibt an, ob bei einer Messwiederholung die Testpersonen dieselbe Rangreihe bezüglich des untersuchten Merkmals ausweisen, d.h. sich das Verhältnis der Ausprägungen zueinander nicht verändert hat. Es kann aber insgesamt zu einer (gleichartigen) Veränderung aller Testwerte gekommen sein, sichtbar durch eine Verschiebung der Winkelhalbierenden. Die *absolute Stabilität* berücksichtigt darüber hinaus auch, inwiefern die Messwerte bei T1 und T2 übereinstimmen, d.h. von der absoluten Höhe her gleich ausfallen. Liegen alle Messwerte im Streudiagramm auf der Winkelhalbierenden ($r_{T1T2} = 1.00$), liegt ein perfekter linearer Zusammenhang vor.

Unterscheiden sich die Werte einer Person bei einer Testwiederholung stark, könnte dies auch durch einen Messfehler bedingt sein. Diese mögliche Fehlerquelle kann man mittels Berechnung des *Reliable Change Index*‘ (*RCI*; Bühner, 2011, S. 206 ff) überprüfen.

$$RCI = \frac{x_{v2} - x_{v1}}{\sigma_{EDiff}}$$

Zur Berechnung des *RCI* wird die Differenz von Post- (x_{v2}) und Ausgangswert (x_{v1}) einer Person gebildet und durch den Standardfehler der Differenz (σ_{EDiff}) der beiden Werte dividiert. Den Standardfehler der Differenz erhält man durch

$$\sigma_{EDiff} = \sqrt{2 \cdot (\sigma_{Ex})^2}.$$

Wobei σ_{Ex} der Standardfehler ist, den man durch

$$\sigma_{Ex} = \sigma_1 \sqrt{1 - \rho_{tt}}$$

erhält. In die Bestimmung des *RCI* gehen somit auch die Standardabweichung der Werte in der Prämessung (σ_1) sowie die Retest-Reliabilität eines Tests (ρ_{tt}) mit ein.

Der *RCI* stellt einen z-Wert dar. Mit seiner Hilfe und der Standardnormalverteilung kann dann die Signifikanz ermittelt werden. Entspricht der Betrag des *RCI* dem Betrag des kritischen z-Werts oder liegt er darüber, ist die Veränderung von T1 zu T2 nicht auf eine mangelnde Messgenauigkeit der Messung zurückzuführen, sondern durch andere Faktoren bedingt.

Ein Faktor, der zu einer Verringerung der Retest-Reliabilität führen kann, ist die Unreliabilität der verwendeten Testverfahren. Da die Retest-Reliabilität in dem vorliegenden Fall auch von der internen Konsistenz des Tests selbst abhängig ist, die nicht vollkommen ist, soll die Retest-Reliabilität daher zusätzlich um die Unreliabilität der verwendeten Testverfahren bereinigt werden. Hierzu wird – angelehnt an Lienert und Raatz (1994) – folgende Minderungskorrektur zur Erfassung der korrigierten Retest-Reliabilität ($\alpha_{tt,korr.}$) verwendet:

$$\alpha_{tt,korr.} = \frac{r_{tt}}{\sqrt{\alpha_{t1} \cdot \alpha_{t2}}}$$

Hierzu wird der Korrelationskoeffizient, der die Retest-Reliabilität (r_{tt}) angibt, anhand der internen Konsistenz der verwendeten Testverfahren (α_{t1} und α_{t2}) relativiert. Lienert und Raatz (1994) empfehlen für Zwecke der Testinterpretation möglichst nur diese Formel zu verwenden.

6.2.2.3 Skalenhomogenität der Subskalen Fleiß und Ordnung

Zur Überprüfung der AGS wird darüber hinaus eine konfirmatorische Faktorenanalyse (confirmatory factor analysis; CFA; Bühner, 2011) gerechnet. Zunächst soll für jeden der beiden Gewissenhaftigkeitsaspekte eine eigene CFA gerechnet werden, um zu überprüfen, inwiefern die Items jeder Subskala homogen sind. Anschließend wird ein zweifaktorielles Modell spezifiziert, das sowohl den Fleiß- als auch den Ordnungsaspekt enthält und überprüft, inwieweit das theoretische Modell (s. Abschnitt 2.1.1.1) empirisch in den Daten wiederzufinden ist.

Die Berechnung der CFA erfolgt mittels der Software IBM AMOS Graphics 19.

Das Messmodell der CFA ergibt sich dabei aus dem aus der Theorie abgeleiteten Modell der latenten Variablen und den Items der AGS. Latente Variablen sind grundsätzlich nicht beobachtbar und werden durch die sog. manifesten oder beobachtbaren Variablen (= Items) abgebildet. Die angenommenen Beziehungen der manifesten Variablen zu den latenten Variablen sind schematisch im Messmodell (s. Abbildung 6-1) dargestellt.

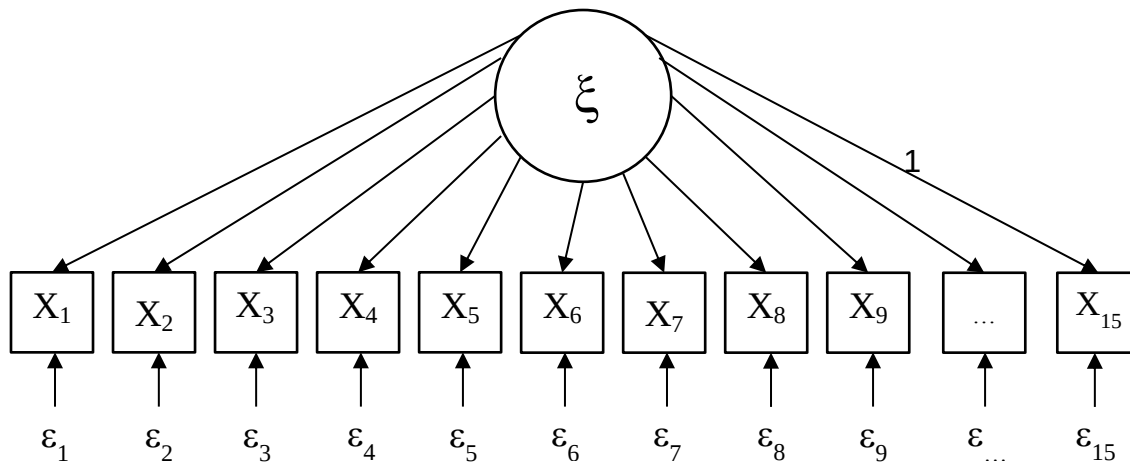


Abbildung 6-1: Schematische Darstellung des Messmodells zur Prüfung der Skalenhomogenität des Gewissenhaftigkeitsaspekts Fleiß bzw. Ordnung.

Im Messmodell wird die Itemantwort einer Person (X_{vi}) auf einem Item durch die individuelle Ausprägung der latenten Variable (ξ_v) und einen Fehler (ε_{vi}) erklärt.

$$X_{vi} = \lambda \xi_v + \varepsilon_{vi}$$

Konkret bedeutet dies, dass die Antworten der Probanden auf die 15 Items (X), die jeweils dem Fleiß- bzw. Ordnungsaspekt (ξ) zugeordnet sind, gemeinsam mit einer Fehlervarianz (ε) die latente Variable Fleiß bzw. Ordnung vorhersagen. Eine zugrundeliegende allgemeine Annahme dabei ist, dass die Fehler der einzelnen Items nicht miteinander kovariieren und die Fehler auch nicht mit der latenten Variable korrelieren. In diesem Falle steht (v) für den Variablenindex, während (i) den Personenindex repräsentiert.

Das Ziel einer CFA ist es, die empirische Varianz- bzw. Kovarianzmatrix durch die geschätzten Parameter (d.h. Ladungen, Kovarianzen und Fehlervarianzen) der linearen Gleichungen so gut wie möglich zu reproduzieren (Bühner, 2011, S. 381 ff). Es soll sich eine möglichst geringe Diskrepanz ergeben zwischen empirischer, d.h. aus der Stichprobe geschätzter, Populationskovarianzmatrix und der sog. implizierten Populationskovarianzmatrix, die sich aus den Parameterschätzungen des spezifizierten Modells ergibt.

Obwohl die empirischen Varianzen bzw. Kovarianzen bekannt sind, lassen sich keine eindeutigen Werte für Fehlervarianzen und Ladungen berechnen, d.h. die Parameter des Modells sind *nicht identifiziert* (Bühner, 2011). Um eine zuverlässige Schätzung der Parameter vornehmen zu können, wurde die Ladung auf das erste Item jeder der beiden Subskalen auf $\lambda = 1$ gesetzt.

Bühner (2011) zufolge ist die am häufigsten verwendete Schätzmethode die *Maximum-Likelihood-Methode* (ML). Es handelt sich hierbei um eine effektive Methode zur Schätzung von Populationskennwerten auf Basis der empirischen Kennwerte aus einer Untersuchungstichprobe. Die Populationsparameter werden so geschätzt, dass die Wahrscheinlichkeit für das Auftreten der beobachteten Stichprobendaten maximal wird. Voraussetzung für die Durchführung der ML-Methode ist eine multivariate Normalverteilung und Intervallskalenniveau der Items. Liegt keine multivariate Normalverteilung vor, führt dies zu überhöhten Schätzungen der χ^2 -Statistik und zu moderaten bis schwerwiegenden Unterschätzungen der Standardfehler. D.h. ein in der Population eigentlich passendes Modell wird durch den Test zu häufig abgelehnt und die Korrelationen zwischen den Variablen im Modell werden aufgrund der fälschlicherweise zu geringen Standardfehler signifikant (Bühner, 2011). Allerdings soll die Methode dem Autor zufolge vergleichsweise robust gegenüber Verletzungen der multivariaten Normalverteilung sein. Die ML-Methode soll daher zur Überprüfung der Modellparameter in der Kontrollgruppe herangezogen werden.

Für die Stichprobe der Bewerber soll die *Asymptotically Distribution-Free-Methode* (ADF) verwendet werden. Die ADF-Methode benötigt keine Verteilungsannahmen und ist dann zu empfehlen, wenn im Modell bspw. ordinale manifeste Variablen enthalten sind (Bühner, 2011). Vor allem bei großen Stichproben komme diese Methode dem Autor zufolge zu besseren Schätzungen und einer genaueren Schätzung des Modell-Fits im Vergleich zur ML-Methode.

Die Modellpassung wird mittels verschiedener Statistiken und Kennwerte überprüft:

Die X^2 -Statistik gibt das Gesamtmaß der Unterschiede zwischen implizierter und beobachteter Kovarianzmatrix an. Formal überprüft der X^2 -Test die Nullhypothese, ob die empirische Kovarianzmatrix der im Modell angenommenen entspricht:

$$H_0 = \Sigma_{\text{Modell}} = \hat{\Sigma}_{\text{empirisch}}$$

Ist der Wert signifikant, sind die Abweichungen zwischen implizierter und beobachteter Kovarianzmatrix nicht allein auf den Zufall rückführbar. Die Nullhypothese wird verworfen. Das heißt, das theoretisch angenommene Modell und die Struktur der Daten unterscheiden sich in bedeutsamer Weise. Die Beibehaltung der Nullhypothese spricht also dafür, dass das angenommene Modell dasjenige ist, welches die wahren kausalen Beziehungen zwischen den Variablen erklären kann.

Die X^2 -Statistik wird über das Minimum der Fit-Funktion (F_{Min}) berechnet und ist ein Kennwert für eine gewichtete Abweichung zwischen beobachteter und implizierter Kovarianzmatrix:

$$X^2 = (N - 1) \cdot F_{\text{Min}}$$

Die Teststärke der X^2 -Statistik steigt mit zunehmender Stichprobengröße (N) an, was einerseits zu kleineren Stichprobenfehlern und genaueren Parameterschätzungen führt, andererseits aber auch schon bei kleinen Abweichungen vom Modell zu einer Verwerfung des Modells führen kann (Bühner, 2011). Bei kleinen Stichproben kann dieser Effekt auch dazu führen, dass größere Abweichungen von einem perfekten Modell unentdeckt bleiben. Aus diesen Gründen sollen weitere Fit-Indizes, die die Güte der Passung der Daten in einem Modell – unabhängig von der Stichprobengröße – erfassen, zur Beurteilung der Modellgüte herangezogen werden.

Zusätzliche Fit-Indizes geben neben X^2 an, wie gut oder schlecht sich in den Daten ein theoretisch angenommenes Modell widerspiegelt. Es gibt verschiedene Fit-Indizes, wovon vor allem der *RMSEA* (Root-Mean-Square-Error of Approximation) und der *SRMR* (Standardized-Root-Mean-Residual) für die Persönlichkeitsforschung empfohlen werden (vgl. Bühner, 2011). Diese Fit-Indizes liefern unterschiedliche Informationen über den Modell-Fit.

Der *RMSEA* ist der am häufigsten angegebene Fit-Index. Es handelt sich hierbei um einen Indikator dafür, wie schlecht ein Modell die Daten beschreibt. Der Wertebereich geht von 0 bis 1. Ein hoher Wert steht hier für einen schlechten Modell-Fit. Der *RMSEA* stellt die durchschnittliche Abweichung der Daten vom Modell pro Freiheitsgrad (df) dar:

$$RMSEA = \sqrt{\frac{X^2 - df}{(N - 1) \cdot df}}$$

Da die X^2 -Statistik aber auch quadrierte Abweichungen enthält, wird zur Vereinfachung der Interpretation die Quadratwurzel gezogen. In Kombination mit dem *SRMR* (s.u.) schlägt Bühner (2011) die Verwendung folgender Cut-off-Werte zur Bewertung des *RMSEA* vor. Von einem guten Modell-Fit kann man demzufolge bei großen Stichproben ($N > 250$) ausgehen, wenn der *RMSEA* < 0.06 ist. Bei kleineren Stichproben ($N \leq 250$) ist von einem guten Modell-Fit auszugehen, wenn der *RMSEA* < 0.08 ist. Schmit et al. (1995) zufolge gilt ein

$RMSEA < 0.05$ als Indikator für einen guten Modell-Fit, bei einem $RMSEA < 0.08$ kann man den Autoren zufolge aber noch von einem gültigen Modell-Fit ausgehen.

Der $SRMR$ ist ebenfalls ein Indikator dafür, wie schlecht sich das angenommene, theoretische Modell in den Daten wiederfinden lässt. Der Kennwert kennzeichnet die standardisierte durchschnittliche Abweichung (= Residuum) zwischen der beobachteten und der implizierten Korrelationsmatrix (Bühner, 2011, S. 427). Als Cut-off-Wert in Kombination mit dem $RMSEA$ gilt dem Autor zufolge ein $SRMR < 0.11$ als Indikator für einen guten Modell-Fit.

Je nach Schätzmethode unterscheiden sich diese Fit-Funktionen und können somit auch zu unterschiedlichen Ergebnissen in der Modelltestung führen (Bühner, 2011).

Neben der Überprüfung der Homogenität der beiden Subskalen soll mit den o.g. Methoden auch das zweifaktorielle Modell, das den AGS zugrunde liegt, überprüft werden. Dieses ist schematisch in Abbildung 6-2 dargestellt.

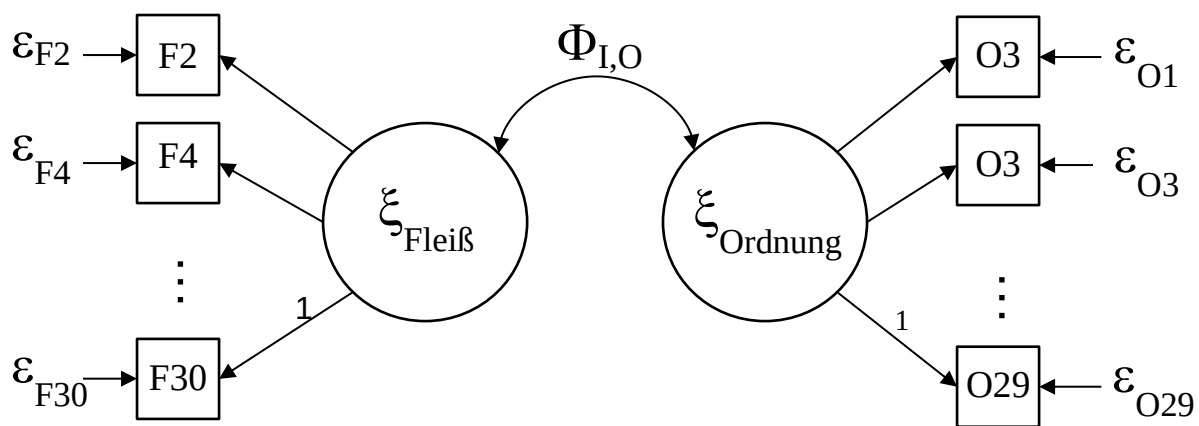


Abbildung 6-2: Schema des zweifaktoriellen Modells, das die latenten Variablen Fleiß und Ordnung enthält.

Vor der Berechnung einer CFA müssen einige Voraussetzungen in den Daten für diese Analyseschritte erfüllt sein bzw. überprüft werden. Einige Bedingungen können bereits an dieser Stelle als erfüllt angesehen werden und benötigen keiner weiteren Analysen. Hierzu zählt das Skalenniveau der zu untersuchenden Variablen. Um die ML-Methode anwenden zu können, muss ein Intervallskalenniveau in den Daten vorliegen, was mit der 5-stufigen Likert-Skala der AGS gegeben ist. Die ADF-Methode erfordert lediglich Ordinalskalenniveau, was ebenfalls erfüllt ist. Darüber hinaus sollte die Anzahl der Indikatoren pro latenter Variable angemessen groß sein, da diese sich ebenfalls auf die Ergebnisse einer CFA auswirken. Bühner (2011) zufolge sollte jede latente Variable mit mindestens vier Items spezifiziert werden. Da die AGS mit je 15 Items jede der beiden Subskalen Fleiß und Ordnung erfassen, kann diese Voraussetzung ebenfalls als erfüllt angesehen werden. Die Stichproben, mit deren Daten die CFA durchgeführt werden, sollten ausreichend groß sein, denn in kleinen Stichproben treten häufiger Schätzprobleme auf. Bühner (2011) empfiehlt hierzu ein Mindest-N von $N = 200$. Für erste Berechnungen an kleineren Stichproben mittels ML-Methode sind dem Autor zufolge aber auch kleinere Stichproben mit ca. $N = 100$ Probanden zulässig. Auch diese Voraussetzungen erfüllen die Stichproben (s. Kapitel 7). Eine Ausreißer- bzw. Extremwertanalyse wurde bereits vorab in der MVA (s. 6.2.1.1) vorgenommen und soll daher nicht noch einmal durchgeführt werden, zumal die Auswahl bzw. der Ausschluss von Probanden nicht ausschließlich aufgrund statistischer Kennwerte erfolgen sollte. Die in den Stichproben auftretenden Werte stellen natürliche Variationen im Antwortverhalten der Probanden dar. Eine Elimination dieser Kennwerte würde somit die Generalisierbarkeit der Ergebnisse dieser Studie einschränken.

Zu den Voraussetzungen, die vor der CFA noch überprüft werden müssen, zählt eine Verteilungsprüfung der Daten. Hierzu sollte bei der Anwendung der ML-Methode die Verteilung der Daten hinsichtlich ihrer multivariaten Normalverteilung überprüft werden, was mit Hilfe des Mardia-Tests vorgenommen wird (Bühner, 2011). Liegt keine multivariate Normalverteilung vor, schlägt der Autor das Bollen-Stine-Bootstrap-Verfahren zur Korrektur des p -Wertes für die χ^2 -Statistik vor. Sehr hoch miteinander korrelierende Items können vor allem im Rahmen von ML-Schätzungen zu Schätzproblemen führen. Um dies ausschließen zu können, sollte vor der CFA auch eine Kollinearitätsanalyse durchgeführt werden, um sehr hohe Korrelationen ($r > 0.85$) zwischen zwei oder mehr Items ausschließen zu können.

6.2.3 Konstruktvalidierung der AGS

Der letzte Schritt zur Überprüfung der Tauglichkeit der AGS als Prädiktor stellt eine Überprüfung der *Konstruktvalidität* dar. Es soll überprüft werden, wie stark die Faktoren der AGS bzw. die Gewissenhaftigkeitsaspekte Fleiß und Ordnung miteinander zusammenhängen. Darüber hinaus sollen die Zusammenhänge der Gewissenhaftigkeitsaspekte Fleiß und Ordnung mit der konstruktnahen Gewissenhaftigkeitsdimension sowie den konstruktfernen Dimensionen Neurotizismus und Offenheit für Erfahrungen aus dem NEO-FFI (Borkenau & Ostendorf, 1993; 2008) und ebenfalls divergenten Leistungsmaßen (Skalen 3 und 4 des LPS; Horn, 1983) überprüft werden. Diese Überprüfung soll mittels einer Multi-Trait-Multi-Method-Analyse (= MTMM; Campbell & Fiske, 1959) erfolgen. Zudem soll die faktorielle Zuordnung der Skalen Fleiß und Ordnung durch eine konfirmatorische Faktorenanalyse überprüft werden.

Die anderen beiden Validitätsarten – Inhaltsvalidität sowie Kriteriumsvalidität – sind in Kapitel 4 („Entwicklung der AGS“) bereits besprochen worden bzw. werden in den Abschnitten 6.3 ff sowie 8.3 ff behandelt.

Allgemein versteht man unter der Validität einer Skala den „*Grad der Genauigkeit, mit der ein Verfahren das misst, was es messen oder vorhersagen soll*“ (Rammstedt, 2004, S. 16). Auch wenn ein Verfahren sich als objektiv und reliabel erweist, muss es nicht valide sein. Reliable, aber invalide Messungen sind potenziell kritischer anzusehen, da auf die falschen Kriteriumsvariablen Rückschlüsse gezogen werden. Daher ist eine hohe Reliabilität nicht notwendigerweise auch ein guter Indikator für die Validität eines Persönlichkeitsmaßes (Schimmack, 2010).

Die Konstruktvalidität überprüft, inwiefern ein Test auch die Eigenschaft misst, die er messen soll (Bühner, 2011). Am häufigsten werden hierzu a priori konkrete Erwartungen über den Zusammenhang des vorliegenden Tests bzw. Testentwurfs mit konstruktverwandten bzw. *konvergenten* und konstruktfernen bzw. *diskriminanten* oder auch *divergenten* Tests zu formulieren. Die Konstruktvalidität kann somit in *konvergente Validität* und *diskriminante Validität* unterteilt werden.

Ein Ansatz zur systematischen Überprüfung der Konstruktvalidität ist der von Campbell und Fiske (1959) entwickelte Multi-Trait-Multi-Method-Ansatz (= MTMM), der auch in dieser Studie verwendet werden soll. Diese Analysemethode erlaubt es, differenzierte Aussagen über die beiden Kategorien der Konstruktvalidität – die konvergente sowie diskriminante Validität – eines Instruments vorzunehmen. Die MTMM-Analyse wird mittels bivariater Korrelationen berechnet und führt zu verschiedenen Korrelations-Matrizen. Der höchste Zusammenhang wird zwischen zwei Messungen desselben Merkmals mit derselben Methode erwartet (= Monotrait-Monomethoden). Eine etwas geringere Korrelation wird zwischen zwei Messungen desselben Merkmals mit unterschiedlichen Methoden erwartet (= Monotrait-Heteromethoden). Eine dazu vergleichsweise noch geringere Korrelation wird bei der Messung zweier unterschiedlicher Merkmale mit derselben Methode (= Heterotrait-Monomethoden; HTMM) erwartet. Die geringste Korrelation sollten die Messungen zweier Unterschiedlicher

Merkmale mit zwei unterschiedlichen Erhebungsmethoden ausweisen (=Heterotrait-Heteromethoden; HTHM). Die Messungen sollten jeweils unabhängig voneinander sein, damit die gefundenen Korrelationen lediglich auf die kausalen Effekte (z.B. Konstruktgleichheit) zurückzuführen sind und man davon ausgehen kann, dass sämtliche anderen potenziellen kausalen Effekte, die Korrelationen zwischen Variablen produzieren könnten, gleich null sind (Schimmack, 2010). Campbell und Fiske (1959) machten viele Unabhängigkeitsannahmen über die Methoden, mit denen Persönlichkeitseigenschaften untersucht werden sollten. Nur wenn diese Unabhängigkeitsannahmen korrekt sind, können die Korrelationen zwischen den interessierenden Variablen verwendet werden, um die Variation im Sinne einer perfekten Validität der latenten Persönlichkeitseigenschaft zu bestimmen. Wobei man davon ausgehen muss, dass es – vor allem in Feldstichprobenuntersuchungen – niemals eine perfekte Unabhängigkeit geben wird. Leichte Verletzungen der Unabhängigkeitsbedingung werden daher zugelassen und sollten nur zu kleinen Biases in der Schätzung der Validitätskoeffizienten führen. Aus diesem Grund sind für die Heterotrait-Heteromethoden-Matrix keine Korrelationen von $r = 0$ zu erwarten.

Die konvergente Validität bezeichnet die Übereinstimmung des untersuchten Instruments mit anderen Verfahren zur Messung des interessierenden Konstrukts. Als Daumenregel für den Zusammenhang von Tests mit gleichem oder ähnlichem Gültigkeitsbereich gilt Bühner (2011) zufolge eine Korrelation (s. 6.3.1) von $r > .50$. Die Interkorrelationen der Skalen sollten dabei kleiner als die Reliabilitäten sein. Eigentlich reicht es bereits aus, die konvergente Validität zur Bestimmung der Konstruktvalidität zu untersuchen, die Bestimmung der diskriminanten Validität dient lediglich dem Ziel, mögliche Fehlereinflüsse identifizieren zu können (Schimmack, 2010). In dieser Untersuchung dient die Gewissenhaftigkeitsskala des NEO-FFI (Borkenau & Ostendorf, 2008) als konstruktkonvergentes Kriterium, um die inhaltliche Nähe der Gewissenhaftigkeitsaspekte Fleiß und Ordnung zu überprüfen. Die beiden Aspekte Fleiß und Ordnung sollen ebenfalls miteinander verglichen werden.

Die diskriminante Validität beschreibt die Unterscheidungsleistung des untersuchten Instruments hinsichtlich der Erfassung anderer, „fremder“ Konstrukte. Die Korrelation mit Tests aus einem anderen Gültigkeitsbereich sollte der Daumenregel von Bühner (2011) zufolge nicht größer als $r < .40$ sein. Als konstruktdivergente Verfahren sollen die Skalen Neurotizismus und Offenheit aus dem NEO-FFI (Borkenau & Ostendorf, 1993, 2008) (HTMM) in die Analysen eingehen. Die LPS-Skalen 3 und 4 (HTHM) werden an einer späteren Stelle im Rahmen der Kriteriumsvalidität (vgl. 6.3. ff) sowie der inkrementellen Validitätsanalyse (s. Abschnitt 6.4) als Vergleichswerte hinzugezogen.

Zentrale Voraussetzung zur Durchführung einer MTMM-Analyse ist, dass zwischen den gemeinsam analysierten Variablen keine Multikollinearität herrscht.

Man kann neben der MTMM-Analyse aber auch mittels konfirmatorischer Faktorenanalysen die Konstruktvalidität einer Skala überprüfen. Dies hat gegenüber der MTMM-Analyse den Vorteil, dass konstruktnahe und konstruktferne Inhaltsbereiche voneinander getrennt werden können. Hierzu zählt u.a. auch die Prüfung des vor der Testkonstruktion aufgestellten Testmodells (Bühner, 2011). Auch Borkenau und Ostendorf (1991) sprechen sich für eine zusätzliche Überprüfung der Struktur des Konstrukts aus, denn eine Evaluation von MTMM-Matrizen per Augenschein berge zahlreiche Nachteile. Zeitgemäßer sei den Autoren zufolge eine Spezifikation und Überprüfung einschlägiger Strukturgleichungsmodelle. Die auf diese Weise erhaltene χ^2 -Statistik gibt dann an, ob das empirische Korrelationsmuster mit den Annahmen zur Faktorstruktur (s. Abb. 6-3) verträglich ist. Aus diesem Grund wird zur Überprüfung der Konstruktvalidität der AGS auch eine konfirmatorische Faktorenanalyse berechnet. Die hierzu notwendigen Voraussetzungen und entsprechenden Kennwerte zur Modellüberprüfung sind in Abschnitt 6.2.2 ausführlich dargestellt.

Abbildung 6-3 veranschaulicht schematisch das Modell zur Überprüfung der konvergenten Validität der jeweils 15 Items der beiden Gewissenhaftigkeitsaspekte Fleiß (F) und Ordnung (O) mit den zwölf Items der Gewissenhaftigkeitsdimension (GH) aus dem NEO-FFI (Borkenau & Ostendorf, 2008) mittels konfirmatorischer Faktorenanalyse.

Die Faktoren Fleiß, Ordnung sowie Gewissenhaftigkeit stellen in diesem Modell die latenten Variablen (ξ) dar, deren Indikatoren jeweils die Items jeder Fragebogenskala sind. Da die Überprüfung des Zusammenhangs der Konstrukte ein Ziel dieser Arbeit ist, werden Korrelationen zwischen den Variablen zugelassen.

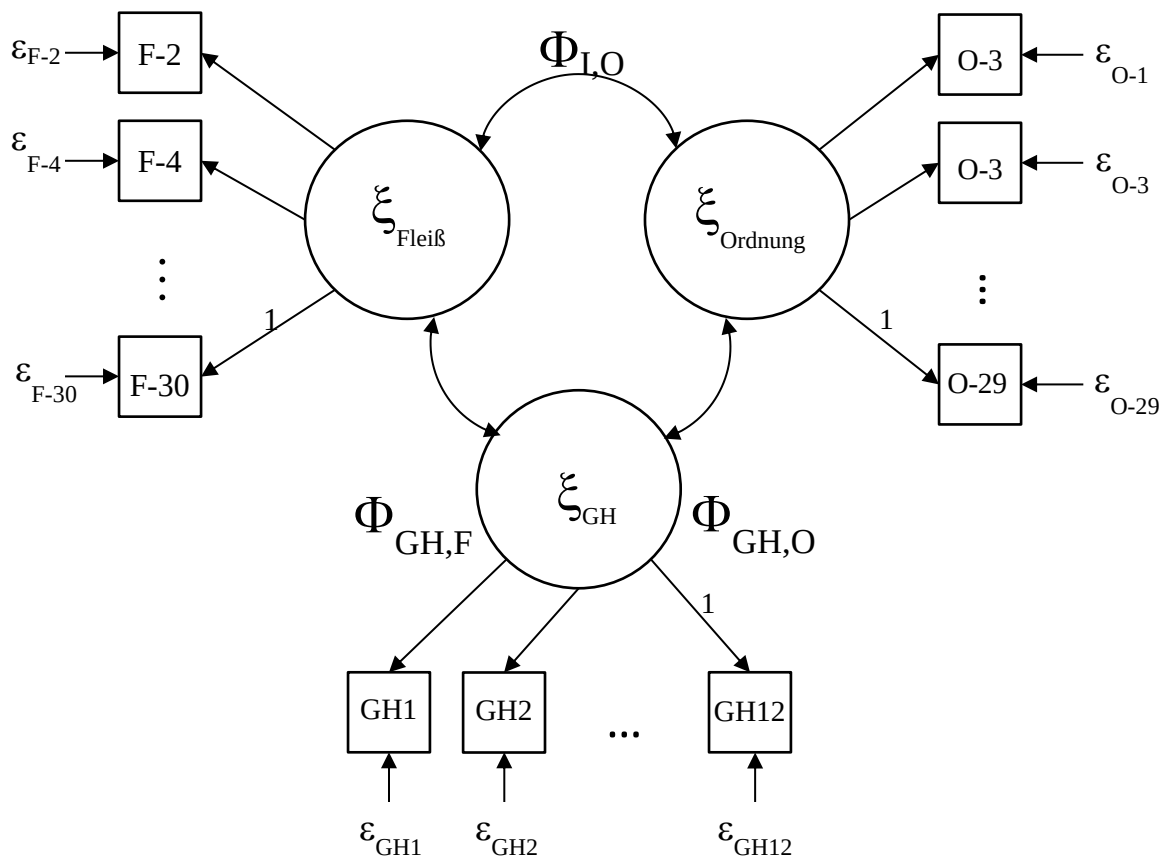


Abbildung 6-3: Schematische Darstellung des Modells zur Überprüfung der konvergenten Validität mit der Gewissenhaftigkeitsdimension aus dem NEO-FFI.

Ein allgemeiner Nachteil sowohl der MTMM- Analyse als auch der CFA ist allerdings, dass ein Test nur mit anderen Tests verglichen wird, deren Inhaltsvalidität (theoretisch) ebenfalls unzureichend ist bzw. sein kann. Da es sich bei den verwendeten Kriteriumsdaten jedoch um allgemein akzeptierte und viel beforschte Verfahren handelt, wird davon ausgegangen, dass dies in der vorliegenden Untersuchung keinen negativen Einfluss auf die Gültigkeit der Ergebnisse hat.

6.3 Vorhersage von Ausbildungs- bzw. Schulleistung

Im folgenden Abschnitt werden die Methoden, die zur Überprüfung der Vorhersagekraft der Schul- bzw. Berufsschulnoten und Vorgesetztenbeurteilungen verwendet wurden, erläutert. Mit Hilfe allgemeiner Korrelationsanalysen (s. Abschnitt 6.3.1) sollen vorab einfache Zusammenhänge zwischen den verschiedenen Variablen aufgezeigt werden. In Abschnitt 6.3.2 werden hierzu notwendige Korrekturformeln vorgestellt. In Abschnitt 6.3.3 werden nachfolgend die einzelnen Schritte zur Überprüfung der Varianzaufklärung der verschiedenen Prädiktoren erläutert.

Die Wirkung der untersuchten Prädiktorvariablen auf die jeweilige Kriteriumsvariable soll in gegenseitiger Abhängigkeit erfolgen. Zur Überprüfung des modellierten Zusammenhangs und Analyse der Wirkweise der Vorhersage- und Kriteriumsvariablen sollen daher sowohl Moderations- als auch Mediationsanalysen durchgeführt werden, welche in den Abschnitten 6.3.4 und 6.3.5 erläutert sind.

6.3.1 Korrelationsanalysen

Zur Überprüfung der einfachen bivariaten Zusammenhänge zwischen den dispositionellen Variablen Fleiß, Ordnung, Selbstwirksamkeit und den wahrgenommenen eigenen Fähigkeiten mit den Leistungsvariablen Schul- bzw. Berufsschulnoten sowie der Vorgesetztenbeurteilung sollen zuerst Korrelationsanalysen berechnet werden. Diese können auch im Sinne einer Kriteriumsvalidität interpretiert werden. Bühner (2011) zufolge handelt es sich bei der Kriteriumsvalidität um die Stärke der Übereinstimmung eines Instruments mit einem von dem Instrument unabhängigen Außenkriterium, das insbesondere für den Anwendungsnutzen des Verfahrens bedeutsam ist.

Voraussetzung für die Berechnung der Korrelationskoeffizienten ist eine bivariate Normalverteilung der Grundgesamtheit, wobei diese schwer nachzuprüfen ist und der Signifikanztest für Korrelationskoeffizienten zudem äußerst robust sowohl gegenüber Verletzungen der Verteilungsannahmen als auch gegenüber Verletzungen des vorausgesetzten Intervallskalenniveaus ist. Laut Bortz (1993) kann ungeachtet der Verteilungseigenschaften in einer Population stets ein Produkt-Moment-Korrelationskoeffizient berechnet werden. Dieser berechnet sich aus dem Anteil der Kovarianz (Kov) an der Gesamtvarianz (V) und „kennzeichnet als deskriptives Maß die Enge des in der Stichprobe angetroffenen, linearen Zusammenhangs zwischen zwei Merkmalen bzw. als Determinationskoeffizient r^2 den Anteil der gemeinsamen Varianz, der auf die lineare Beziehung zurückgeht“ (Bortz, 1993, S. 197) bzw. dient als „Maß für die Stärke des linearen Zusammenhangs zwischen zwei Zufallsvariablen x und y “ (Rasch & Kubinger, 2006, S. 245). In dieser Untersuchung soll daher der Pearson-Korrelationskoeffizient (r_{xy}) zur Überprüfung der Zusammenhänge zwischen den Variablen berechnet werden:

$$r_{xy} = \frac{Kov(x,y)}{\sqrt{V(x)V(y)}}$$

Die Werte von r_{xy} liegen stets zwischen -1 und +1, wobei ein Wert von $r_{xy} = 0$ bedeutet, dass zwischen den beiden Variablen kein Zusammenhang besteht bzw. diese unabhängig voneinander sind. Besteht zwischen den Variablen jedoch ein positiver (negativer) Zusammenhang, so gehen hohe Ausprägungen der Variablen x tendenziell mit hohen (niedrigen) Ausprägungen in y einher.

Werden aus der Grundgesamtheit Stichproben gezogen, die stark systematisch vorselektiert sind, was in dieser Untersuchung für die Bewerberstichprobe der Fall ist, ist es möglich, dass die Variationsbreite in dieser stark eingeschränkt ist und es somit zu einer Unterschätzung der Populationskorrelation durch die Stichprobenkorrelation kommt. Für diesen Fall gibt es Korrekturformeln, die im folgenden Abschnitt 6.3.2 beschrieben werden. Allgemein ist hierzu jedoch auch anzumerken, dass die gefundenen Ergebnisse keinen Anspruch auf Gültigkeit in einer allgemeinen Grundgesamtheit erheben, sondern vor allem im Kontext der Leistungsvorhersage von Schulabgängern Gültigkeit besitzen sollen.

Ergebnisse von Korrelationsanalysen können Bortz (1993) zufolge auf vier verschiedene Arten interpretiert werden. Eine Möglichkeit ist, unidirektional von einem Einfluss von Variable A auf Variable B bzw. B auf A auszugehen. Zum anderen ist aber auch eine wechselseitig kausale Beeinflussung der Variablen möglich. Auch ein kausaler Einfluss einer

Drittvariablen ist denkbar in dem Sinne, dass ein Zusammenhang vorgetäuscht wird, wenn die beiden untersuchten Merkmale von einem dritten abhängen. Allerdings liefert der Korrelationskoeffizient keine Informationen darüber, welche dieser Interpretationsmöglichkeiten die Richtige ist. Bortz (1993) zufolge sind in den meisten sozialwissenschaftlichen Studien die korrelativen Zusammenhänge auf den Einfluss von Drittvariablen zurückzuführen, die beide Variablen beeinflussen. Dem Autor zufolge ist „eine Korrelation zwischen zwei Variablen eine notwendige, aber keine hinreichende Voraussetzung für kausale Abhängigkeiten“ (Bortz, 1993, S. 217) und können daher nur als Koinzidenzen interpretiert werden, in dem Sinne, dass sie Hinweise liefern, zwischen welchen Merkmalen Kausalbeziehungen bestehen könnten. Aussagen über lineare Zusammenhänge können mittels grafischer Darstellung der Daten gemacht werden. Zur Überprüfung von kausalen Beziehungen sind allerdings weitere Analysen notwendig, die in den folgenden Abschnitten 6.3.4 sowie 6.3.5 erläutert werden.

Zu bedenken ist zudem, dass eine einfache Korrelation r auch keine Aussage über die Bedeutsamkeit eines Zusammenhangs erlaubt, beispielsweise in dem Sinne, dass es sinnvoll erscheint, mit Hilfe einer Regressionsanalyse einen unbekannten Kriterium y aus einem Prädiktor x vorherzusagen. So unterliegt die (Berufs-) Schulleistung als Kriterium beispielsweise einer gewissen Streuung, die dazu führt, dass diese nicht bei allen (Berufs-) Schülern gleich ausgeprägt ist (Rasch & Kubinger, 2006). Ein Indikator für die Bedeutsamkeit eines Zusammenhangs ist das *Bestimmtheitsmaß* r^2 , das durch Multiplikation des Korrelationskoeffizienten r_{xy} mit sich selbst erzeugt wird. Das Bestimmtheitsmaß r^2 gibt Auskunft darüber, wie groß der Anteil der Varianz der Variablen y ist, der durch die Varianz der Variablen x erklärt werden kann. Die Werte von r^2 liegen zwischen 0 und 1. Je näher der Wert an 1 liegt, desto mehr Varianz kann erklärt werden. Ein Wert von $r^2 = .50$ zeigt an, dass die Hälfte der Varianz erklärt wird. Laut Rasch und Kubinger (2006) spricht man in diesem Fall von einem mittleren Zusammenhang. Bei Multiplikation von r^2 mit 100 kann die Stärke des Zusammenhangs in Prozent angegeben werden.

Bühner (2011) gibt als Daumenregel für Korrelationswerte im Sinne einer prädiktiven Validität an, dass die Zusammenhänge mindestens $r > .20$ sein sollten. Görlich und Schuler (2007) führen an, dass die DIN 33430 die Kriteriumsvaliditäten mit Orientierungswerten zwischen $r = .30$ bis $.55$ vergleichsweise hoch ansetze. Allerdings unterscheide die DIN 33430 auch nicht zwischen verschiedenen Arten von Kriterien (z.B. Schul- oder Berufsschulnoten vs. innerbetriebliche Leistung).

6.3.2 Kriterien der Dateneignung

In der Planung und Durchführung von Validierungsstudien müssen immer auch Artefakte bzw. Fehlerquellen in den Daten berücksichtigt werden, die zu einer Verringerung der Validitätskennwerte führen können. So wurden in dieser Untersuchung in der Bewerberstichprobe vor allem Bewerber eingestellt, die in den Testverfahren hohe Werte erzielten. Zudem kann es aufgrund der Unreliabilität der verwendeten Prädiktoren zu einer weiteren Verringerung der Aussagekraft der gefundenen Ergebnisse gekommen sein. Aus diesem Grund sollen im Folgenden kurz Maßnahmen zur Korrektur der Range Restriction (6.3.5.1) sowie Unreliabilität der Prädiktorvariablen (6.3.5.2) dargestellt werden.

6.3.2.1 Range Restriction

Da in dieser Studie auch Daten analysiert werden, die zeitlich deutlich nach dem ersten Testtag (T1) bzw. nach der Erhebung der Prädiktorvariablen erhoben wurden, sollten mögliche Dropout- und Selektionseffekte berücksichtigt werden, die die Aussagekraft der Ergebnisse in Bezug auf die Kriteriumsvariable beeinflussen könnten. Bei der Bewerberstichprobe zum zweiten Testzeitpunkt (T2) handelt es sich nämlich um eine stark selektierte Subgruppe, die

explizit anhand der Prädiktorvariablen ausgewählt wurde. Es ist daher davon auszugehen, dass die Varianz der Daten in dieser Subgruppe deutlich geringer ist als in der Gesamtgruppe der Ausbildungsbewerber, was zu einer geringeren Vorhersagekraft bzw. Varianzeinschränkung (= *Range Restriction*) der Daten in Bezug auf das Kriterium führen kann (Lievens et al., 2009; Schmidt et al., 2008; Wiberg & Sundström, 2009). Aus diesem Grund soll eine Range-Restriction-Korrektur vorgenommen werden. Da die Probanden der Bewerberstichprobe sowohl auf Basis der dispositionellen Variablen als auch (weiterer) Leistungstestdaten sowie Ergebnissen aus Bewerbungsgesprächen und Interaktionsübungen ausgewählt wurden, handelt es sich um eine direkte und indirekte Range Restriction, weshalb *Thorndikes case 2 Korrekturformel* zur Ermittlung des korrigierten Korrelationskoeffizienten (σ_{xy})

$$\sigma_{xy} = \frac{s_{x.ges} \cdot r_{xy}}{\sqrt{s_{x.ges}^2 \cdot r_{xy}^2 + s_x^2 - s_x^2 \cdot r_{xy}^2}}$$

angewendet werden soll. In dieser Formel bezeichnet r_{xy} den gefundenen empirischen Zusammenhang zwischen dem Prädiktor und dem Kriterium in der eingeschränkten Analysestichprobe. Mit s_x ist die Streuung der Daten der Prädiktorvariable in der eingeschränkten Stichprobe gemeint, während $s_{x.ges}$ die Streuung der Prädiktorvariable in der Gesamtstichprobe erfasst. Wiberg und Sundström (2009) verglichen die oben angeführte Formel zur Korrektur der Range Restriction mit einer Imputationsformel. Den Autorinnen zufolge ist *Thorndikes case 2 Korrektur* eine gute und viel verwendete Methode, um die wahre Vorhersagekraft der Daten aus Auswahl-situationen zu schätzen. Der wahre Korrelationskoeffizient wird hierbei immer noch etwas unterschätzt.

6.3.2.2 Minderungskorrektur

Alle psychologischen Testverfahren besitzen einen bestimmten Grad an Unreliabilität. Der Zusammenhang zwischen Testwerten und einem Kriterium hängt rechnerisch allerdings von der Reliabilität ab. D.h. misst man die interessierenden Merkmale nur ungenau, kann der systematische messfehlerfreie Zusammenhang nur unzureichend geschätzt werden. Unterscheidet sich die Reliabilitäten der Testkennwerte zudem zwischen und innerhalb der Stichproben, sind auch die Korrelationen zwischen den Stichproben nicht vergleichbar. Erst wenn man die Korrelationen in beiden Stichproben anhand der Reliabilitätskennwerte aufgewertet hat, sind die Korrelationen vergleichbar. Dazu berechnet man die *einfache* oder *doppelte Minderungs- oder Attenuitätskorrektur* (vgl. Bühner, 2011, S. 212). Da Reliabilitätsschätzungen meist stichprobenabhängig sind, ist eine Minderungskorrektur sinnvoll, um die Korrelationen der wahren Werte von zwei Messungen zu schätzen.

Eine Minderungskorrektur gibt somit an, wie stark zwei Variablen miteinander korrelieren, wenn diese perfekt messen würden (Lienert & Raatz, 1992). Der Zusammenhang zwischen Prädiktor und Kriterium ohne Minderungskorrektur wird *operationale Validität* genannt (Lievens, Ones & Dilchert, 2009). Diese unkorrigierte prädiktive Validität unterschätzt den Autoren zufolge die Vorhersagekraft der dispositionellen Variablen in Bezug auf die Kriteriumsvariablen.

Die doppelte Minderungskorrektur berücksichtigt die Reliabilität bzw. Messgenauigkeit beider Messwerte, also von Prädiktor (r_{xx}) und Kriterium (r_{yy}) (Lienert & Raatz, 1992):

$$r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} \cdot r_{yy}}}$$

Bei der einfachen Minderungskorrektur wird nur ein Messwert korrigiert. Görlich und Schuler (2007) führen an, dass bezüglich der IHK-Zwischenprüfungsnoten bzw. IHK-Zwischenzeugnispunkte eine Reliabilität von $r_{yy} = .80$ angenommen werden kann. Werden die Validitäten um den Kriteriumsmessfehler korrigiert, so steigt die Validität bzw. die aufgeklärte Varianz

an. Die Reliabilitätskennwerte der Prädiktorvariablen sind bekannt bzw. werden im Rahmen der Skalenüberprüfung (s. 6.2.2.1) berechnet.

Da es allerdings aus praktischen Gesichtspunkten nicht hilfreich ist, davon auszugehen, es gäbe perfekt messgenaue Schulnoten, soll in dieser Studie neben der doppelten ($r_{xy'}$) und der Kriteriumskorrektur (r_{xy}) auch die einfache Minderungskorrektur zur Ermittlung einer korrigierten Prädiktorkorrelation ($r_{x'y}$) verwendet werden:

$$r_{xy'} = \frac{r_{xy}}{\sqrt{r_{yy}}} \quad \text{bzw.} \quad r_{x'y} = \frac{r_{xy}}{\sqrt{r_{xx}}}$$

Die einfache Minderungskorrektur ist somit im letzten Fall auch ein Indikator dafür, ob ein Messinstrument noch verbessert werden sollte. Wenn die Vorhersagekraft des Tests bzw. einer Skala gering ist, könnte durch eine Verbesserung der Messgenauigkeit auch eine höhere Vorhersagekraft erreicht werden. Da die in dieser Studie verwendeten Testverfahren bzw. Prädiktorvariablen alle eine ausreichend hohe Reliabilität aufweisen, sollte es jedoch nicht zum sog. *Verdünnungsparadoxon* kommen. Diesem zufolge wird der korrigierte Zusammenhang umso höher, je geringer die Messgüte der verwendeten Instrumente ist, sodass es sich letztlich nur noch um einen artifiziellen Wert handelt, der mit dem wahren Zusammenhang der untersuchten Variablen keine Ähnlichkeit mehr aufweist bzw. nicht in die Realität übertragbar ist.

Um eine Minderungskorrektur durchführen zu können, sollten die Skalen keine korrelierten Messfehler aufweisen und die Voraussetzungen für den jeweiligen Reliabilitätskoeffizienten erfüllen (Bühner, 2011).

Allerdings warnt Bühner (2011) auch vor einer einfachen Interpretation der korrigierten Testwerte als „wahre Werte“, da bei der Anwendung einer Minderungskorrektur nur ein „optimales Szenario“ erstellt wird, das u.U. stark von der Realität abweichen kann. Es ist zudem möglich, dass minderungskorrigierte Korrelationen größer als eins werden, wenn die Reliabilität der untersuchten Testkennwerte unterschätzt wird. Das geschieht dem Autor zufolge allerdings meist nur in Fällen, wenn die Voraussetzungen für die Berechnung der Reliabilitätskoeffizienten verletzt sind.

6.3.3 Regressionsanalysen

Zur Modellüberprüfung sollen im Folgenden Moderations- (s. 6.3.4) und Mediationsanalysen (s. 6.3.5) berechnet werden, die nach dem Ansatz von Baron und Kenny (1986) methodisch auf Regressionsanalysen basieren. Das Ziel einer Regressionsanalyse ist die Überprüfung der Varianzaufklärung zur Ermittlung der Vorhersagekraft einer Variablen x (Prädiktor) in Bezug auf eine Variable y (Kriterium). Diese wird mittels einer Regressionsgeraden geschätzt, die den Zusammenhang zwischen x und y beschreibt. Diese beschreibt allgemein die Funktion:

$$\hat{y} = b \cdot x + a$$

Der Regressionskoeffizient wird mit b bezeichnet und steht für die Steigung der Geraden. Das a in der Gleichung bezeichnet den Achsenabschnitt, bei der die Regressionsgerade die y -Achse schneidet. Eine positive Steigung b zeigt an, dass die Werte y mit größer werdenden x -Werten ebenfalls anwachsen. Entsprechend zeigt eine negative Steigung b an, dass die y -Werte bei größer werdenden Werten von x tendenziell kleiner werden (Bortz, 1998). Allerdings hängt ein Kriterium nicht nur von einem Prädiktor allein ab, sondern von vielen Variablen, sodass die einfache lineare Regressionsgleichung um weitere Regressionskoeffizienten b_1 bis b_k erweitert werden kann:

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_k \cdot x_k + a_{1.23\dots k}$$

Die Partialregressionskoeffizienten (b_1 bis b_k) werden auch *b-Gewichte* der einzelnen Prädiktorvariablen genannt. Sie stellen die relativen Gewichte der Prädiktorvariablen in der Vorhersage dar und können mit Hilfe von *t*-Tests auf Signifikanz überprüft werden.

Es ist auch möglich, die Regressionsgerade in einer standardisierten Form zu beschreiben:

$$z_{\hat{y}} = \beta_1 \cdot z_1 + \beta_2 \cdot z_2 + \dots + \beta_k \cdot z_k$$

Der Vorteil der Standardisierung ist, dass die Beta-Gewichte nur Werte zwischen -1 und +1 annehmen und wie Korrelationskoeffizienten interpretiert werden können. Die additive Konstante a entfällt hier, da das standardisierte Kriterium ($z_{\hat{y}}$) einen Mittelwert von null hat. Auch die Regressionskonstanten können mittels *t*-Test auf Signifikanz überprüft werden. Die multiple Regression selbst wird mittels F-Test auf Signifikanz getestet.

Voraussetzung für die Berechnung einer Regressionsanalyse ist, dass sowohl die Prädiktors als auch Kriteriumsvariablen intervallskaliert und normalverteilt sind. Darüber hinaus sollten die zu den x-Werten gehörenden y-Werte homogen (Homoskedastizität muss vorliegen) und unabhängig sein.

Es ist möglich, dass das an der Untersuchungsstichprobe ermittelte R^2 den bestehenden Zusammenhang in der Population überschätzt. Je kleiner die Untersuchungsstichprobe und je größer die Anzahl der Prädiktoren (K) ist, desto wahrscheinlicher ist eine Überschätzung von R^2 . Olkin und Pratt (1958) schlugen zur Kontrolle dieses Effekts eine Schrumpfungskorrektur vor:

$$\hat{R}^2 = 1 - \left(\frac{N - 3}{N - K - 2} \right) \cdot \left((1 - R^2) + \left(\frac{2}{N - K} \right) \cdot (1 - R^2)^2 \right)$$

Von einem sogenannten *Supressor-Effekt* spricht man, wenn ein Prädiktor die Varianzaufklärung verbessert, ohne dass er selbst jedoch mit dem untersuchten Kriterium korreliert.

6.3.4 Moderationsanalysen

Eine Moderationsanalyse überprüft, inwiefern die Wirkungsintensität einer Prädiktorvariable auf eine Kriteriumsvariable von einer zweiten (Moderator-) Variable abhängig ist (Müller, 2007). Dem Autor zufolge besteht eine Moderatorbeziehung dann, „wenn eine qualitative oder quantitative Variable die Richtung oder die Stärke einer Beziehung zwischen einer Prädiktor- und einer Prognosevariablen beeinflusst“ (S. 246). Abbildung 6-4 veranschaulicht grafisch, welche der in dieser Studie untersuchten Variablen als Moderator im Vorhersage-

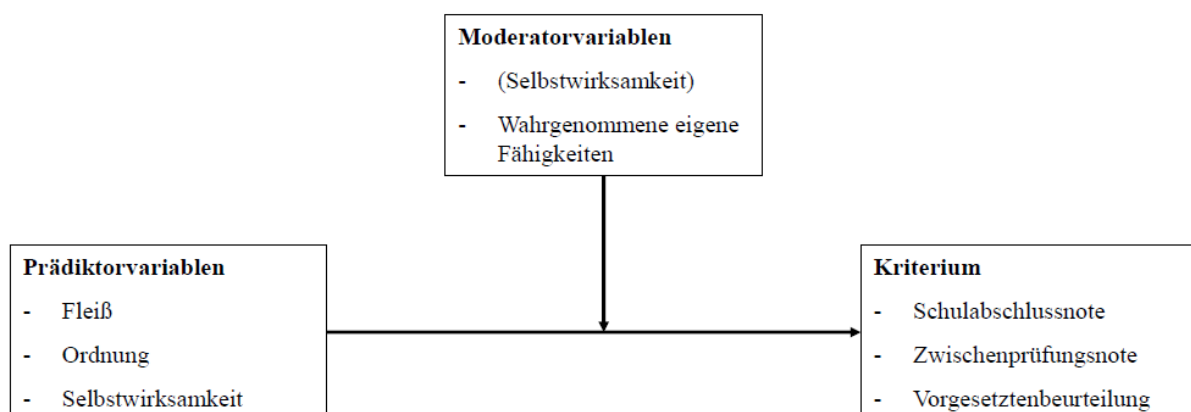


Abbildung 6-4: Darstellung möglicher Moderatorbeziehung/en im Untersuchungsmodell

modell fungieren könnten. Moderationsanalysen werden nach dem Ansatz von Baron und Kenny (1986) mittels Regressionsanalysen (s. Abschnitt 6.3.2) berechnet. Die Moderatorbeziehung wird hierbei einfach berechnet durch die Erweiterung der Regressionsgleichung um einen Interaktionsterm ($x_i x_j$), der die in Interaktion stehenden Variablen multiplikativ miteinander verbindet:

$$y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 x_2 + e$$

Es ist auch möglich, dass mehr als zwei Variablen miteinander interagieren. Die Modellgleichung wird dann um die entsprechenden Interaktionsterme ergänzt bzw. erweitert.

Die Ausprägung der Moderatorvariablen beeinflusst demzufolge das von der resultierenden Regressionsgerade erreichte Niveau und deren Steigung und führt zu zwei Verbesserungen. Man kann die vorhergesagten Wirkungszusammenhänge in Bezug auf die unterschiedlichen Ausprägungen der Prädiktorvariablen differenzieren und zudem die Güte der Vorhersage der Regression verbessern bzw. Fehlschätzungen verhindern.

Vor der Durchführung einer Moderationsanalyse sollten einige Voraussetzungen überprüft werden. So sollten Homogenität, Normalität sowie Linearität in den Daten vorliegen. Ebenfalls sollte vorab die Multikollinearität der Daten überprüft werden und ggf. Variablen oder Ausreißer, die die Regressionen stark beeinflussen könnten, vor den Analysen ausgeschlossen werden. Um eine ausreichende statistische Aussagekraft zu erreichen, sollte die Stichprobe fünf bis zehn Datensätze pro Prädiktorvariable nicht unterschreiten.

Der erste Schritt der Moderationsanalyse ist eine multiple Regression (s. 6.3.2) aller interessierenden Variablen. In diesem ersten Schritt bleiben die möglichen Interaktionen noch unberücksichtigt. In den folgenden Berechnungen wird das Modell dann schrittweise um die Interaktionsterme erweitert und die Anpassungsgüte der verschiedenen Modelle mittels des Gütemaßes R^2 ermittelt. Übersteigt das R^2 der Gleichung, die den Interaktionsterm enthält ($= R_2^2$), das R^2 des Modells ohne Interaktion ($= R_1^2$), wird anschließend die Signifikanz der Interaktion mit k Prädiktoren mittels eines F-Tests überprüft:

$$F = \frac{(R_2^2 - R_1^2)/(k_2 - k_1)}{(1 - R_2^2)/(N - k_2 - 1)}$$

Wird der F -Test signifikant, werden die Interaktionsterme in das Modell übernommen (Müller, 2007). Wird der F -Test nicht signifikant, werden die Interaktionsterme aus der Gleichung gelöscht.

Die Stärke des Moderatoreffekts (f^2) kann als Verhältnis der durch die Interaktion aufklärten systematischen Varianz relativ zur nicht erklärten Varianz des Kriteriums bestimmt werden.

Da verschiedene Modelle berechnet werden, soll am Ende das Modell ausgewählt werden, welches die beste Anpassung gewährt. Bei komplexen Modellen kann es allerdings sein, dass sich die Modelle in ihrer Anpassungsgüte nicht sehr voneinander unterscheiden. In der Regel wählt man dann das Modell aus, das die beste theoretische Verankerung besitzt, auch wenn es eine etwas schlechtere Anpassungsgüte aufweist (Müller, 2007).

6.3.5 Mediationsanalysen

Mediationsbeziehungen lassen sich besonders dann berichten, wenn in einem zu untersuchenden Modell Prozesse abgebildet werden, deren Einflussgrößen durch Kausalbeziehungen miteinander verbunden sind. Da die verschiedenen dispositionellen Variablen und Leistungs- bzw. Erfolgsindikatoren in dieser Studie einen starken Zusammenhang miteinander aufweisen sollten, soll mittels Mediationsanalysen überprüft werden, auf welche Weise die Variablen

miteinander interagieren bzw. inwiefern Selbstwirksamkeit oder wahrgenommene eigene Fähigkeiten zwischen Gewissenhaftigkeitsaspekten und den Kriteriumsvariablen vermitteln.

Müller (2007) zufolge wird eine Variable „dann als Mediator bezeichnet, wenn eine Prädiktorvariable zuerst einen Effekt auf die Mediatorvariable bewirkt und diese wiederum die Prognosevariable beeinflusst“ (S. 255). In einem dritten Schritt muss die Mediatorvariable in einer multiplen Regression mit der unabhängigen Variablen als signifikanter Prädiktor der abhängigen Variablen bestehen bleiben. Die Wirkung der Prädiktorvariablen kann dabei vollständig, aber auch nur teilweise über die Mediatorvariable (= partielle Mediation) übertragen werden. In letzterem Fall wirkt die Prädiktorvariable teilweise über die Mediatorvariable und teilweise auf direktem Weg. Abbildung 6-5 veranschaulicht modellhaft, wie die Variablen in dieser Studie in Beziehung zueinander stehen könnten.

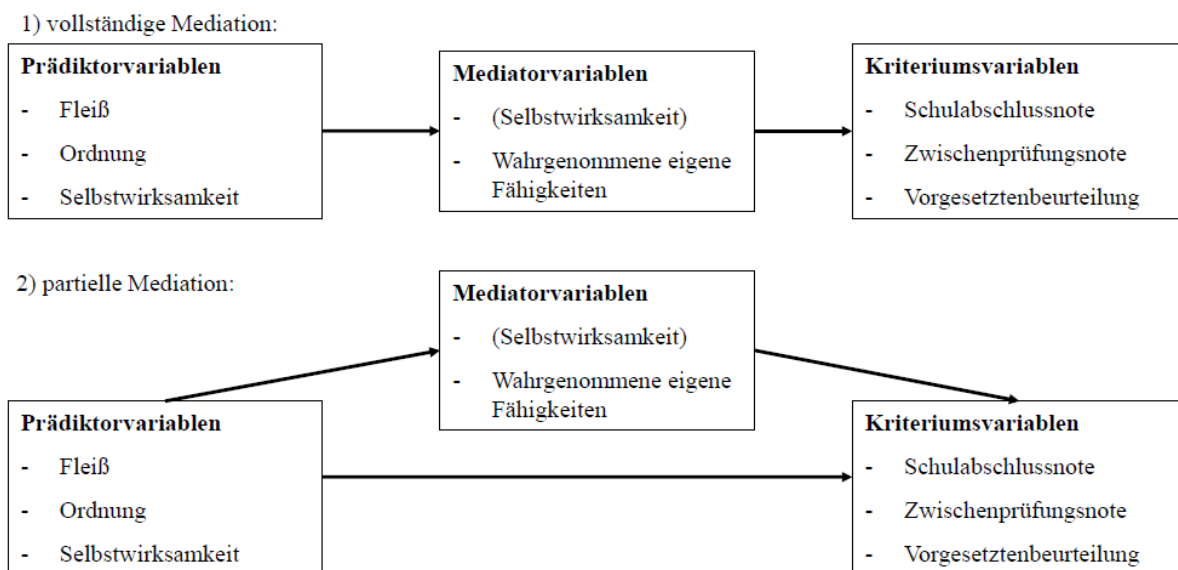


Abbildung 6-5: Darstellung möglicher Mediationsbeziehung/en im Untersuchungsmodell

Zur Überprüfung der Mediationsbeziehungen soll in dieser Studie die *Causal-Steps-Methode* nach Baron und Kenny (1986) angewendet werden, da diese auch eine Untersuchung partieller Mediationsbeziehungen zulässt.

Bei der Causal-Steps-Methode werden zuerst einfache Regressionen (s. 6.3.2) gerechnet, um zu überprüfen, inwiefern ein signifikanter Zusammenhang zwischen Prädiktor und Prognosevariable besteht. Die Mediationsvariablen bleiben in diesem ersten Schritt noch unberücksichtigt. In einem zweiten Schritt wird dann mittels Korrelationsanalysen (s. 6.3.1) der Zusammenhang zwischen Prädiktor und Mediator ermittelt. Diese muss signifikant sein. Im dritten Schritt wird dann eine Regression vom Prädiktor (X) und Mediator (Med) auf die Kriteriumsvariable (Y) mit der Gleichung

$$Y = a + b_1 \cdot X + b_2 \cdot Med + e$$

berechnet. Der Koeffizient b_2 muss signifikant sein, um von einer Mediation ausgehen zu können. Der vierte und letzte Schritt ist die Überprüfung, ob es sich um eine vollständige oder partielle Mediation handelt. Dies kann man am Wert des Koeffizienten b_1 ablesen. Wenn $b_1 = 0$ ist, handelt es sich um eine vollständige Mediation. Ist $b_1 \neq 0$, handelt es sich um eine partielle Mediation (Müller, 2007).

Das Ausmaß einer Mediation wird als „Reduzierung des Effektes der Prädiktorvariable auf die Prognosevariable“ definiert (Müller, 2007, S. 257). Allerdings gilt es zu beachten, dass das Ausmaß, mit dem der Effekt von X auf Y reduziert wird, nicht gleichzeitig bedeutet, dass es auch zu einer Änderung der erklärten Varianz oder der F - oder p -Werte kommt. Das F für

die Beziehung von Prädiktor- und Kriteriumsvariable kann durchaus stark sinken, selbst wenn der Mediator keinen Einfluss auf den Output ausübt (Müller, 2007).

Der Sobel-Test nach Baron und Kenny (1986) überprüft die Gültigkeit der Mediationsanalyse. Beim Sobel-Test handelt es sich im Prinzip um einen spezialisierten t -Test, der überprüft, inwiefern die Abnahme des Effektes des Prädiktors (Präd.) durch die Hinzunahme des Mediators (Med.) signifikant ist. Wenn dieser Effekt signifikant ist, handelt es sich um eine Mediation. Mittels

$$z = \frac{B_{\text{Präd.}} \cdot B_{\text{Med.}}}{\sqrt{(B_{\text{Med.}}^2 \cdot SE_{B_{\text{Präd.}}}^2 + B_{\text{Präd.}}^2 \cdot SE_{B_{\text{Med.}}}^2)}}$$

wird überprüft, inwiefern der indirekte Pfad der unabhängigen Variablen über den Mediator das Kriterium signifikant ist. Die Regressionskoeffizienten von Prädiktor und Mediator sind hierbei mit B beschrieben, die zugehörigen Standardfehler mit $SE B$. Sofern $z > 1,96$ ist, werden die z -Werte mit $p < .05$ assoziiert. In diesem Falle wird der Sobel-Test signifikant. Eine Mediation liegt vor.

Allerdings tendiert der Sobel-Test dazu, nur in großen normalverteilten Stichproben signifikant zu werden. Dies ist der Fall, da der Test voraussetzt, dass $B_{\text{Präd}}$ und B_{Med} jeweils normalverteilt sind, was in kleineren Stichproben nicht so leicht erreicht wird.

6.4 Vergleich der AGS mit der Gewissenhaftigkeitsskala des NEO-FFI

Um die Vorhersagekraft der AGS mit einem allgemeinen Gewissenhaftigkeitsmaß vergleichen zu können, soll im Folgenden die inkrementelle Validität der AGS bzw. der Gewissenhaftigkeitsaspekte Fleiß und Ordnung über die Vorhersagekraft der Gewissenhaftigkeitsdimension des NEO-FFI (Borkenau & Ostendorf, 1993, 2008) hinaus ermittelt werden. Für die arbeitsbezogen formulierte Skala wird dementsprechend ein besserer Model-Fit erwartet.

Die inkrementelle Validität ist der (prozentuale) Zuwachs der prädiktiven Validität, der durch Hinzunahme eines weiteren Prädiktors zu einer Messung resultiert, bzw. der Beitrag eines Tests zur Verbesserung der Vorhersage eines Kriteriums über einen anderen Test hinaus (Bühner, 2011). Zur Überprüfung der Varianzaufklärung und der inkrementellen Validität der dispositionellen Variablen sollen hierarchische Regressionsanalysen (s. 6.3.2) berechnet werden.

Eine Variable besitzt inkrementelle Validität, wenn ihre Aufnahme in die Regressionsgleichung den Anteil der aufgeklärten Varianz (R^2) am Kriterium erhöht, d.h. die Vorhersage verbessert.

6.5 Weitere angewandte Testverfahren

Zur Überprüfung des Einflusses der Kontrollvariablen wurden neben den oben bereits beschriebenen Korrelationsanalysen weitere Analysemethoden herangezogen, die im Folgenden kurz dargestellt werden sollen. Hierzu zählen t -Tests (6.5.1), Levene-Tests (6.5.2), Tests auf Normalverteilung (6.5.3) sowie Multikollinearitätsanalysen (6.5.4).

6.5.1 t -Tests

Zur Überprüfung, inwiefern die Studienteilnehmer sich in bestimmten Merkmalen signifikant voneinander unterscheiden, sollen t -Tests berechnet werden, beispielsweise zur Feststellung möglicher Geschlechterunterschiede. Ein t -Test prüft anhand der Mittelwerte (M_x) aus zwei verbundenen oder unabhängigen Stichproben, inwiefern diese Mittelwerte sich signifikant

voneinander unterscheiden oder gleich sind. Dies geschieht mittels der sogenannten *t*-Statistik:

$$t = \frac{M_1 - M_2}{s(M_1 - M_2)}$$

Hierbei wird die Bedeutsamkeit der Abweichung der beiden Mittelwerte bzw. deren Differenz in Relation zur Streuung der Mittelwertdifferenz ($s(M_1 - M_2)$) berechnet (Bortz, 1993).

Voraussetzung für die Berechnung eines *t*-Tests ist, dass beide Stichproben aus normalverteilten Grundgesamtheiten mit gleichen Varianzen stammen. Die Gleichheit der Varianzen kann mittels Levene-Test (s. Abschnitt 6.5.2) überprüft werden, während die Normalverteilungsannahme mittels Kolmogorov-Smirnov- oder Saphiro-Wilk-Test (Abschnitt 6.5.3) erfolgen kann.

Wird der *t*-Test signifikant, deutet dies auf Unterschiede in den Stichproben hin.

6.5.2 Levene-Test

Der Levene-Test wird zur Überprüfung der Varianzhomogenität zweier Stichproben verwendet. Dieser Test zur Überprüfung der Gleichheit der Varianzen ist relativ unempfindlich gegenüber einer Verletzung der Normalverteilungsvoraussetzung. Als Prüfwert wird ein *F*-Wert berechnet, mit dem die Wahrscheinlichkeit angegeben ist, mit der eine Gleichheit der Varianzen in der Grundgesamtheit vorliegt. Wird das Ergebnis des Levene-Tests nicht signifikant ($p > .20$), so wird die Nullhypothese beibehalten, dass sich die Varianzen nicht signifikant voneinander unterscheiden.

6.5.3 Tests auf Normalverteilung

Zur Überprüfung der Normalverteilung der untersuchten Variablen stehen in der SPSS-Software der *Kolmogorov-Smirnov-Test* und der *Saphiro-Wilk-Test* zur Verfügung. Der Kolmogorov-Smirnov-Test überprüft die Nullhypothese bzw. Annahme, eine Normalverteilung läge den Daten zugrunde. Die *Irrtumswahrscheinlichkeit* p wird auch als *Asymptotische Signifikanz* bezeichnet. Um einen Beta-Fehler zu vermeiden, wird das Signifikanzniveau entsprechend auf $p < .01$ hochgesetzt. Wird der Kolmogorov-Smirnov-Test signifikant bzw. ist die Asymptotische Signifikanz kleiner als $p < .10$, muss man von einer Abweichung von der Normalverteilung ausgehen. In diesem Falle sollten im Folgenden keine Tests durchgeführt werden, die eine Normalverteilung der Daten voraussetzen. Der Saphiro-Wilk-Test ist für kleinere Stichproben ($N < 50$) besser geeignet, kann aber auch in größeren Stichproben eingesetzt werden. Wenn $p > .05$ ist, ist von normalverteilten Daten auszugehen. Wenn $p < .05$ ist, weicht die Verteilung der Daten signifikant von einer Normalverteilung ab.

6.5.4 Multikollinearitätsanalysen

Eine Überprüfung der bzw. eine nicht-vorliegende Multikollinearität ist eine Voraussetzung für die Interpretierbarkeit der Ergebnisse aus Regressionsanalysen. Multikollinearität kann die Ergebnisse einer multiplen Regression in dreifacher Weise beeinträchtigen. Zum einen gefährdet extreme Multikollinearität die rechnerische Genauigkeit der Regressionsgewichte, was zu einer erschwerten Interpretierbarkeit der *b*-Gewichte führen kann. Zum anderen kann Multikollinearität außerdem eine Verzerrung der Teststatistiken verursachen (Bortz, 1993).

Zum Nachweis der Multikollinearität werden in der Literatur verschiedene Ansätze aufgeführt. Die am häufigsten verwendete Methode ist eine Überprüfung der bivariaten Korrelationen zwischen den Prädiktoren. Hohe Interkorrelationen ($r > .80$) stellen einen Hinweis auf das Vorliegen von Multikollinearität dar (Bühner, 2011).

6.5.5 Teststärke

Die *Teststärke* bzw. *Power* eines Tests gibt die Wahrscheinlichkeit ($P = 1 - \beta$) an, mit der ein Signifikanztest bei Gültigkeit der H_1 auch zu einem signifikanten Ergebnis führt (Bortz & Döring, 1995). Die Wahrscheinlichkeit gibt an, inwiefern die in der Stichprobe vorhandenen Unterschiede durch ein statistisches Verfahren aufgedeckt werden können. Mit zunehmendem Stichprobenumfang wächst die Power eines Tests. Als akzeptable Power wird ein Wert von $P = .80$ angestrebt. Die zugehörigen optimalen Stichprobenumfänge zum Erhalt einer guten Power für verschiedene Signifikanztests können für einen näherungsweisen Indikator in Tabellen nachgesehen werden (z.B. Bortz & Döring, 1995, S. 575). Die exakte Berechnung von Teststärken bzw. optimalen Stichprobenumfängen ist mit dem Programm G*Power 3 von Faul, Erdfelder, Lang und Buchner (2007) möglich.

Poweranalysen sowie die Ermittlung optimaler Stichprobenumfänge sollen in der vorliegenden Studie verwendet werden, um die Aussagekraft der gefundenen Ergebnisse besser einordnen zu können. Vor allem die Aussagekraft der Korrelationskoeffizienten soll auf diese Weise evaluiert werden.

6.6 Nicht angewandte Testverfahren

Eine weitere Möglichkeit, die angenommenen Zusammenhänge zwischen den Untersuchungsvariablen zu überprüfen, wäre das Berechnen von Strukturgleichungsmodellen gewesen. Strukturgleichungsmodelle sind strukturprüfende multivariate Verfahren, die kausale Zusammenhänge testen bzw. schätzen.

Von einer Verwendung von Strukturgleichungsmodellen wurde abgesehen, da diese einen eher konfirmatorischen Charakter besitzen und die Überprüfung der Modellannahmen in dieser Untersuchung eher explorativen Charakter besaß. Die verwendeten Faktoren- und Regressionsanalysen können als Sonderfälle von Strukturgleichungsmodellen angesehen werden. Eine Strukturanalyse besteht im ersten Schritt aus einer Faktorenanalyse zur Bestimmung der Anzahl der latenten Variablen. Dies wurde in vorhergehenden Untersuchungen bereits durchgeführt (Kasper, in Vorbereitung; Struckmeyer, 2009). Eine nachfolgende CFA bestätigt das Messmodell in einem zweiten Schritt. Da dies bereits im Rahmen der Skalenüberprüfung der AGS weitgehend geschieht, ist dieser Schritt ebenfalls nicht mehr notwendig. Der dritte Schritt beinhaltet eine Testung des Strukturmodells sowie eine nachfolgende Testung verschiedener Vorhersagemodelle. Dies soll in der vorliegenden Untersuchung mittels verschiedener Moderations- und Mediationsanalysen geschehen.

Da man sich bei der Verwendung von Strukturgleichungsmodellen zudem im Vorwege bereits über die genauen Wirkweisen der gemessenen Indikatoren auf die Kriteriumsvariablen im Klaren sein sollte, erschien das Berechnen von Strukturgleichungsmodellen an dieser Stelle nicht als beste Methode. Die explorative Berechnung verschiedener Regressionsmodelle wurde daher vorgezogen.

7. Durchführung

In diesem Kapitel werden die Durchführung der Feldstudie sowie der Kontrollgruppenuntersuchungen beschrieben. Hierbei wird in den verschiedenen Abschnitten zwischen der Bewerberstichprobe (7.1), der Gruppe der daraufhin eingestellten Bewerber bzw. der Retest-Stichprobe (7.2) sowie der Kontrollgruppe (7.3) unterschieden. Es werden in jedem Abschnitt jeweils die Rekrutierung der Studienteilnehmer beschrieben, zentrale Merkmale der Stichproben bzw. Subgruppen genannt, der Ablauf der jeweiligen Testung geschildert und es wird auf Probleme bei der Rekrutierung und Testdurchführung eingegangen.

7.1 Bewerberstichprobe

Um möglichst genaue Aussagen über die Anwendbarkeit und Nützlichkeit der Skala im Auswahlkontext treffen zu können bzw. eine möglichst gute Übertragbarkeit der Ergebnisse zu gewährleisten, wurde das Ziel verfolgt, eine praxisnahe Stichprobe zu untersuchen. Dies wurde durch die Zusammenarbeit mit Moldzio & Partner – Institut für Personalauswahl und einem der Kunden der Unternehmensberatung ermöglicht.

7.1.1 Rekrutierung der Bewerberstichprobe

Als Untersuchungsteilnehmer fungierten Bewerber für eine kaufmännische bzw. technische Ausbildung bei einem großen deutschen Industrieunternehmen. Der Auswahlprozess des Unternehmens setzte sich aus einem mehrstufigen Verfahren zusammen. Von den ursprünglich insgesamt etwa 2.000 Interessenten, die sich direkt bei dem betreffenden Unternehmen per E-Mail oder Post um einen Ausbildungsplatz beworben hatten, wurden anhand der Zeugnisnoten und Bewerbungsanschreiben ca. 40 % der Bewerber zu einem von mehreren regelmäßig durchgeführten halbtägigen Gruppentests zur Vorauswahl (s. Abschnitt 7.1.3) eingeladen, die zwischen August 2009 und Februar 2011 stattfanden. Die Testtage waren vom Ablauf her für technische und kaufmännische Bewerber gleich, die verwendeten Testverfahren unterschieden sich jedoch teilweise. Die Vorauswahl der Bewerber hierfür erfolgte durch die Unternehmen selbst. Alle Teilnehmer dieser sog. „Testtage“ willigten vor Beginn der Testung ein, dass ihre Daten für wissenschaftliche Zwecke gespeichert und in anonymisierter Form weiterverwendet werden dürfen. Nach dem Testtag wurde eine Auswahl potenziell interessanter bzw. geeigneter Kandidaten zu einem sog. „Bewerbertag“ eingeladen, an dem die verbleibenden Bewerber durch die Ausbildungsverantwortlichen interviewt wurden und einige Gruppenaufgaben, wie z.B. eine Diskussion, bewältigen mussten. Anschließend an den Bewerbertag entschied sich das Unternehmen für ca. 45 Bewerber je Ausbildungsrichtung, denen jeweils ein Vertragsangebot zugesendet wurde. Die auf diese Weise eingestellten Bewerber des Testjahres 2009 bildeten die Grundgesamtheit der Retest-Stichprobe in der vorliegenden Untersuchung (s. Abschnitt 7.2).

7.1.2 Beschreibung der Bewerberstichprobe

Die Bewerberstichprobe ($N_{Bew} = 740$) setzt sich aus den Teilnehmern der kaufmännischen ($N_{BewK} = 397$) und technischen Testtage ($N_{BewT} = 343$) zusammen.

Insgesamt nahmen 278 (38 %) weibliche und 462 (62 %) männliche Bewerber an den Testtagen teil (s. Tab. 7-1). Betrachtet man die beiden Substichproben der unterschiedlichen Ausbildungsbereiche getrennt voneinander, gab es im kaufmännischen Bereich mit 236 (59 %) Mädchen und 161 (41 %) Jungen mehr weibliche als männliche Bewerber um einen Ausbildungsplatz. Im technischen Bereich bewarben sich mit 301 (88 %) männlichen Testteilnehmern demgegenüber deutlich mehr Jungen um eine Ausbildung als Mädchen ($n_{BewT} = 42$; 12 %).

Tabelle 7-1: Geschlechterverteilung in der Bewerberstichprobe

	Kaufleute	Techniker	Gesamt
weiblich	236	42	278
männlich	161	301	462
gesamt	397	343	740

Auch bezüglich der Altersstruktur unterscheiden sich die Substichproben. Im kaufmännischen Bereich waren die Teilnehmer durchschnittlich 18,18 ($SD= 1,27$) Jahre alt. Die Bewerber im technischen Bereich waren mit durchschnittlich 17,06 Jahren ($SD= 2,11$) vergleichsweise jünger. Abbildung 7-1 veranschaulicht die Verteilung der Altersstruktur der zwei Subpopulationen. Der jüngste Teilnehmer war zum Testzeitpunkt 14 Jahre alt, die älteste Teilnehmerin 34. Insgesamt waren die meisten Bewerber zwischen 15 und 21 Jahre alt.

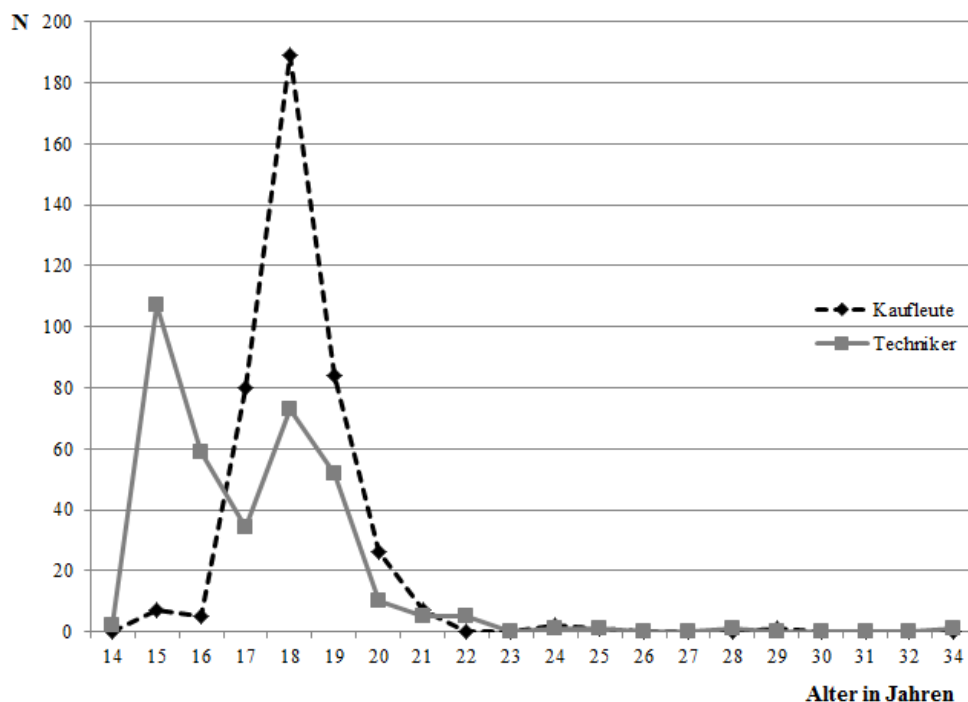


Abbildung 7-1: Vergleich der Altersverteilung der kaufmännischen und technischen Bewerber

Die unterschiedlichen Altersstrukturen der beiden Substichproben lassen sich anhand der verschiedenen Niveaus der Schulbildung der Bewerber erklären (vgl. Tab. 7-2). Es zeigte sich, dass die kaufmännischen Kandidaten insgesamt über ein höheres Bildungsniveau verfügten als die technischen Ausbildungsbewerber, somit länger zur Schule gegangen und folglich auch vergleichsweise älter waren.

Tabelle 7-2: Schulbildungsniveau in der Bewerberstichprobe

	Kaufleute	Techniker	Gesamt
Hauptschulabschluss	0	33	33
Realschulabschluss	13	160	173
Fachhochschulreife	158	32	190
Abitur	226	118	344
Gesamt	497	343	740

Im kaufmännischen Bereich strebten 297 Bewerber (70,3 %) eine Ausbildung als Industriekaufmann bzw. -kauffrau an, 45 interessierten sich für eine duale Ausbildung im Bereich Trade Management (11,3 %), 16 Kandidaten (4,0 %) bewarben sich als Informatikkaufleute und

weitere zwölf Kandidaten (3,0 %) strebten eine Ausbildung als Wirtschaftsinformatiker an. Sieben Bewerber (1,8 %) strebten eine Tätigkeit als Sozialversicherungsfachangestellter an, sechs Kandidaten (1,5 %) interessierten sich für einen dualen Studiengang „International Business Administration“ (IBA) und vier Teilnehmer (1,0 %) bewarben sich um eine duale Ausbildung im Bereich Betriebswirtschaft. Jeweils ein Bewerber (0,1 %) interessierte sich für eine Ausbildung zum Wirtschaftsingenieur bzw. als Werbekauffrau/ -mann. Die restlichen 26 Bewerber (6,5 %) konnten keiner der genannten Kategorien zugeordnet werden und wurden in der Kategorie „Sonstige“ zusammengefasst.

Im technischen Bereich bewarben sich die meisten Testteilnehmer um eine Tätigkeit als Elektroniker für Betriebstechnik ($n=93$; 27,1 %). Mit 79 Bewerbern (23,0 %) interessierten sich ebenfalls viele für eine Ausbildung zum Industriemechaniker. 50 Kandidaten (14,6 %) bewarben sich als Mechatroniker. Eine Ausbildung zum Zerspanungsmechaniker wollten 22 Teilnehmer (6,4 %) beginnen, 17 Bewerber (5,0 %) interessierten sich für eine Ausbildung zum Elektroniker und 12 Bewerber (3,5 %) strebten eine Tätigkeit als Werkzeugmechaniker an. Jeweils zwei Kandidaten (0,6 %) bewarben sich als potenzielle Elektroanlagenmonteure bzw. Fachinformatiker. Im Bereich der dualen Studiengänge strebten 30 Bewerber (8,7 %) ein Studium im Bereich Maschinenbau an, acht Personen (2,3 %) wollten Elektrotechnik studieren. Fünf Teilnehmer (1,5 %) interessierten sich darüber hinaus für eine Ausbildung zur Fachkraft für Lebensmitteltechnik. Insgesamt 23 Bewerber (6,7 %) konnten keiner der genannten Berufsbilder zugeordnet werden und wurden unter der Kategorie „Sonstige“ zusammengefasst.

7.1.3 Ablauf der Bewerbertestungen

Die Bewerber wurden nach der Analyse der Bewerbungsunterlagen durch das jeweilige Unternehmen per Brief und/ oder E-Mail zu einem Testtag eingeladen. Die Testunterlagen wurden vor dem Testtermin durch die Unternehmensberatung Moldzio & Partner – Institut für Personalauswahl vorbereitet und zur Verfügung gestellt. Die Testungen selbst wurden vor Ort bei dem auswählenden Unternehmen von einer durch Moldzio & Partner in der Testdurchführung geschulte Fachkraft durchgeführt. Je nach Ausbildungsbereich gab es zwei verschiedene Testpakete bzw. Testabläufe.

Die Gruppenstärke variierte je nach Termin von einer bis hin zu 38 getesteten Personen je Testtag. Die Testungen dauerten jeweils ca. drei Stunden.

Die Bewerber für eine technische Ausbildung bearbeiteten zuerst Ausschnitte aus verschiedenen Leistungstests. Hierzu zählten die Skalen 3 und 4 sowie 7 bis 10 des Leistungsprüfungssystems (LPS; Horn, 1983) sowie das Arbeitsblatt 4 des Intelligenz-Struktur-Tests 2000 (IST-2000; Amthauer et al., 1999). Anschließend waren aus einem berufsbezogenen Leistungstest für technische und handwerkliche Fähigkeiten (AZUBI-TH; Görlich & Schuler, 2007) die Untertests „Technisches Verständnis“, „Berechnen von Grundrissen“ sowie „Berichtskorrektur“ zu bearbeiten. Den Abschluss der Testbatterie bildeten zwei Persönlichkeitstests, der NEO-FFI (Borkenau & Ostendorf, 2008) und die AGS.

Das kaufmännische Testpaket bestand ebenfalls aus verschiedenen Untertests des LPS (Skalen 3, 4 und 8) sowie des IST-2000 (Aufgabenblätter 4 bis 6). Außerdem bearbeiteten die Bewerber ein Lückendiktat zur Erfassung ihrer Rechtschreibkenntnisse (Rechtschreibtest - Neue Rechtschreibregelung; Bulheller & Häcker, 2001). Den Abschluss der Testbatterie bildeten wiederum die Persönlichkeitsinventare NEO-FFI und AGS.

Die bearbeiteten Testpakete wurden durch Moldzio & Partner – Institut für Personalauswahl ausgewertet und in SPSS eingegeben. Eine Rückmeldung der Testergebnisse an das Unternehmen sowie die getesteten Bewerber erfolgte ca. eine Woche nach Testeingang.

7.1.4 Probleme bei der Rekrutierung und Testdurchführung

Da es sich bei der Bewerberstichprobe um eine Feldstichprobe handelte, wurde die Rekrutierung dieser Stichprobe ausschließlich durch das auswählende Unternehmen vorgenommen, bei dem auch die Testungen im Rahmen der Auszubildendenauswahl durchgeführt wurden. Die zuständigen Personalverantwortlichen berichteten von keinerlei systematischen Problemen bei der Auswahlprozedur. Lediglich in Ausnahmefällen seien einige der eingeladenen Bewerber nicht zum Auswahltest erschienen. Die Vor- und Nachteile von Felduntersuchungen in Bezug auf die Besonderheiten der vorliegenden Studie werden in Abschnitt 9.1 („Bewertung des methodischen Vorgehens“) detailliert besprochen.

Auch während der Testungen gab es keine Probleme oder Zwischenfälle. Lediglich in zwei Fällen kam es bei der Bearbeitung der Leistungstestskalen zu Verwechslungen von Testmaterialien, was jedoch bei der weiteren Auswertung der Testunterlagen berücksichtigt wurde. Die Verwechslungen führten jedoch bei keinem der Kandidaten zu einer signifikanten Verzerrung der Gesamtestleistung, weshalb dies im weiteren Verlauf der Untersuchung nicht weiter berücksichtigt wird. Für die vorliegende Untersuchung der Persönlichkeitseigenschaften Gewissenhaftigkeit bzw. Fleiß und Ordnung sowie Selbstwirksamkeit hatte dies allerdings keine Relevanz. Die entsprechenden Datensätze wurden daher mit verwendet.

Insgesamt gab es während der Testungen nur wenige Nachfragen oder Störungen. Dies ist einerseits den gut geschulten Testleitern und andererseits der – sicherlich – hohen persönlichen Relevanz der Auswahl situation für die Testteilnehmer zuzuschreiben, in der sich jeder Bewerber versucht, von seiner besten Seite zu zeigen und nicht „unangenehm“ aufzufallen.

7.2 Retest-Stichprobe

Um Aussagen über die Stabilität der Merkmale Ordnung und Fleiß über einen längeren Zeitraum tätigen zu können, das theoretische Modell bzw. die Vorhersagekraft der AGS überprüfen zu können sowie einige, zum Testzeitpunkt noch nicht erhobene, Kontrollvariablen zu erfassen, wurde ein Teil der Bewerberstichprobe nach ca. einem Jahr noch einmal mit einem Testpaket befragt. Diese zweite Erhebung wird in den folgenden Abschnitten beschrieben.

7.2.2 Beschreibung der Retest-Stichprobe

Insgesamt nahmen 90 Auszubildende an der Nachtestung bei dem Hausgerätehersteller teil. Von den bearbeiteten Fragebogenpaketen konnten allerdings lediglich 63 Fragebogenpakete der Zweitbefragung eindeutig den Daten der Ersterhebung (Bewerberstichprobe) zugeordnet werden. Alle folgenden Berechnungen und Analysen der Retestanalyse beziehen sich daher auf diese 63 Datensätze.

Die Retest-Stichprobe setzte sich aus 31 technischen (49,2 %) und 32 kaufmännischen Auszubildenden (50,8 %) zusammen. Beide Ausbildungsbereiche waren somit in etwa gleich stark in der Retest-Stichprobe vertreten. Insgesamt nahmen 18 weibliche (28,6 %) und 45 männliche Auszubildenden (71,4 %) an der Nachbefragung teil. Der geringere Anteil der weiblichen Testteilnehmer ist in dem vergleichsweise niedrigen Anteil weiblicher Auszubildenden im technischen Bereich begründet (siehe Tabelle 7-3).

Tabelle 7-3: Geschlechterverteilung in der Retest-Stichprobe

	Kaufleute	Techniker	Gesamt
Weiblich	16	2	18
Männlich	16	29	45
Gesamt	32	31	63

Die Teilnehmer der Nachbefragung waren durchschnittlich 18,28 Jahre alt (SD 1,89). Abbildung 7-2 veranschaulicht die Altersgruppenverteilung der Retest-Stichprobe getrennt nach Ausbildungsbereichen. Der Trend, dass die technischen Auszubildenden vergleichsweise jünger als die Kaufleute waren, setzte sich in dieser Substichprobe fort und war sogar noch deutlicher erkennbar.

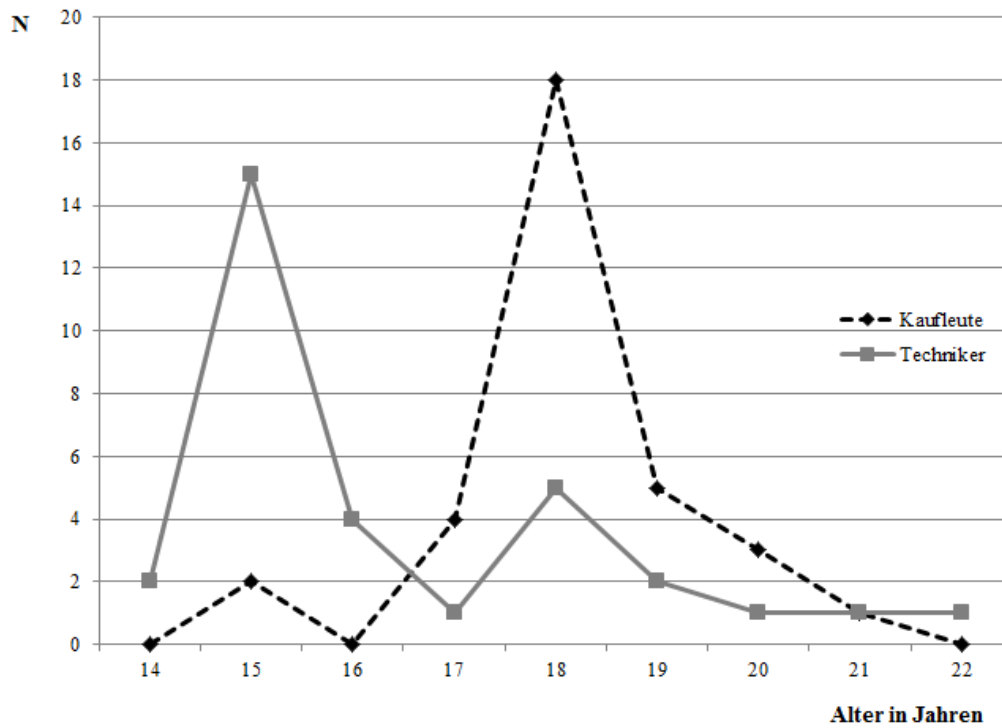


Abbildung 7-2: Vergleich der Altersverteilung der kaufmännischen und technischen Auszubildenden (Retest-Stichprobe)

Die Ursache für diesen Alterseffekt kann man, wie unter 7.1.2 bereits erläutert, in den unterschiedlichen schulischen Bildungshintergründen der Auszubildenden vermuten. Tabelle 7-4 verdeutlicht, dass 93,7 % der kaufmännischen Auszubildenden Abitur oder Fachabitur gemacht haben, während im technischen Bereich lediglich 41,9 % der Auszubildenden über ein Fachabitur oder Abitur verfügten. Die Mehrheit der Auszubildenden in diesem Bereich hatte einen Realschulabschluss (54,8 %).

Tabelle 7-4: Bildungsniveau der Retest-Stichprobe

	Kaufleute	Techniker	Gesamt
Hauptschulabschluss	0	1	1
Realschulabschluss	2	17	19
Fachhochschulreife	13	4	17
Abitur	17	9	26
Gesamt	32	31	63

Bezüglich des Ausbildungsberufes wurde seitens des Unternehmens nicht stringent zwischen Kaufleuten und Technikern unterschieden. So fanden sich in einigen Ausbildungsberufen sowohl Kandidaten wieder, die am kaufmännischen, als auch solche, die am technischen Testtag teilgenommen hatten. Die Mehrheit der Getesteten gab an, eine Ausbildung zum Elektroniker für Betriebstechnik zu absolvieren ($N= 37$; 58,7 %). Diese Gruppe setzte sich aus 28 Kaufleuten und neun Technikern zusammen. Fünf technische (7,9 %) und ein kaufmännischer Auszubildender (1,6 %) gaben an, zum Industriemechaniker ausgebildet zu werden. Jeweils drei Auszubildende im technischen Bereich (4,8 %) wurden eigenen Angaben zufolge

zum Zerspanungs- oder Werkzeugmechaniker bzw. Elektroanlagenmonteur ausgebildet. Weiterhin gaben drei technische Testteilnehmer (4,8 %) an, einen dualen Studiengang im Bereich Engineering zu besuchen, zwei weitere Techniker (3,2 %) studierten im Bereich Elektrotechnik. Als Elektroniker ließen sich zwei Kaufleute (3,2 %) ausbilden. Zwei technische (3,2 %) und ein kaufmännischer Auszubildender (1,6 %) konnten keiner der Kategorien zugeordnet werden. Eine Testperson (1,6 %) machte keine Angaben über ihren Ausbildungsberuf.

37 der Testteilnehmer (58,7 %) gaben an, über berufliche Vorerfahrung zu verfügen, während 23 Auszubildende (36,5 %) eigenen Angaben zufolge keinerlei berufliche Vorerfahrungen hatten. Drei Testteilnehmer (4,8 %) machten zu ihrer beruflichen Vorerfahrung keine Angaben. Tabelle 7-5 veranschaulicht die berufliche Vorerfahrung der Auszubildenden.

Tabelle 7-5: Berufliche Vorerfahrung in der Retest-Stichprobe

	Kaufleute	Techniker	Gesamt
ohne berufliche Vorerfahrung	9	14	23
mit beruflichen Vorerfahrung	23	14	37
keine Angaben	0	3	3
gesamt	32	31	63

Nach ca. eineinhalb Jahren Ausbildung wurden durch die Ausbildungsbeauftragten des Unternehmens die Leistungsdaten und Vorgesetztenbeurteilungen der Auszubildenden übermittelt. Nicht für alle Auszubildenden gab es jedoch Leistungsindikatoren. Bei den Kaufleuten lagen Berufsschulnoten und Vorgesetztenbeurteilungen von 31 Auszubildenden vor. Aufgrund des kaufmännischen Curriculums der zuständigen Industrie- und Handelskammer (IHK) hatten die Zwischenprüfungen jedoch noch nicht stattgefunden. Dementsprechend lagen keine Kriteriumsvariablen für diese Auszubildenden vor. Auch für einige der Auszubildenden, die sich im dualen Studium befanden, konnten keine Leistungsindikatoren ermittelt werden.

Im technischen Ausbildungsbereich hatten zwei Auszubildende bis zum Zeitpunkt der Nacherhebung ihre Lehre abgebrochen, ein dritter wurde in eine Zweigstelle in ein anderes Bundesland versetzt, sodass von diesem ebenfalls keine Daten verfügbar waren. Ein weiterer Auszubildender war zwischenzeitlich verstorben. Die Mehrheit der technischen Auszubildenden hatte die Zwischenprüfung abgelegt, sodass von jeweils 26 Technikern Angaben zur schriftlichen und praktischen Prüfungsnote vorlagen. In Bezug auf das Gesamtergebnis lagen sogar 30 Ergebnisse vor. Drei Auszubildende legen ihre Zwischenprüfung erst im Herbst 2012 ab. Die durchschnittliche Berufsschulnote lag von 35 Auszubildenden im technischen Bereich vor. Zudem wurden 31 Vorgesetztenbeurteilungen übermittelt. Auch im technischen Ausbildungsbereich gab es von denjenigen Auszubildenden, die ein duales Studium absolvierten, wiederum teilweise keine Leistungsindikatoren.

7.2.3 Ablauf der Testung

Die neu beginnenden Auszubildenden wurden durch das Unternehmen im September 2010 im Rahmen eines Ausbildungswochenendes miteinander bekannt gemacht sowie auf die anstehende Ausbildung vorbereitet. Im Rahmen einer allgemeinen Informationsveranstaltung wurden den Auszubildenden ein informierendes Anschreiben (s. Anhang B1) sowie das Testpaket zur Bearbeitung ausgehändigt. Die kaufmännischen und technischen Auszubildenden wurden hierbei getrennt voneinander befragt, wobei der Ablauf der Testung in beiden Stichproben gleich ablief. Die Auszubildenden bearbeiteten das Fragebogenpaket selbstständig und gaben dieses danach an die Testleiter zurück, die diese per Post gesammelt an die Autorin der vorliegenden Studie zurückschickten. Die Datenauswertung und Zuordnung der Testpakete der

Auszubildenden mit den Daten der Bewerberstichprobe wurde durch die Autorin der vorliegenden Studie vorgenommen.

7.2.4 Probleme bei der Rekrutierung, Testdurchführung und -auswertung

Die Rekrutierung der Stichprobe selbst war mit nur wenig Aufwand verbunden, da alle neu beginnenden Auszubildenden befragt werden konnten. Von den Test- bzw. Ausbildungsleitern wurden keine Probleme bei der Testdurchführung berichtet.

Allerdings konnten bei der Auswertung bzw. Zuordnung der Daten nicht alle Fragebogenpakete eindeutig einem Auszubildenden zugeordnet werden, da einige Teilnehmer der Befragung die erste Seite des Fragebogenpakets entfernten, auf der die Befragten unter anderem auch einige persönliche Daten, wie beispielsweise ihr Geburtsdatum und Geschlecht, angeben sollten. Über die Gründe für dieses Vorgehen einzelner Testkandidaten liegen keine Informationen vor. Um Unstimmigkeiten mit dem Betriebsrat des Unternehmens zu vermeiden, wurde jedoch von einer Ursachenklärung abgesehen.

Die Fragebogen bzw. Versuchspersonen, deren Fragebogenpakete nicht zugeordnet werden konnten, wurden aus den Analysen ausgeschlossen.

Ebenso konnten nur für insgesamt 70 der getesteten Auszubildenden Leistungs- und Erfolgsindikatoren ermittelt werden. Dies lag an den spezifischen Curricula der Berufsschulen bzw. der ausbildenden Fachhochschulen. Einige Auszubildende hatten zum Zeitpunkt der Erhebung der Kriteriumsdaten im Frühjahr 2012 ihre Zwischenprüfungen noch nicht abgelegt. Darüber hinaus kam es aufgrund von Lehrabbrüchen sowie eines Todesfalles zu Drop-Out in den Daten.

7.3 Kontrollgruppe

Da es sich bei der Bewerberstichprobe und der daraus resultierenden Retest-Stichprobe ausschließlich um Bewerber bzw. Auszubildende aus einem realen (Bewerbungs-) Setting handelte und die meisten Angaben aus Fragebögen auf Selbsteinschätzungen basierten, bestand die Möglichkeit, dass die Bewerber die Fragebögen anders beantwortet haben könnten, als sie es in einer anderen Situation getan hätten. Daher wurden zum Vergleich Schüler verschiedener Abschlussklassen, d.h. eine demografisch vergleichbare, unabhängige Stichprobe, herangezogen, die ein sehr ähnliches Fragebogenpaket wie die Auszubildenden in der Retest-Stichprobe bearbeitete. Dies war notwendig, um mögliche konfundierende Effekte wie beispielsweise sozial erwünschtes Antwortverhalten kontrollieren und die Fragebogen- und Leistungsdaten miteinander vergleichen zu können.

7.3.1 Rekrutierung der Kontrollgruppe

Zur Rekrutierung geeigneter Kontrollgruppen wurden gezielt Schul- bzw. Abteilungsleiter einiger Schulen im Großraum Hamburg angesprochen. Dies geschah mittels eines Anschreibens und einer detaillierteren Studienerläuterung (s. Anhänge B2 und B3), das die Schulleiter/-innen entweder per E-Mail oder persönlich ausgehändigt bekamen. Nach einigen Tagen wurden die Entscheidungsträger der jeweiligen Schulen telefonisch kontaktiert und um Rückmeldung gebeten. Auswahlkriterium für die Kontrollgruppe war, dass es sich bei den zu testenden Personen um Schüler aus Abschlussklassen handelte und die betreffenden Klassenlehrer sich mit der Durchführung der ca. einstündigen Testung sowie der Weiterleitung der Abschlussnoten der Schüler nach Schuljahresende einverstanden erklärten.

7.3.2 Beschreibung der Kontrollgruppe

Insgesamt nahmen 97 Schüler an der Kontrollgruppenuntersuchung teil. Hiervon kamen 25 Schüler (25,8 %) von einer Gemeinschaftsschule, wo sie die 13. Klasse besuchten und sich auf das Abitur vorbereiteten. Die restlichen Schüler stammten von einer Hamburger Stadtschule, auf der 23 (23,7 %) von ihnen eine Realschulklasse besuchten und die 49 weiteren Schüler (50,5 %) in einer Hauptschulklasse waren. Alle Schüler bereiteten sich auf ihren im gleichen Schuljahr anstehenden Schulabschluss vor.

Die Kontrollgruppe bestand aus 45 Schülerinnen (46,4 %) und 52 Schülern (53,6 %). Tabelle 7-6 beinhaltet die Geschlechterverteilung unterteilt nach Schulart.

Tabelle 7-6: Geschlechterverteilung in der Schüler-Kontrollgruppe

	Hauptschule	Realschule	Gymnasium	Gesamt
weiblich	18	11	16	45
männlich	31	12	9	52
gesamt	49	23	25	97

Die Geschlechterverteilung entspricht der allgemeinen Geschlechterverteilung an Schulen. In der Hauptschule waren mit 31 Schülern vergleichsweise mehr (63,3 %) männliche als weibliche Schüler (36,7 %). Die Gruppe der Realschüler bestand zu 47,8 % aus Schülerinnen und zu 52,2 % aus Schülern, was vergleichsweise ausgeglichen ist. In der Subgruppe der Gymnasiasten war die Geschlechterverteilung im Vergleich mit der Hauptschülergruppe umgekehrt. Hier waren die weiblichen Schülerinnen mit 16 Mädchen (64,0 %) im Vergleich zu den Jungen ($N=9$; 36,0 %) überrepräsentiert.

Ein Geschlechtereffekt spiegelte sich auch in den Durchschnittsnoten der Kontrollgruppe wider. Insgesamt gaben die Schüler an, im letzten Zeugnis einen Notendurchschnitt von 2,84 ($SD=0,54$) erreicht zu haben. Hier waren die Leistungen der Mädchen mit einem Notendurchschnitt von 2,63 ($SD=0,49$) um eine Drittel-Note besser als die der Jungen ($M=3,01$; $SD=0,52$).

Das durchschnittliche Alter der Schüler betrug 16,28 Jahre ($SD=1,73$). Das vergleichsweise geringe Durchschnittsalter ist mit dem großen Anteil der Hauptschüler in der Kontrollgruppe erklärbar. Abbildung 7-3 veranschaulicht die Altersverteilung der Kontrollgruppe. Es ist zu

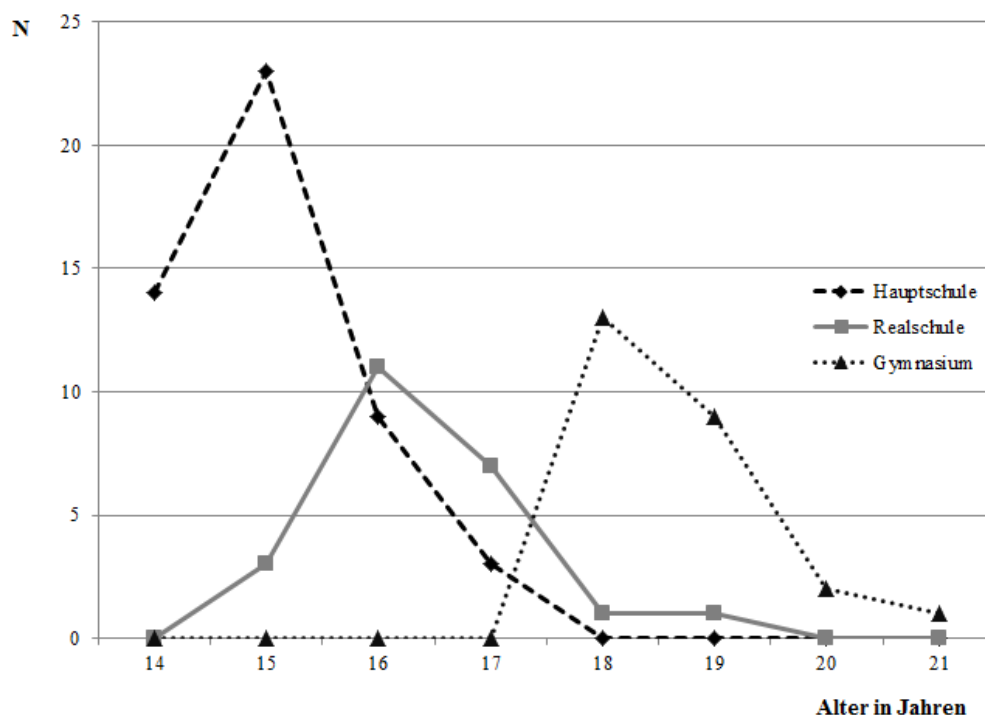


Abbildung 7-3: Vergleich der Altersverteilung in der Kontrollgruppe

erkennen, dass die meisten Hauptschüler 15 Jahre alt und somit die jüngsten Probanden in der Kontrollgruppe waren. Die Realschüler waren im Durchschnitt etwa ein Jahr älter als die Hauptschüler (16 Jahre) und die Gymnasiasten waren in der Regel zwischen 18 und 19 Jahre alt.

Die Mehrheit der Schüler ($N = 89$; 91,8 %) gab an, bereits berufliche Erfahrungen gemacht zu haben. Dabei unterschieden sich die Schulformen nur wenig. 44 (89,8 %) der Hauptschüler, 21 (91,3 %) der Realschüler und 24 (96,0 %) der Gymnasiasten hatten bereits im Rahmen eines Praktikums oder Aushilfsjobs Erfahrungen in der Berufswelt gesammelt.

Auf die Frage, ob sie bereits einen Ausbildungs- oder Berufswunsch hätten, antworteten 65 Schüler (67,0 %) mit „ja“. Tabelle 7-7 veranschaulicht, wie viele der Schüler bereits eine Idee haben, was sie beruflich nach der Schule machen möchten. Über alle Schulformen hinweg hatte die Mehrheit der Schüler bereits Vorstellungen darüber, was sie später beruflich einmal machen möchten.

Tabelle 7-7: Berufsvorstellungen in der Schüler-Kontrollgruppe

	Hauptschule	Realschule	Gymnasium	Gesamt
mit Berufswunsch	35	14	16	65
ohne Berufswunsch	14	9	9	32
gesamt	49	23	25	97

Die durchschnittlichen Abschlussnoten der Schüler wurden für die Gymnasialklasse am Schuljahresende im Sommer 2011 an die Versuchsleiterin übermittelt. Zu allen 25 Gymnasiasten waren Schulnoten vorhanden (100 %). Tabelle 7-8 gibt einen Überblick über die verfügbaren Kriteriumsdaten der Kontrollgruppe.

Tabelle 7-8: Überblick über die verfügbaren Kriteriumsdaten in der Schüler-Kontrollgruppe

	Hauptschule	Realschule	Gymnasium	Gesamt
getestet	49	23	25	97
Schulabschlusszeugnisse	23	21	25	69
fehlend	26	2	0	28

Die Abschlusszeugnisse Schülerstichprobe der Stadtteilschule wurden von der Testleiterin nach Abschluss des Schuljahres 2010/ 2011 im Sommer 2011 eigenhändig kopiert. Im Laufe des Schuljahres hatten zwei Schüler die Realschulklasse verlassen, sodass in dieser Subgruppe den ursprünglich 23 getesteten Schülern 21 Zeugnisse bzw. Schulabschlussnoten (91,3 %) zugeordnet werden konnten. In der Subgruppe der Hauptschüler war es im Laufe des Schuljahres zu einer größeren Fluktuation gekommen. Hier konnten am Ende des Schuljahres lediglich 23 Abschlusszeugnisse (46,9 %) den ursprünglich 49 getesteten Schülern zugeordnet werden.

7.3.3 Ablauf der Kontrollgruppentestung

Zur Testung der Kontrollgruppen wurden im Vorwege Termine mit den betreffenden Klassenlehrern abgestimmt. Diese fanden im Oktober und Dezember 2010 statt.

Die Testung wurde jeweils durch den Klassenlehrer in der Klasse angekündigt, der in der Regel während der Testung im Hintergrund anwesend blieb. Lediglich die Klassenlehrerin der Abiturienten verließ während der Testung die Klasse. Nach der Einleitung durch die betreffende Lehrkraft, stellte sich die Versuchsleiterin vor und erklärte den Zweck bzw. die Ziele der Untersuchung. Die Instruktion der Schüler erfolgte mündlich, war aber auch auf der ersten Seite des Testpaketes nachzulesen (s. Anhang B4). Den Schülern wurde im Rahmen der In-

struktion zuerst ein kurzer Überblick über Verfahren und Instrumente zur Auswahl von Auszubildenden gegeben. Dann wurde der Zweck der Studie erläutert und um die Mithilfe der Schüler geworben. Es wurde betont, dass die Testung gleichzeitig auch eine Möglichkeit für die Schüler darstellt, einen praktischen Einblick in persönlichkeitsdiagnostische Auswahlverfahren zu erhalten. Anschließend füllten die Schüler eine Codierungs-Liste (s. Anhang B5) aus, die der anonymisierten Zuordnung ihrer Fragebogenpakete mit den Abschlussnoten am Ende des Schuljahres diene.

Im Anschluss daran bearbeiteten die Schüler individuell die Fragebogenpakete. Zwischenfragen waren erlaubt und wurden von der Versuchsleiterin beantwortet.

Für die Beantwortung der Fragebogenpakete benötigten die Schüler zwischen 15 und 35 Minuten. Die Haupt- und Realschüler benötigten für die Bearbeitung des Fragebogenpakets deutlich mehr Zeit als die Abiturienten. Auch für diese Beobachtung könnten die geringeren Deutschkenntnisse der Haupt- und Realschüler als Erklärungsansatz herangezogen werden. Demgegenüber schienen sich diese Schüler allerdings augenscheinlich auch sehr ernsthaft mit den Fragen auseinander zu setzen, worin auch die längere Bearbeitungsdauer begründet sein könnte. Insgesamt wird die längere Bearbeitungszeit jedoch nicht als kritisch angesehen, da alle Versuchspersonen in der Kontrollgruppe das Fragebogenpaket augenscheinlich ernsthaft bearbeiteten.

Die fertig ausgefüllten Fragebögen wurden von den Versuchsleitern eingesammelt und die Übereinstimmung der Codes auf der Liste mit denen auf den Fragebögen nochmals überprüft. Die Codierungslisten wurden anschließend den Klassenlehrern übergeben, die diese bis zum Ende des Schuljahres im Sommer 2011 verwahrten und zur Rückmeldung der Abschlussnoten der Schüler verwendeten.

7.3.4 Probleme bei der Rekrutierung und Testdurchführung

Allgemein erwies sich die Suche nach geeigneten Abschlussklassen als schwierig, da viele Schulen bzw. Lehrer darüber klagten, viele Forschungsanfragen zu bekommen, und anmerkten, dass durch die Teilnahme an außercurricularen Untersuchungen die ohnehin recht knappe Unterrichtszeit weiter verkürzt würde. Insbesondere für die Abschlussjahrgänge wurde seitens der Schulleitung und Lehrer der Bedarf an Unterricht betont und folglich der Ausfall von Unterrichtszeit in der Regel als Absagegrund für die Untersuchungsanfrage angeführt. Insbesondere die Rekrutierung geeigneter Abiturjahrgänge erwies sich aus diesen Gründen als schwierig. Aufgrund dieser Problematik sind in der Kontrollgruppe die Hauptschulklassen stärker repräsentiert als in der Bewerberstichprobe.

Bei der Organisation der Testungen erwies sich die Kommunikation mit einigen Schulleitern als schwierig. So geschah es, dass in einem Fall die zu testenden Klassen bzw. Klassenlehrer nicht über das Kommen der Testleiterin informiert wurden. Die betreffenden Testungen wurden in diesem Falle kurzfristig auf einen anderen Termin verschoben, sodass sich für die Durchführung der Untersuchung und die Datenqualität aus dieser Problematik heraus jedoch keine Nachteile ergaben.

Während der Testdurchführung kam es in einigen Klassen mehrfach zu Störungen durch die Schüler, die durch Ermahnungen seitens der Testleiterin oder Klassenlehrer unterbunden wurden.

Bei der Testung der Real- und Hauptschulklassen war auffällig, dass es vermehrt zu Nachfragen kam, die den Inhalt bzw. die semantische Bedeutung einzelner Fragebogenitems betraf. Insbesondere das Item 10 der AGS („*Andere beschreiben mich als einen sehr eifrigen Menschen.*“) war für die Schüler schwer zu verstehen, da diese die Bedeutung des Wortes „eifrig“ nicht kannten. Alle Nachfragen der Schüler wurden durch eine einfache Umschreibung des

Wortes bzw. der Bedeutung des Items durch die Testleiterin beantwortet. Hierdurch kam es jedoch zu einiger Unruhe in den Klassen. Das teilweise geringe Itemverständnis fiel auch den anwesenden Klassenlehrern auf, die die Ursache hierfür darin sahen, dass die betreffenden Worte nicht zum Wortschatz der getesteten Schüler gehörten. Ein weiterer Grund hierfür kann der Migrationshintergrund einiger Testpersonen sein. Viele der getesteten Haupt- und Realschüler waren augenscheinlich keine Deutsch-Muttersprachler oder mehrsprachig aufgewachsen.

Auffällig war auch der hohe Drop-Out der Hauptschüler (53 %). In der Subgruppe der Gymnasiasten kam es demgegenüber zu keinerlei fehlenden Daten. Zwar schaffte in dieser Gruppe einer der Schüler nicht das Abitur, dennoch gab es hierzu eine statistisch verwertbare Information. In der Subgruppe der Hauptschüler sowie zu den zwei fehlenden Realschülern gab es jedoch keine Informationen über Ursachen und Gründe für diese hohe Fluktuation. Aufgrund eines Sabbaticals des Schulleiters bzw. Ansprechpartners an der betreffenden Schule konnte dieser Problematik jedoch nicht weiter nachgegangen werden.

8. Ergebnisse

Die Ergebnisdarstellung in diesem Kapitel folgt dem in Abschnitt 6 beschriebenen Ablauf der Analysen. Im ersten Abschnitt (8.1) werden die vorbereitenden Datenanalysen dokumentiert. In Abschnitt 8.2 sind danach die Ergebnisse der Skalenüberprüfung der AGS beschrieben, bevor im Anschluss daran in Abschnitt 8.3 die Vorhersagekraft der Variablen Fleiß, Ordnung, Selbstwirksamkeit und wahrgenommene eigene Fähigkeiten in verschiedenen Modellen überprüft wird. Abschnitt 8.4 enthält die Ergebnisse der Überprüfung der inkrementellen Validität der AGS. In Abschnitt 8.5 werden die Ergebnisse der Überprüfung der Kontrollvariablen berichtet. Der letzte Abschnitt (8.6) fasst die Ergebnisse der vorliegenden Studie noch einmal zusammen.

8.1 Vorbereitende Datenanalysen

Vor der Überprüfung der Hypothesen mussten die vorliegenden Datensätze auf die verschiedenen Analysen vorbereitet werden. Neben der Rekodierung negativ gepolter Items betraf dies v.a. die Analyse und Behandlung der fehlenden Werte, welche in den folgenden Abschnitten detaillierter beschrieben ist.

8.1.1 Analyse und Behandlung fehlender Werte

Die folgenden drei Abschnitte beschreiben die Analyse und Behandlung fehlender Werte in den beiden Stichproben dieser Untersuchung. Zunächst werden die Fälle identifiziert, die einen zu hohen Anteil fehlender Werte aufweisen (Abschnitt 8.1.1.1) und folglich aus der Untersuchung ausgeschlossen werden müssen. Danach wird der Fehlendmechanismus dargestellt (Abschnitt 8.1.1.2), bevor im dritten Abschnitt (8.1.1.3) die Ersetzung der fehlenden Werte beschrieben wird.

8.1.1.1 Elimination von Fällen mit zu hohem Anteil fehlender Werte

In die Analysen zu den fehlenden Werten gingen jeweils sämtliche Selbsteinschätzungsbögen ein, die über ein mehrfach gestuftes Antwortformat verfügten (= NEO-FFI, Selbstwirksamkeitsskala, AGS, wahrgenommene eigene Fähigkeiten). Im Anhang (s. Appendix A) können detailliertere Statistiken zu den einzelnen Fragebögen eingesehen werden.

Tabelle 8-1 gibt einen Überblick darüber, wie die fehlenden Werte in der Bewerberstichprobe ($N_{Bew}=760$) verteilt waren. Insgesamt gab es in dieser Stichprobe 94 unvollständige Datensätze (12,3 %). Hiervon wies die Mehrheit der Fälle ($n_{Bew}=74$; 9,7 %) jedoch einen Fehleranteil von weniger als 5 % Missings auf. Mit 87,7 % war der Anteil der vollständig bearbeiteten Testverfahren relativ hoch. Lediglich 20 Kandidaten (2,6 %) wiesen mehr als 5 % fehlende Werte auf und wurden entsprechend von den weiteren Analysen ausgeschlossen. Die ausgeschlossenen Fälle sind im Anhang in Tabelle A-6 aufgelistet. Nach der Elimination der 20 kritischen Fälle verblieben $N_{Bew}=740$ Datensätze in der Bewerberstichprobe.

Auffällig war die Diskrepanz zwischen technischen und kaufmännischen Bewerbern bezüglich der fehlenden Werte. So wiesen die technischen Bewerber ($N_{Bew,t}=358$) mit insgesamt 60 unvollständigen Datensätzen (anteilig 16,8 %) im Gegensatz zu den kaufmännischen Bewerbern ($N_{Bew,k}=402$) mit insgesamt 34 nicht vollständig bearbeiteten Datensätzen (anteilig 8,5 %) deutlich mehr Missings auf. Dies spiegelte sich auch in den ausgeschlossenen Datensätzen wider. Von den 20 ausgeschlossenen Fällen entfielen 15 (75 %) auf die technischen Bewerber, während nur 5 (25 %) der kaufmännischen ausgeschlossen werden mussten.

Tabelle 8-1: Fehlende Werte Bewerberstichprobe

Anzahl fehlender Werte pro Fall	Anteil fehlender Werte pro Fall	<i>n</i> (absolute Häufigkeit)	Kumuliertes <i>n</i>	Gültige Prozente	Kumulierte Prozente
32	29,1 %	3	3	0,4 %	0,4 %
31	28,2 %	1	4	0,1 %	0,5 %
17	15,3 %	3	7	0,4 %	0,9 %
16	14,4 %	5	12	0,7 %	1,6 %
14	12,7 %	7	19	0,9 %	2,5 %
9	8,1 %	1	20	0,1 %	2,6 %
5	3,6 %	1	21	0,1 %	2,7 %
3	2,7 %	3	24	0,4 %	3,1 %
2	1,8 %	11	35	1,4 %	4,5 %
1	0,9 %	59	94	7,8 %	12,3 %
0	0,0 %	666	760	87,7 %	100,0 %

Anmerkungen: $N = 760$. Lies erste Zeile: Bei 3 Bewerbern (0,4 % der Stichprobe) liegt der Anteil der fehlenden Werte bei 29,1 %.

In der Kontrollgruppe ($N_{KG} = 97$) gab es insgesamt 16 unvollständige Datensätze. Keiner dieser Fälle überschritt jedoch die kritische 5 %-Grenze, sodass keiner der Datensätze aus den Analysen ausgeschlossen werden musste. Tabelle 8-2 gibt Auskunft über den Anteil der Missings in der Kontrollgruppe.

Tabelle 8-2: Fehlende Werte Kontrollgruppe

Anzahl fehlender Werte pro Fall	Anteil fehlender Werte pro Fall	<i>n</i> (absolute Häufigkeit)	Kumuliertes <i>n</i>	Gültige Prozente	Kumulierte Prozente
2	3,4 %	1	1	1,0 %	1,0 %
1	1,7 %	15	16	15,5 %	16,5 %
0	0,0 %	81	97	83,5 %	100,0 %

Anmerkungen: $N = 97$.

Auch in der Reteststichprobe ($N_R = 63$) wurde für die Daten der erneuten Befragung eine gesonderte Analyse der fehlenden Werte vorgenommen. Tabelle 8-3 gibt Auskunft über den Anteil der fehlenden Werte in der Retest-Stichprobe. Insgesamt gab es 7 unvollständige Datensätze (11 %), wovon sich allerdings keiner über der kritischen 5 %-Grenze befand. Für die weiteren Analyseschritte verblieb somit der gesamte Datensatz mit $N_R = 63$ Fällen.

Tabelle 8-3: Fehlende Werte Retest-Stichprobe

Anzahl fehlender Werte pro Fall	Anteil fehlender Werte pro Fall	<i>n</i> (absolute Häufigkeit)	Kumuliertes <i>n</i>	Gültige Prozente	Kumulierte Prozente
5	3,0 %	1	1	1,6 %	1,6 %
2	1,2 %	1	2	1,6 %	3,2 %
1	0,6 %	5	7	7,9 %	11,1 %
0	0,0 %	56	63	88,9 %	100,0 %

Anmerkungen: $N = 63$.

Fehlende Werte in den Leistungsdaten, z.B. fehlende Abschlussnoten bei den Schülern der Kontrollgruppe oder nicht mitgeteilte Berufsschulnoten der Auszubildenden konnten nicht ersetzt oder simuliert werden und wurden daher wie Drop-Out-Daten behandelt. Insgesamt lagen Berufsschulnoten von $n_{Bew} = 62$ Auszubildenden bzw. Schulabschlusszeugnisse von $n_{KG} = 69$ Schülern vor, die den entsprechenden Datensätze zugeordnet werden konnten.

8.1.1.2 Analyse des Fehlendmechanismus

Die Analyse des Fehlendmechanismus erfolgte zuerst durch Sichten der kritischen Fälle. Hierbei fiel auf, dass sämtliche Probanden in der Bewerberstichprobe, deren Anteil fehlender Werte höher als 5 % war, jeweils eine bzw. zwei Seiten im NEO-FFI überblättert bzw. gar nicht bearbeitet hatten. Nach der Löschung dieser 20 Personen konnte kein Zusammenhang mehr zwischen den Missings und den erhobenen Variablen gefunden werden. Auch für die gelöschte Person der Retest-Stichprobe traf dies zu. Eine Berechnung des MCAR-Tests war nicht möglich, da keine Variable mehr als 5 % Missings aufwies. Es kann also davon ausgegangen werden, dass ein MCAR- oder zumindest MAR-Mechanismus vorlag.

8.1.1.3 Ersetzungen fehlender Werte

Wie in Abschnitt 6.1.3 dargestellt, kam für die verbleibenden unvollständigen Datensätze die beschriebene Form der Mittelwertimputation zum Einsatz. Insgesamt verblieben im Bewerberdatensatz ($N_{Bew}=740$) 95 zu ersetzende Datenpunkte. In der Kontrollgruppe ($N_{KG}=97$) waren 17 Datenpunkte zu ersetzen. In der Reteststichprobe ($N_R=63$) mussten zwölf Datenpunkte ergänzt werden. In den verschiedenen Stichproben veränderten sich die Mittelwerte und Standardabweichungen durch die Mittelwertimputation nur geringfügig im erweiterten Nachkommastellenbereich. Der Test auf Vergleich der Mittelwerte wurde sowohl für die beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung als auch für das Konstrukt der Selbstwirksamkeit in keinem Fall signifikant. Somit kann davon ausgegangen werden, dass die Mittelwertimputation zu keinen gravierenden Messverzerrungen führte. Dem Anhang (s. Appendix, A1) sind hierzu detailliertere Analysen zu entnehmen.

8.2 Überprüfung der Skaleneigenschaften der AGS

Der erste Teil dieser Studie beinhaltete die psychometrische Überprüfung der AGS bzw. deren Items, die in den folgenden Abschnitten dargestellt wird. Zuerst werden in den folgenden drei Abschnitten die allgemeinen Itemeigenschaften (8.2.1) der beiden Subskalen der AGS beschrieben, bevor die Ergebnisse der Reliabilitäts- sowie Validitätsüberprüfungen (8.2.2 und 8.2.3) dargestellt werden.

8.2.1 Analysen auf Itemebene

Die erste Fragestellung zielte auf die Überprüfung der Items der beiden Gewissenhaftigkeitsskalen Fleiß und Ordnung ab. Die AGS sollten nach den Kriterien der Itemtrennschärfe und Itemschwierigkeit zwei Faktoren abbilden, die den definierten Gewissenhaftigkeitsaspekten Fleiß und Ordnung entsprechen (vgl. Hypothese 1a). Die Ergebnisse hierzu sind in den folgenden beiden Abschnitten dargestellt. Zuerst erfolgte eine Analyse der Itemeigenschaften der beiden Skalen Ordnung und Fleiß bezüglich allgemeiner deskriptiver Daten (8.2.1.1). Danach werden die Itemschwierigkeiten (8.2.1.2) und Trennschärfeindizes (8.2.1.3) getrennt für die Stichprobe der Bewerber ($N_{Bew}=740$) sowie die der Schüler ($N_{KG}=97$) dargestellt.

8.2.1.1 Deskriptiva

Die Itemmittelwerte, Standardabweichungen sowie minimale und maximale Antworten der Bewerberstichprobe ($N_{Bew}=740$) sowie der Kontrollgruppe der Schüler ($N_{KG}=97$; in Klammern dahinter) sind in Tabelle 8-4 aufgeführt. Es ist auffällig, dass die Itemmittelwerte der Bewerberstichprobe durchweg höher waren als die Itemmittelwerte der Kontrollgruppe. Für die Skala Fleiß betrug der Mittelwert in der Bewerberstichprobe ($N_{Bew}=740$) $M=3,07$, während dieser Skalenmittelwert in der Schülerkontrollgruppe ($N_{KG}=97$) mit $M=2,67$ deutlich geringer ausfiel. Bezüglich des Ordnungsaspektes zeigte sich ein ähnliches Muster. Der Ska-

lenmittelwert betrug in der Bewerberstichprobe ($N_{Bew}=740$) $M=3,12$, während er in der Kontrollgruppe der Schüler ($N_{KG}=97$) mit $M=2,77$ geringer ausfiel. Die Standardabweichungen betreffend zeigte sich in der Bewerberstichprobe ein etwas homogeneres Antwortverhalten.

Tabelle 8-4: Itemkennwerte der AGS

Item	<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>
Ordnung 1	3,00 (2,69)	.66 (.77)	0 (1)	4 (4)
Fleiß 2	3,22 (2,80)	.57 (.73)	0 (1)	4 (4)
Ordnung 3	3,22 (2,78)	.69 (1.10)	0 (0)	4 (4)
Fleiß 4	2,85 (2,34)	.61 (.84)	1 (0)	4 (4)
Ordnung 5	2,78 (2,43)	.79 (.95)	1 (0)	4 (4)
Fleiß 6	2,36 (2,26)	.67 (.78)	0 (0)	4 (4)
Ordnung 7	3,82 (3,33)	.42 (.92)	1 (1)	4 (4)
Fleiß 8	3,36 (2,88)	.56 (.83)	2 (1)	4 (4)
Ordnung 9 *	3,26 (2,56)	.79 (1.03)	0 (0)	4 (4)
Fleiß 10	2,85 (2,33)	.69 (.83)	0 (0)	4 (4)
Ordnung 11	3,30 (2,84)	.55 (.69)	1 (1)	4 (4)
Fleiß 12	3,12 (2,78)	.63 (.83)	1 (1)	4 (4)
Ordnung 13	3,16 (2,72)	.69 (.98)	1 (0)	4 (4)
Fleiß 14	3,33 (3,06)	.57 (.86)	1 (0)	4 (4)
Ordnung 15 *	3,32 (2,80)	.74 (1.24)	0 (0)	4 (4)
Fleiß 16	3,23 (2,73)	.68 (.93)	0 (0)	4 (4)
Ordnung 17 *	3,21 (2,78)	.84 (1.13)	0 (0)	4 (4)
Fleiß 18	3,27 (3,07)	.55 (.71)	1 (1)	4 (4)
Ordnung 19	2,89 (2,61)	1.03 (1.10)	0 (0)	4 (4)
Fleiß 20	2,80 (2,44)	.63 (.74)	1 (0)	4 (4)
Ordnung 21	2,85 (2,79)	.70 (.72)	1 (1)	4 (4)
Fleiß 22	2,98 (2,53)	.59 (.74)	1 (1)	4 (4)
Ordnung 23	3,23 (2,98)	.60 (.92)	0 (0)	4 (4)
Fleiß 24	3,26 (2,81)	.55 (.82)	1 (1)	4 (4)
Ordnung 25 *	3,02 (2,67)	.94 (1.18)	0 (0)	4 (4)
Fleiß 26	3,10 (2,76)	.64 (.88)	0 (0)	4 (4)
Ordnung 27	2,85 (2,53)	.70 (.91)	1 (1)	4 (4)
Fleiß 28	3,01 (2,64)	.67 (.78)	1 (0)	4 (4)
Ordnung 29	2,88 (2,98)	.80 (.96)	0 (0)	4 (4)
Fleiß 30	3,32 (2,67)	.59 (.73)	1 (1)	4 (4)

Anmerkungen: *M*= Itemmittelwert; *SD*= Standardabweichung; *Min*= Minimum; *Max*= Maximum; $N_{Bew}=740$ Ausbildungsbewerber bzw. $N_{KG}=97$ Kontrollgruppe. Die Werte der Kontrollgruppe stehen in Klammern; * umgepolte Items wurden vor der Analyse rekodiert.

So waren die mittleren Standardabweichungen im Durchschnitt etwas geringer ausgeprägt, wenn man die Bewerberstichprobe (Fleiß: $SD_{Bew}=0,36$; Ordnung: $SD_{Bew}=0,37$) mit der Kontrollgruppe (Fleiß: $SD_{KG}=0,41$; Ordnung: $SD_{KG}=0,47$) verglich. Auffällig war diesbezüglich, dass die Standardabweichungen vor allem bei den negativ gepolten Items (s. Ordnungs-Items O9, O15, O17 und O25) größer waren als bei den restlichen positiv gepolten Items. Dieser Effekt zeigte sich in beiden Stichproben, wobei er in der Kontrollgruppe, die ohnehin ein variableres Antwortverhalten zeigte, stärker ausgeprägt war. Dies ist auf die bestehende Range Restriction in der Bewerberstichprobe zurückzuführen (s. Abschnitt 6.3.5.1).

Bezüglich der niedrigsten und höchsten Antworten (s. *Min* und *Max*) zeigten sich vergleichbare Tendenzen in beiden Stichproben. Zumeist wurde die gesamte Bandbreite der Antwortskala genutzt, wobei eine leichte Tendenz hin zu positiveren Antworten zu beobachten war (s. hierzu auch 8.2.1.2 Itemschwierigkeiten). Bei den Items O7, F8, O11, F12, F18, O21, F22, F24, O27 und F30 wurde das Antwortformat nicht voll genutzt. Die Antwortkategorie „starke Ablehnung“ wurde in beiden Stichproben nicht gewählt.

Eine Überprüfung der Mittelwertsunterschiede mittels t -Tests ergab, dass sich die Mittelwerte der Auszubildenden und der Kontrollgruppe signifikant voneinander unterschieden. Sowohl für den Fleiß- ($t(165) = 7,109$, $p < .01$) und den Ordnungsaspekt ($t(165) = 5,348$, $p < .01$) als auch das Selbstwirksamkeitskonstrukt ($t(165) = 6,666$, $p < .01$) ergaben sich signifikante Unterschiede für die beiden untersuchten Stichproben.

Die Abbildungen 8-1 und 8-2 enthalten Verteilungen der Mittelwerte der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung differenziert nach Stichprobe (schwarz= Ausbildungsbewerber, $N_{Bew} = 740$; grau= Kontrollgruppe, $N_{KG} = 97$). Sowohl für den Fleiß- als auch für den Ordnungsaspekt sind die Verteilungskurven in beiden Stichproben rechtssteil, d.h. nach rechts verschoben.

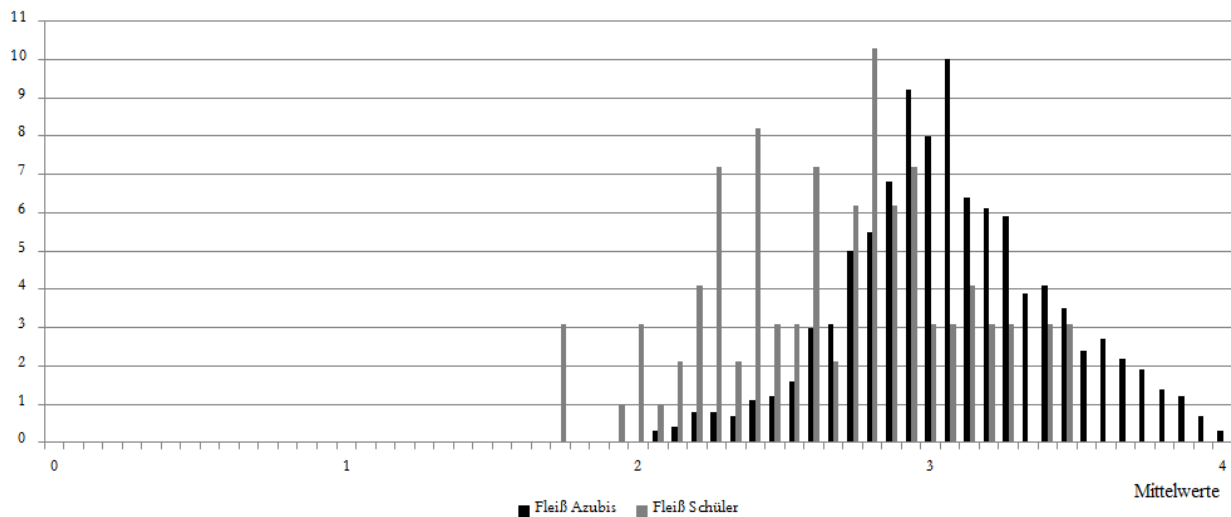


Abbildung 8-1: Häufigkeitsverteilung der Skalenmittelwerte der Fleißskala für die Bewerberstichprobe ($N_{Bew} = 740$) und die Kontrollgruppe ($N_{KG} = 97$)

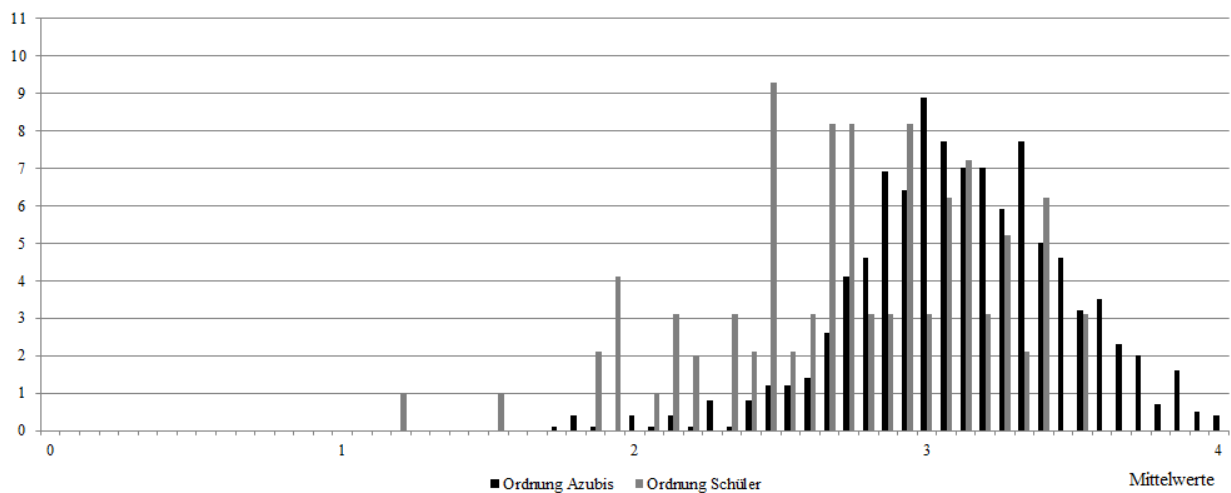


Abbildung 8-2: Häufigkeitsverteilung der Skalenmittelwerte der Ordnungsskala für die Bewerberstichprobe ($N_{Bew} = 740$) und die Kontrollgruppe ($N_{KG} = 97$)

Die Überprüfung der Normalverteilung der Testwerte ergab für die beiden Stichproben unterschiedliche Ergebnisse. Die Asymptotische Signifikanz betrug in der Kontrollgruppe ($N_{KG} = 97$) $p_{KG} = 0,51$ für die Fleißskala bzw. $p_{KG} = 0,48$ für die Ordnungsskala. Es kann daher von einer Normalverteilung der Daten in der Kontrollgruppe ausgegangen werden. Für die Stichprobe der Ausbildungsbewerber ($N_{Bew} = 740$) zeigte sich ein gegenteiliger Effekt. Die Asymptotische Signifikanz betrug $p_{Bew} = 0,00$ für die Fleißskala bzw. $p_{Bew} = 0,01$ für die Ordnungsskala und wurde somit signifikant. Die Daten der Bewerberstichprobe waren diesen Ergebnissen zufolge nicht normalverteilt. Allerdings ist das Ergebnis des Signifikanztests

stark von der Stichprobengröße abhängig. Bei großen Stichproben lehnt der Test auch bei geringen Abweichungen von der Normalverteilung die Normalverteilungshypothese ab (Bühner, 2011). Daher sollten zur Bestimmung der Normalverteilung neben der grafischen Überprüfung mittels Histogramm auch weitere Indikatoren herangezogen werden.

Bezüglich der Schiefe (v) ergibt sich für die Stichprobe der Ausbildungsbewerber ($N_{Bew} = 740$) für den Fleißaspekt mit $v_{Bew} = 0,07$ eine kaum merkliche Rechtsschiefe. Die Ordnungsskala weist mit $v_{Bew} = -0,39$ eine linksschiefe Verteilung auf. Gemäß Lienert und Raatz (1994) können bei dieser Stichprobengröße ($N_{Bew} > 400$) die gefundenen Abweichungen allerdings als praktisch insignifikant gelten. Die Rohwerteverteilung der Ausbildungsbewerber konnte daher dennoch als normalverteilt angesehen werden. Die Werteverteilungen der Kontrollgruppe waren mit $v_{KG} = -0,15$ für die Fleißskala und $v_{KG} = -0,71$ für die Ordnungsskala jeweils etwas bis mäßig linksschief.

Der Exzess betrug in der Kontrollgruppe ($N_{KG} = 97$) $E = 2,53$ für die Rohwerteverteilung der Fleißskala. Die zugehörige Ordnungsskala wies einen Exzess von $E = 3,54$ auf. Beide Werte liegen zwischen $+2,00$ und $+4,00$, was ein weiteres Indiz für die Normalverteilung der Messwerte in der Kontrollgruppe der Schüler ist. Für die Ausbildungsbewerber ($N_{Bew} = 740$) ergab sich für die Fleißskala ein Exzess von $E = 3,03$, was fast einer perfekten Normalverteilung entspricht. Für die Verteilung der Ordnungsskala zeigte sich ein Exzess von $E = 3,85$. Auch dieser Wert lag zwischen den kritischen Grenzen von $+2,00$ und $+4,00$. Die Befunde zu den Wölbungen der Verteilungskurven sprachen somit für eine Annahme der Normalverteilungsfunktion in den Daten sowohl in der Bewerberstichprobe als auch in der Kontrollgruppe.

Die Befunde der Verteilungsprüfungen deuten eher auf eine Normalverteilung der Daten hin. Methodisch könnte man nun mittels einer z-Transformation der Variablen eine gesicherte Normalverteilung herstellen. Zwar wäre damit eine Voraussetzung zur Berechnung vieler statistischer Verfahren gegeben, andererseits sind transformierte Daten vielfach schlechter interpretierbar als nicht transformierte (Tabachnik & Fidell, 2001). Auch mit Blick auf die ohnehin herrschende eingeschränkte Interpretierbarkeit aufgrund der Stichproben- und Untersuchungseigenschaften (s. 9.5 Limitationen) wird daher im Folgenden auf eine Transformation der Daten verzichtet und von der Normalverteilung der Daten ausgegangen.

8.2.1.2 Itemschwierigkeiten

Tabelle 8-5: Itemschwierigkeiten der AGS

Item	<i>P</i> Ausbildungs- bewerber	<i>P</i> Kontroll- gruppe	Item	<i>P</i> Ausbildungs- bewerber	<i>P</i> Kontroll- gruppe
Ordnung 1	.75	.67	Fleiß 2	.81	.70
Ordnung 3	.81	.70	Fleiß 4	.71	.59
Ordnung 5	.70	.61	Fleiß 6	.59	.57
Ordnung 7	.96	.83	Fleiß 8	.84	.72
Ordnung 9	.82	.64	Fleiß 10	.71	.58
Ordnung 11	.83	.71	Fleiß 12	.78	.70
Ordnung 13	.79	.68	Fleiß 14	.83	.77
Ordnung 15	.83	.70	Fleiß 16	.81	.68
Ordnung 17	.80	.70	Fleiß 18	.82	.77
Ordnung 19	.72	.65	Fleiß 20	.70	.61
Ordnung 21	.71	.70	Fleiß 22	.75	.63
Ordnung 23	.81	.75	Fleiß 24	.82	.70
Ordnung 25	.76	.67	Fleiß 26	.78	.69
Ordnung 27	.71	.63	Fleiß 28	.75	.66
Ordnung 29	.72	.75	Fleiß 30	.83	.67

Anmerkungen: P = Itemschwierigkeit; $N_{Bew} = 740$ Ausbildungsbewerber bzw. $N_{KG} = 97$ Kontrollgruppe; * umgepolte Items wurden vor der Analyse rekodiert.

Tabelle 8-5 gibt einen Überblick über die Itemschwierigkeiten der AGS. In der Bewerberstichprobe reichten die Itemschwierigkeiten von $P = .59$ bis $P = .96$. Eine besonders geringe Schwierigkeit ($P > .80$) wiesen in dieser Stichprobe die Items F2, O3, O7, F8, O9, O11, F14, O15, F16, O17, F18, O23, F24 und F30 auf. Die mittlere Itemschwierigkeit der Ordnungsskala betrug $P = .78$, die der Fleißskala $P = .77$. In der Kontrollgruppe gingen die Itemschwierigkeiten von $P = .57$ bis $P = .83$. Eine besonders niedrige Schwierigkeit wies in dieser Stichprobe das Item O7 auf. Die mittlere Itemschwierigkeit der Ordnungsskala betrug für die Kontrollgruppe $P = .69$, für die Fleißskala zeigt sich eine mittlere Itemschwierigkeit von $P = .67$. Insgesamt zeigte kein Item eine hohe Schwierigkeit ($P < .20$) auf. Tendenziell wiesen die Items eher eine niedrigere Schwierigkeit auf ($P > .50$). In der Kontrollgruppe fielen die Itemschwierigkeiten etwas niedriger aus als in der Bewerberstichprobe, d.h. den Aussagen wurde weniger stark zugestimmt.

Nach den Kriterien der Itemschwierigkeit bilden die beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung der AGS somit zwei Faktoren ab. Hypothese 1a kann somit als teilweise bestätigt angesehen werden.

8.2.1.3 Trennschärfen

Zur Ermittlung der Trennschärfekoeffizienten der einzelnen Items wurden die korrigierten Item-Skala-Korrelationen (r_{it}) berechnet, die in Tabelle 8-6 aufgelistet sind. Für die Bewerberstichprobe ($N_{Bew} = 740$) ergab sich ein mittlerer Trennschärfekoeffizient von $r_{it} = .39$ ($r_{it} = .17$ bis $.61$) für die Ordnungsskala. Als gering ($r_{it} < .30$) waren die Trennschärfen der Items O7 ($r_{it} < .23$), O19 ($r_{it} < .21$) und O29 ($r_{it} < .17$) zu bezeichnen. Die Items der Fleißskala wiesen einen höheren mittleren Trennschärfekoeffizienten von $r_{it} = .51$ ($r_{it} = .35$ bis $.63$) auf. Alle Items dieser Skala wiesen eine mittlere bis hohe Trennschärfe auf ($r_{it} > .30$).

In der Kontrollgruppe ($N_{KG} = 97$) konnte für den Ordnungsaspekt ein mittlerer Trennschärfekoeffizient von $r_{it} = .36$ ($r_{it} = .08$ bis $r_{it} = .65$) ermittelt werden. Die Korrelationskoeffizienten der Items O7 ($r_{it} = .08$), O9 ($r_{it} = .17$), O17 ($r_{it} = .26$), O19 ($r_{it} = .23$), O21 ($r_{it} = .27$) und O29 ($r_{it} = .19$) wiesen auf eine geringe Trennschärfe ($r_{it} < .30$) der Items in dieser Stichprobe hin. Bezüglich der Fleißskala wurde eine etwas höhere mittlere Trennschärfe von $r_{it} = .42$ ($r_{it} = .08$ bis $r_{it} = .57$) berechnet. Auf dieser Skala wies lediglich das Item F4 ($r_{it} = .08$) eine als gering zu klassifizierende Trennschärfe auf.

Tabelle 8-6: Korrigierte Trennschärfekoeffizienten der Items für die beiden AGS-Skalen Ordnung und Fleiß

Item	r_{it}	Item	r_{it}
Ordnung 1	.34 (.32)	Fleiß 2	.49 (.47)
Ordnung 3	.58 (.65)	Fleiß 4	.46 (.08)
Ordnung 5	.35 (.31)	Fleiß 6	.38 (.34)
Ordnung 7	.23 (.08)	Fleiß 8	.54 (.39)
Ordnung 9*	.30 (.17)	Fleiß 10	.45 (.49)
Ordnung 11	.40 (.33)	Fleiß 12	.59 (.54)
Ordnung 13	.57 (.53)	Fleiß 14	.50 (.35)
Ordnung 15*	.61 (.55)	Fleiß 16	.49 (.43)
Ordnung 17*	.40 (.26)	Fleiß 18	.56 (.51)
Ordnung 19	.21 (.23)	Fleiß 20	.48 (.42)
Ordnung 21	.38 (.27)	Fleiß 22	.54 (.52)
Ordnung 23	.46 (.54)	Fleiß 24	.63 (.45)
Ordnung 25*	.47 (.45)	Fleiß 26	.35 (.34)
Ordnung 27	.41 (.46)	Fleiß 28	.59 (.32)
Ordnung 29	.17 (.19)	Fleiß 30	.54 (.57)

Anmerkungen: r_{it} = korrigierte Trennschärfe; $N_{Bew} = 740$ Ausbildungsbewerber bzw. $N_{KG} = 97$ Kontrollgruppe; Werte der Kontrollgruppe stehen in Klammern; * umgepolte Items wurden vor der Analyse rekodiert.

Tendenziell wies die Kontrollgruppe ($N_{KG}=97$) etwas geringere Trennschärfekoeffizienten als die Bewerberstichprobe ($N_{Bew}=740$) auf. Kritische Korrelationskoeffizienten in beiden Stichproben wiesen nur die Items O7 ($r_{it}=.23$ bzw. $.08$) sowie O29 ($r_{it}=.17$ bzw. $.19$) auf. Insgesamt kann man von einer mittleren bis guten Trennschärfe der Items in den Gewissenhaftigkeitsskalen Fleiß und Ordnung ausgehen. Demnach kann Hypothese 1a nun vollständig angenommen werden.

8.2.2 Messgüte der Subskalen Fleiß und Ordnung

Der zweite Teil der Itemanalysen zielte auf die Überprüfung der Homogenität der Skalen ab. Die beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung sollten sich als reliabel im Sinne einer hohen internen Konsistenz erweisen (vgl. Hypothese 1b). Die beiden Gewissenhaftigkeitsaspekte der AGS sollten auch zeitlich stabil im Sinne einer hohen Retest-Reliabilität sein (vgl. Hypothese 2). Darüber hinaus sind die Ergebnisse der Überprüfung des theoretischen Modells der AGS dargestellt, was anhand der Überprüfung der Eindimensionalität der Skala sowie der zugehörigen Itemladungen erfolgt (vgl. Hypothesen 3a und 3b). Im Folgenden werden die Ergebnisse der Überprüfung der Messgüte der AGS bezüglich ihrer Reliabilität (8.2.2.1), Retest-Reliabilität (8.2.2.2) sowie der Homogenität (8.2.2.3) dargestellt.

8.2.2.1 Reliabilitätsschätzungen der Subskalen

Sowohl für die Bewerberstichprobe ($N_{Bew}=740$) als auch für die Kontrollgruppe ($N_{KG}=97$) wurde für beide Gewissenhaftigkeitsaspekte Cronbachs α mittels SPSS berechnet. Die Reliabilität des Fleißaspekts war in der Bewerberstichprobe gut ($\alpha_{Bew}=.86$), die Kontrollgruppe wies ebenfalls eine gute Reliabilität der Skala auf ($\alpha_{KG}=.80$). Die zugehörigen standardisierten Alpha-Koeffizienten waren mit $\alpha_{BewS}=.87$ bzw. $\alpha_{KGS}=.81$ nur geringfügig höher, weshalb in beiden Stichproben von identischen Itemvarianzen ausgegangen werden kann.

Für den Ordnungsaspekt ergaben sich etwas geringere Konsistenzen. Sowohl in der Bewerberstichprobe ($\alpha_{Bew}=.78$), als auch in der Kontrollgruppe ($\alpha_{KG}=.75$) wiesen die Items eine befriedigende Konsistenz auf. Auch auf dieser Skala kann von Varianzgleichheit ausgegangen werden, da die standardisierten Alpha-Koeffizienten nur geringfügig höher ($\alpha_{BewS}=.79$) bzw. gleich hoch ($\alpha_{KGS}=.75$) ausfielen als die unstandardisierten Werte. Insgesamt kann man also von einer konsistenten Skala ausgehen. Hypothese 1b wurde demnach bestätigt.

Dem Anhang (s. A2, Tabelle A-10) können die Konsistenzkoeffizienten entnommen werden, die resultieren würden, wenn eines der Items gelöscht worden wäre.

8.2.2.2 Retest-Reliabilität

Zur Bestimmung der Retest-Reliabilität wurden die Merkmalsausprägungen der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie der Selbstwirksamkeitsskala aus den Daten der Bewerberstichprobe (T1) jeweils mit den entsprechenden Ausprägungen der Wiederholungsmessung (T2) korreliert. Die Ergebnisse können Tabelle 8-7 entnommen werden.

Tabelle 8-7: Retestreliabilität AGS und Selbstwirksamkeit

	<i>M</i>	<i>SD</i>	1	2	3	4	5	6
T1 Fleiß	3,09	.35	(.85)					
T1 Ordnung	3,14	.40	.46**	(.82)				
T1 SE	3,05	.36	.79**	.44**	(.84)			
T2 Fleiß	3,07	.32	.51**	.19	.42**	(.81)		
T2 Ordnung	3,08	.35	.17	.46**	.24	.30*	(.72)	
T2 SE	2,86	.34	.33**	.02	.50**	.65**	.22	(.84)

Anmerkungen: $N_R=63$; M = Mittelwert; SD = Standardabweichung; SE = Selbstwirksamkeit; Reliabilitäten (Cronbachs α) stehen in Klammern; ** $p < 0,01$ (zweiseitig); * $p < 0,05$ (zweiseitig).

In der Reteststichprobe ($N_R = 63$) waren die internen Konsistenzen der Gewissenhaftigkeitsaspekte Fleiß ($\alpha_{F,T1} = .85$) und Ordnung ($\alpha_{O,T1} = .82$) sowie der Selbstwirksamkeitsskala ($\alpha_{SE,T1} = .84$) als hoch zu bezeichnen. Auch die internen Konsistenzen der erneuten Befragung (T2) waren zufriedenstellend ($\alpha_{O,T2} = .72$) bis hoch ($\alpha_{F,T2} = .81$ und $\alpha_{SE,T2} = .84$). Die Retest-Reliabilitäten fielen erwartungsgemäß niedriger aus. So waren die Korrelationen sowohl für die beiden Gewissenhaftigkeitsaspekte Fleiß ($r = .51, p < 0,01$) und Ordnung ($r = .46, p < 0,01$) als auch für die Selbstwirksamkeitsskala ($r = .50, p < 0,01$) nur in mittlerem Maße ausgeprägt.

Nachfolgend wurde mittels t -Test für gepaarte Stichproben überprüft, ob die Mittelwerte der Skalen zwischen T1 und T2 signifikant unterschiedlich waren. Sowohl für die Fleißskala ($T = 0,578; p = 0,57$) als auch für den Ordnungsaspekt ($T = 1,220; p = 0,23$) ergab sich kein signifikanter Unterschied der Mittelwerte zwischen T1 und T2. Lediglich für die Selbstwirksamkeitsskala wurde der Vergleich der Mittelwerte signifikant ($T = 4,315, p < .01$).

Da Veränderungen des wahren Werts, die für alle Personen gleich ausfallen, die Retestkorrelation als Schätzung für die Retestreliabilität nicht senken, musste davon ausgegangen werden, dass differenzielle Effekte, d.h. Veränderungen, die für manche Testpersonen den Wert zu T2 erhöhen oder verringern, die geringe Retestkorrelation erklärten. Dies ist in den Streudiagrammen (s. Abbildungen 8-3 bis 8-5) grafisch veranschaulicht.

Die Abbildungen 8-3 bis 8-5 enthalten jeweils die Streudiagramme der Testwerte der Fleiß-, Ordnungs- bzw. Selbstwirksamkeitsskala zu T1 und T2 zur Veranschaulichung der absoluten und relativen Stabilität der Messwerte. Die Winkelhalbierende ist durch die grau gestrichelte Gerade angedeutet. Für Werte, die auf dieser Geraden liegen, liegt absolute Stabilität vor, d.h. die Ausprägungen sind zu T1 und T2 jeweils gleich.

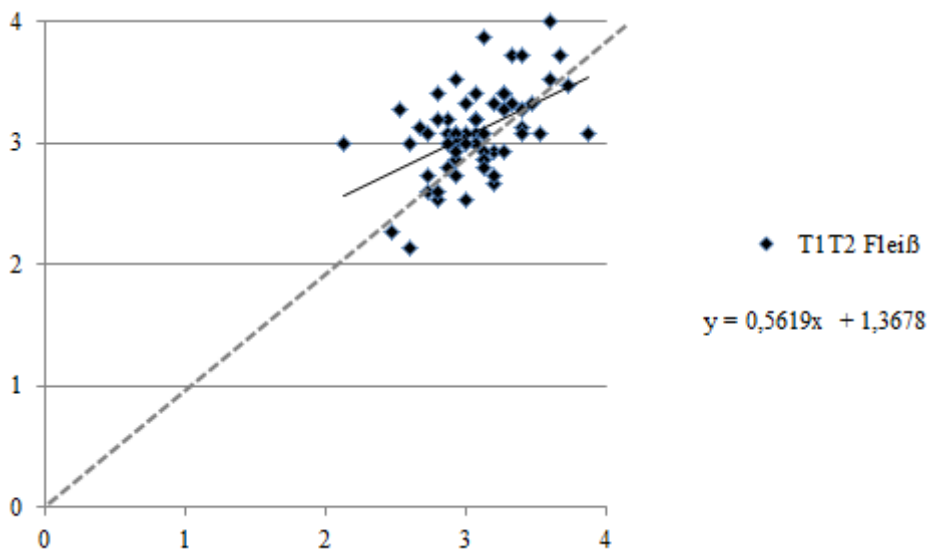


Abbildung 8-3: Streudiagramm der beiden Fleiß-Messungen (X-Achse: T1; Y-Achse: T2).

In allen drei Abbildungen sind jeweils auch die Regressionsgeraden (s. schwarze Trendlinien) eingezeichnet, die den Zusammenhang der Wert von T1 (X-Achse) und T2 (Y-Achse) veranschaulichen. Diese schneiden jeweils die zugehörige Winkelhalbierende.

Wie den Abbildungen 8-3 bis 8-5 zu entnehmen ist, wichen die Probanden sowohl in den Gewissenhaftigkeitsaspekten Fleiß und Ordnung als auch auf der Selbstwirksamkeitsskala sowohl nach unten als auch nach oben ab. Das heißt, es gab Auszubildende, bei denen sich der Messwert zum zweiten Testzeitpunkt (T2) verringert hatte (= Werte unterhalb der Winkelhalbierenden), aber auch welche, bei denen der Testwert angestiegen ist (= Werte oberhalb der Winkelhalbierenden).

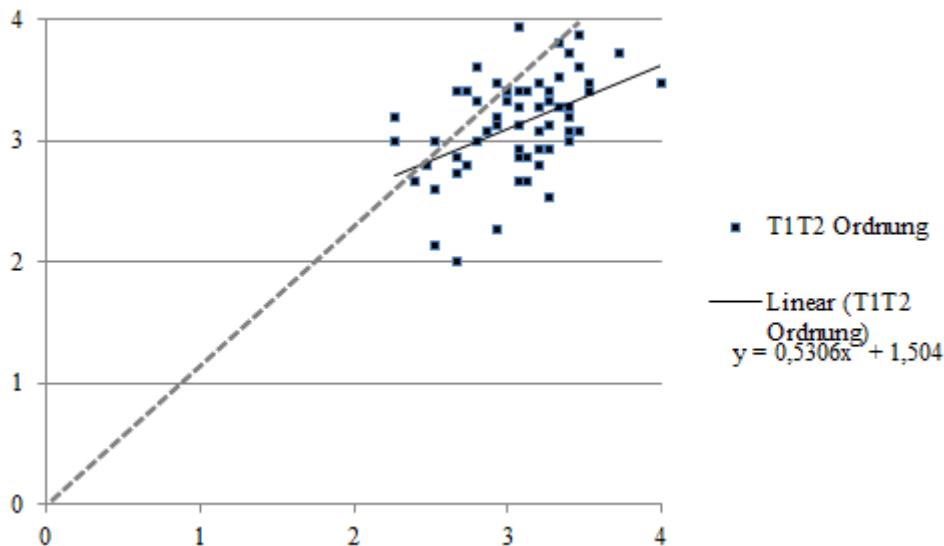


Abbildung 8-4: Streudiagramm der beiden Ordnungs-Messungen (X-Achse: T1; Y-Achse: T2).

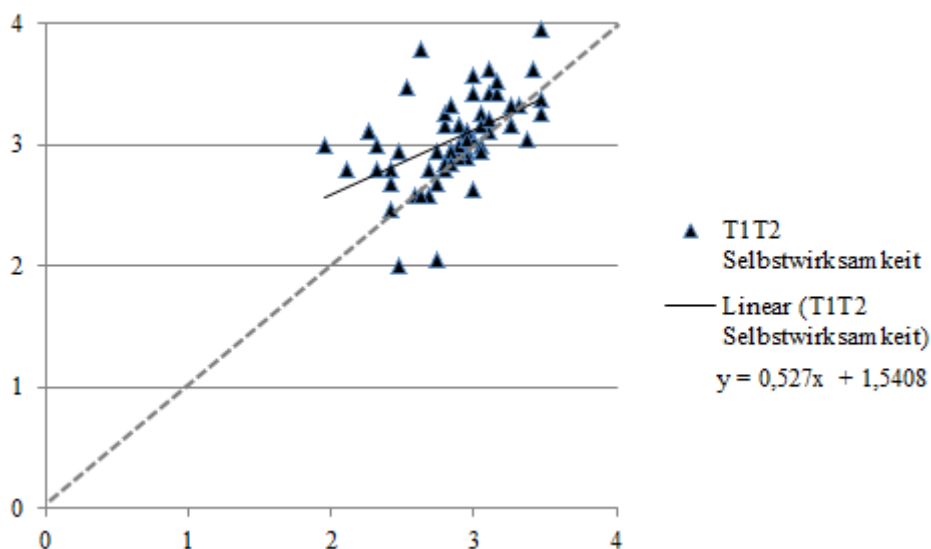


Abbildung 8-5: Streudiagramm der beiden Selbstwirksamkeits-Messungen (X-Achse: T1; Y-Achse: T2).

Aufgrund der Abweichungen der Daten von T1 zu T2 konnte man weder von einer absoluten noch einer relativen Stabilität der Messwerte ausgehen. Es ist zu vermuten, dass die abgebildeten Veränderungen zumindest teilweise auf einen gewissen Anteil unsystematischer Messfehler zurückzuführen sind. Es wird weiter angenommen, dass diese unsystematischen Einflüsse zu beiden Testzeitpunkten (T1 und T2) gleichermaßen auftraten und somit für die verringerten Korrelationen verantwortlich sind.

Um zu überprüfen, ob Messfehler für die geringe Retest-Reliabilität der AGS und Selbstwirksamkeitsskala verantwortlich waren, wurde der *Reliable Change Index* (RCI) für die beiden Gewissenhaftigkeitsskalen Fleiß und Ordnung sowie für die Selbstwirksamkeitsskala wie in 6.2.2.2 beschrieben berechnet. Für die Fleißskala ergab sich $RCI = -0,07$, für die Ordnungsskala $RCI = -0,17$ sowie die Selbstwirksamkeitsskala $RCI = -0,531$. Als Vergleichswert wurde $z = 1,65$ bei einseitiger Testung herangezogen, da aufgrund des hohen Anreizes der Testsituation zu T1 (Auswahltag der Azubis) davon ausgegangen wurde, dass sich die Auszubildenden in dieser Situation eher positiv, d.h. mit höheren Werten eingeschätzt haben. Jeder der drei berechneten RCI-Werte lag somit deutlich unter dem kritischen Wert. Daher bestätigten diese Ergebnisse die Annahmen, dass die geringe Retest-Reliabilität aufgrund mangelnder Mess-

genauigkeit zustande gekommen sein kann, die auch auf unterschiedliche Tendenzen zu T1, Werte zu verfälschen, zurückgeführt werden kann.

Ein Faktor, der die Retest-Reliabilität eines Tests verringert, ist die Unreliabilität des verwendeten Testverfahrens, weshalb die ermittelten empirischen Retest-Korrelationen mit Hilfe der Minderungskorrektur (s. 6.2.2.2) korrigiert wurden. Hiernach ergab sich für den Fleißaspekt eine korrigierte Retest-Reliabilität von $\alpha_{F,korr} = .61$. Für den Ordnungsaspekt konnte eine korrigierte Retest-Reliabilität von $\alpha_{O,korr} = .60$ errechnet werden. Die Retest-Reliabilität der Selbstwirksamkeitsskala stieg nach der Korrektur ebenfalls auf $\alpha_{SE,korr} = .60$ an. Unter Berücksichtigung der Unreliabilität des verwendeten Maßes, konnte somit von einer knapp ausreichenden Retest-Reliabilität für die AGS sowie für die Selbstwirksamkeitsskala ausgegangen werden. Hypothese 2 muss diesen Ergebnissen zufolge nicht verworfen werden.

8.2.2.3 Skalenhomogenität der Subskalen Fleiß und Ordnung

Vor der Durchführung der konfirmatorischen Faktorenanalysen (CFA) wurden einige noch ausstehende Voraussetzungsüberprüfungen vorgenommen. Hierzu zählten die Überprüfung der multivariaten Normalverteilung der Items und eine Überprüfung der Kollinearität der Items der AGS.

Der Test auf multivariate Normalverteilung wurde mit der o.g. AMOS-Software durchgeführt. Im Anhang (s. A-3, Tabelle A-11) sind die Ergebnisse des hierzu durchgeführten Mardia-Tests für die Einzelitems der AGS bzw. deren Subskalen Ordnung und Fleiß aufgeführt. Auf Einzelitemebene ist von einer multivariaten Normalverteilung der Daten in beiden Stichproben auszugehen.

Die multivariate Kurtosis fiel für die Bewerberstichprobe ($N_{Bew} = 740$) mit 50,49 für die Subskala Ordnung und mit 42,94 für Fleiß allerdings sehr hoch aus. Als Indikator für die multivariate Normalverteilung wurde in diesem auch der critical ratio (c.r.) als Prüfgröße herangezogen, der als z-Wert interpretiert werden kann. Dieser war in dieser Stichprobe sowohl für den Fleiß- (25,86) als auch für den Ordnungsaspekt (30,41) deutlich größer als der kritische z-Wert ($\pm 1,96$) bei einer zweiseitigen Prüfung mit $\alpha = 0.05$. Aus diesen Gründen konnte in der Bewerberstichprobe davon ausgegangen werden, dass keine multivariate Normalverteilung der Daten der AGS vorlag. Da für diese Stichprobe allerdings das ADF-Verfahren zur Überprüfung der Skalenhomogenität durchgeführt werden sollte, welches keine Verteilungsannahmen benötigt, wurden diese Ergebnisse für diesen Teil der Analysen nicht weiter berücksichtigt.

Für die Kontrollgruppe der Schüler ($N_{KG} = 97$) fiel die multivariate Kurtosis sowohl für den Gewissenhaftigkeitsaspekt Fleiß (17,83) als auch für den Ordnungsaspekt (17,01) hoch aus, wenn auch deutlich niedriger als in der Bewerberstichprobe. Der critical ratio (c.r.) war für beide Gewissenhaftigkeitsaspekte (Fleiß: 3,71; Ordnung: 3,89) höher als der kritische z-Wert ($\pm 1,96$) bei einer zweiseitigen Prüfung mit $\alpha = 0.05$. Auch in der Kontrollgruppe konnte somit keine multivariate Normalverteilung in den Daten nachgewiesen werden. In diesem Falle soll bei der ML-Methode das Bollen-Stine-Bootstrap-Verfahren angewendet werden (Bühner, 2011, S. 409), welches die Überschreitungswahrscheinlichkeit (p) korrigiert, die anhand der Freiheitsgrade (df) und der χ^2 -Statistik bestimmt wird. Da in den Daten vorab die fehlenden Werte ersetzt wurden (s. 8.1.1), kann das Verfahren durchgeführt werden.

Zur Überprüfung der Kollinearität der AGS wurden die Korrelationsmatrizen der Items für jede der beiden Subskalen herangezogen. Als Richtwert galt, dass die Korrelationen zwischen den Items nicht größer als 0.85 sein sollen (Bühner, 2011, S. 432). Sowohl in der Bewerberstichprobe ($N_{Bew} = 740$) als auch in der Kontrollgruppe ($N_{KG} = 97$) befanden sich alle Interitemkorrelationen der Subskalen Fleiß und Ordnung unterhalb dieser kritischen Grenze

(s. Anhang A-3, Tabellen A-12 und A-13), weshalb die Daten auch in diesem Punkt für die Durchführung einer CFA als geeignet beurteilt werden konnten.

Zur Überprüfung der Skalenhomogenität der beiden Gewissenhaftigkeitsaspekte der AGS wurde für jeden Aspekt ein einfaktorielles Modell mit einer latenten und 15 beobachteten Variablen spezifiziert (s. Abb. 6-1). Die Analysen wurde sowohl mit den Daten der Bewerberstichprobe ($N_{Bew}=740$) als auch mit den Daten der Kontrollgruppe ($N_{KG}=97$) durchgeführt. Die Nullhypothese lautete hierzu, dass das Modell zur Datenstruktur passt. Die Alternativhypothese war dementsprechend, dass das Modell nicht der Struktur in den Daten entspricht. Da in diesem Falle die Geltung der Nullhypothese getestet wurde, galt es, den Beta-Fehler durch eine Erhöhung des Alpha-Niveaus auf $\alpha = 0.10$ zu minimieren.

Die AMOS-Ausgaben der einfaktoriellen CFA-Modelle sind im Anhang (s. A4, Abbildungen A-1 bis A-4) enthalten.

Für die Bewerberstichprobe ($N_{Bew}=740$) betrug der χ^2 -Wert für die Subskala Fleiß mit 90 Freiheitsgraden 232,527. Dieser wurde bei dieser Anzahl von Freiheitsgraden signifikant ($p=0.000$). Somit wurde das Modell durch den Modelltest abgelehnt. Von 2000 Bootstrap-Stichproben passte das Modell in 1999 Fällen besser als in der Bewerberstichprobe. Lediglich in einer Bootstrap-Stichprobe war das Modell schlechter als in der Bewerberstichprobe. Der korrigierte p -Wert wurde mit $p=0.001$ angegeben. Das heißt, die Verletzung der multivariaten Normalverteilung, welche durch den Mardia-Test angezeigt wurde, wirkte sich nur marginal auf den p -Wert aus. Das Modell würde somit auch mit Verteilungskorrektur verworfen werden.

Für die Überprüfung der Modellgüte des Ordnungsaspekts ergaben sich in der Bewerberstichprobe ($N_{Bew}=740$) vergleichbare Ergebnisse. Der χ^2 -Wert mit 90 Freiheitsgraden von 382,252 wurde ebenfalls signifikant ($p=0.000$), was auch für die Subskala Ordnung zu einer Ablehnung des Modells durch den Modelltest führte. Das Modell passte in 2000 von 2000 Bootstrap-Stichproben besser als in der Bewerberstichprobe. Der p -Wert veränderte sich nicht ($p=0.000$). Eine Verteilungskorrektur hatte somit keinen Einfluss auf die Modellgüte.

Für die Kontrollgruppe der Schüler ($N_{KG}=97$) betrug der χ^2 -Wert für die Subskala Fleiß mit 90 Freiheitsgraden 135,436. Dieser wurde bei dieser Anzahl von Freiheitsgraden signifikant ($p=0.001$). Somit wurde das Modell durch den Modelltest abgelehnt. Von 2000 Bootstrap-Stichproben passte das Modell in 1775 Fällen besser als in der Kontrollgruppe. In 225 Bootstrap-Stichproben war das Modell schlechter als in der Kontrollgruppe. Der korrigierte p -Wert wurde mit $p=0.113$ angegeben. Das heißt, die Verletzung der multivariaten Normalverteilung, welche durch den Mardia-Test angezeigt wurde, wirkte sich auf den p -Wert aus. Nach der Korrektur führte dies dazu, dass das Modell als gültig angesehen werden konnte.

Für die Überprüfung der Modellgüte des Ordnungsaspekts ergaben sich in der Kontrollgruppe ($N_{KG}=97$) ähnliche Ergebnisse wie zuvor. Der χ^2 -Wert mit 90 Freiheitsgraden von 127,904 wurde ebenfalls signifikant ($p=0.005$), was auch für die Subskala Ordnung zu einer Ablehnung des Modells durch den Modelltest führte. Das Modell passte in 1687 von 2000 Bootstrap-Stichproben besser als in der Kontrollgruppe. In 131 Bootstrap-Stichproben war das Modell schlechter als in der Kontrollgruppe. Der korrigierte p -Wert wurde mit $p=0.157$ angegeben. Das heißt, die Verletzung der multivariaten Normalverteilung wirkte sich auf den p -Wert aus. Nach der Korrektur konnte das Modell daher als gültig angesehen werden.

Für die Kontrollgruppe ($N_{KG}=97$) lag somit nach der Bollen-Stine-Bootstrap-Korrektur der angenommene Modell-Fit vor. In der Bewerberstichprobe ($N_{Bew}=740$) musste demgegenüber auch nach der Korrektur sowohl für den Fleiß- als auch für den Ordnungsaspekt der angenommene Modell-Fit verworfen werden. Problematisch bei χ^2 -Modelltestungen ist allerdings, dass bei großen Stichproben bereits geringe Abweichungen zwischen der vom Modell impli-

zierten und der empirisch beobachteten Kovarianzmatrix zu einem signifikanten Testergebnis und damit zu einer Ablehnung des Modells führen, obwohl große Stichproben sonst zu einer genaueren Schätzung der Parameter in dem Modell führen. Auch wird der χ^2 -Wert mit zunehmender Anzahl an Items pro Faktor überschätzt (vgl. Bühner, 2011), was wiederum zu einer verstärkten Ablehnung des Modells führt. Daher ist anzunehmen, dass die gefundenen Effekte in der Bewerberstichprobe v.a. auf die Größe der Stichprobe zurückzuführen waren, weshalb im Folgenden zusätzlich einige Fit-Indizes zur Bewertung der Modellgüte herangezogen werden.

In der Bewerberstichprobe ($N_{Bew} = 740$) lag der *RMSEA* des Fleißaspektes bei *RMSEA* = 0,046. Der Cut-off-Wert für große Stichproben ($N > 250$) liegt Bühner (2011) zufolge bei *RMSEA* $\leq$ 0,06, sodass man in diesem Fall davon ausgehen konnte, dass das Modell passt. Der *SRMR* für den Fleißaspekt lag mit *SRMR* = 0,0641 ebenfalls unter der in Bühner (2011) definierten Grenze von *SRMR* $\leq$ 0,11. Insgesamt konnte man für den Fleißaspekt in der Bewerberstichprobe von einem exakten Modell-Fit ausgehen.

Für den Ordnungsaspekt ergab sich in der Bewerberstichprobe ein *RMSEA* von 0,066, womit sich dieser knapp über dem Cut-off-Wert von *RMSEA* $\leq$ 0,06 befand und gegen einen Modell-Fit sprach. Der *SRMR* für die Subskala Ordnung betrug *SRMR* = 0,0930 und lag somit unter der kritischen Grenze von *SRMR* $\leq$ 0,11. Die Befundlage zum Modell-Fit der Ordnungsskala in der Bewerberstichprobe ist somit uneindeutig.

In der Kontrollgruppe ($N_{KG} = 97$) lag der *RMSEA* der Subskala Fleiß bei *RMSEA* = 0,073. Der Cut-off-Wert für kleine Stichproben ($N < 250$) liegt Bühner (2011) zufolge bei *RMSEA* $\leq$ 0,08, sodass man davon ausgehen konnte, dass das Modell passt. Der *SRMR* für den Fleißaspekt lag mit *SRMR* = 0,0814 ebenfalls unter der in Bühner (2011) genannten Grenze von *SRMR* $\leq$ 0,11. Für den Fleißaspekt in der Kontrollgruppe sprachen daher auch die Fit-Indizes für das Vorliegen des angenommenen Modell-Fits.

Für den Ordnungsaspekt ergab sich in der Kontrollgruppe ($N_{KG} = 97$) ein *RMSEA* = 0,066. Dieser Wert befand sich somit unter dem Cut-off-Wert von *RMSEA* $\leq$ 0,08 (bei kleinen Stichproben) und sprach für einen Modell-Fit. Der *SRMR* für den Ordnungsaspekt in dieser Stichprobe lag mit *SRMR* = 0,0857 unter der Grenze von *SRMR* $\leq$ 0,11. Von einem exakten Fit konnte man daher auch für den Ordnungsaspekt in der Kontrollgruppe ausgehen.

Aufgrund der standardisierten Ladungen und Fehlervarianzen konnte man über die eben beschriebenen Ergebnisse hinaus für die beiden Gewissenhaftigkeitsaspekte in beiden Stichproben von lokalem Modell-Fit ausgehen (s. Anhang A4). Auch die Kommunalitäten der Items deuteten bei den Gewissenhaftigkeitsaspekten Fleiß und Ordnung auf homogene Faktoren hin.

Der Modell-Fit der Subskala Fleiß konnte mit $\chi^2(90) = 232,527$ ($p = 0,000$), *RMSEA* = 0,046 (CI90: 0,039 – 0,054) und *SRMR* = 0,0641 in der Bewerberstichprobe ($N_{Bew} = 740$) sowie $\chi^2(90) = 135,436$ ($p = 0,001$), *RMSEA* = 0,073 (CI90: 0,046 – 0,097) und *SRMR* = 0,0814 in der Kontrollgruppe ($N_{KG} = 97$) als bedingt zutreffend angenommen werden. In der Subskala Ordnung konnte von einem Modell-Fit mit $\chi^2(90) = 382,252$ ($p = 0,000$), *RMSEA* = 0,066 (CI90: 0,060 – 0,073) und *SRMR* = 0,0930 in der Bewerberstichprobe ($N_{Bew} = 740$) sowie $\chi^2(90) = 127,904$ ($p = 0,005$), *RMSEA* = 0,066 (CI90: 0,037 – 0,091) und *SRMR* = 0,0857 in der Kontrollgruppe ($N_{KG} = 97$) ebenfalls von einer bedingten Passung des angenommenen Modells ausgegangen werden. Die Daten deuteten an, dass es sich bei den Gewissenhaftigkeitsaspekten Fleiß und Ordnung um homogene Faktoren handelte. Hypothese 3a konnte somit angenommen werden.

Die Überprüfung des zweifaktoriellen Modells (s. Abb. 6-2) wurde ebenfalls mittels konfirmatorischer Faktorenanalysen überprüft.

Alle Items der Subskala Ordnung im zweifaktoriellen Modell der Bewerberstichprobe ($N_{Bew}=740$) luden deutlich auf dem eigenen Faktor ($\lambda_O=0,21-0,69$). Das Item O29 wies eine negative Ladung auf ($\lambda_{O29}=-0,12$). Die Items der Subskala Fleiß luden ebenfalls alle auf den zugehörigen Faktor ($\lambda_F=0,49-0,82$). Die Ladungen der Fleiß-Items fielen dabei deutlich höher aus als die Ladungen der Ordnungs-Items. Allerdings musste das Modell für diese Stichprobe durch den Modelltest abgelehnt werden, da $\chi^2(404)=1966,930$ ($p=0,000$). Die multivariate Kurtosis fiel mit 123,417 sehr hoch aus und deutete gemeinsam mit dem *c.r.*-Wert ($z=38,310$, $p<0,001$) an, dass für diese Analyse keine multivariate Normalverteilung der Daten vorlag. Die Durchführung der Bollen-Stine-Bootstrap-Korrektur führte zu einem korrigierten $p=0,955$. Das heißt, die Verletzung der multivariaten Normalverteilung wirkte sich in dieser Stichprobe stark auf die Modellpassung bzw. die Gültigkeit der Indikatoren aus. Der *RMSEA* war mit $RMSEA=0,072$ für das zweifaktorielle Modell etwas größer als der Grenzwert für große Stichproben ($RMSEA\leq 0,06$). Von einem Modell-Fit konnte somit nach der Korrektur und aufgrund der eindeutigen Faktorladungen ausgegangen werden, auch wenn der *RMSEA* knapp über der von Bühner (2011) genannten Grenze lag.

Für die Kontrollgruppe der Schüler ($N_{KG}=97$) ergaben sich ähnliche Ergebnisse. Sämtliche Items der Subskala Ordnung luden deutlich auf dem eigenen Faktor ($\lambda_O=0,06-0,85$). Die Items der Subskala Fleiß luden ebenfalls alle auf den zugehörigen Faktor ($\lambda_F=0,08-0,66$). Die Ladungen der beiden Faktoren waren dabei vergleichbar. Allerdings musste auch für die Kontrollgruppe der Schüler ($N_{KG}=97$) das Modell durch den Modelltest abgelehnt werden, da der χ^2 -Wert signifikant wurde ($\chi^2(404)=613,053$, $p=0,000$). Die multivariate Kurtosis fiel mit 51,493 wiederum hoch aus und deutete gemeinsam mit dem *c.r.*-Wert ($z=5,787$, $p<0,001$) an, dass auch für die Kontrollgruppe keine multivariate Normalverteilung der Daten vorlag. Die Durchführung der Bollen-Stine-Bootstrap-Korrektur führte zu einem korrigierten $p=0,182$. Die Steigerung des p -Werts deutete somit an, dass die Verletzung der multivariaten Normalverteilung sich in dieser Stichprobe stark auf die Indikatoren der Modellpassung auswirkte. Nach der Korrektur war von einem Modell-Fit auszugehen. Mit 0,073 lag der *RMSEA* für das zweifaktorielle Modell unter dem in Bühner (2011) definierten Grenzwert für kleine Stichproben ($RMSEA\leq 0,08$), was wiederum für das angenommene Modell sprach. Der *SRMR* lag mit $SRMR=0,0989$ unter der in Bühner (2011) genannten Grenze von $SRMR\leq 0,11$. Dies sprach wiederum für die Gültigkeit des angenommenen Modells.

Weitere Kennwerte zur Überprüfung des Modell-Fit können dem Anhang (s. A4, Tabelle A-14 und A-15) entnommen werden.

Insgesamt konnte somit unter Beachtung der Einschränkungen der Testvoraussetzungen und damit einhergehenden bedingten Interpretierbarkeit der Ergebnisse davon ausgegangen werden, dass den AGS ein zweifaktorielles Modell zugrunde liegt. Das Modell musste nicht verworfen werden. Die Fit-Indizes sowie Faktorladungen zeigten an, dass die Abweichungen von passenden Modellen noch innerhalb eines akzeptierten Rahmens lagen. Hypothese 3b kann somit als eingeschränkt bestätigt angesehen werden.

8.2.3 Konstruktvalidierung der AGS

Zur Überprüfung der Konstruktvalidierung der AGS wurde eine Multi-Trait-Multi-Method-Analyse (MTMM) mit den Datensätzen der Bewerberstichprobe ($N_{Bew}=740$) nach Campbell und Fiske (1959) berechnet. Es wurden hohe Zusammenhänge zwischen ähnlichen Konstrukten (Hypothesen 4a und 4b) sowie geringere Zusammenhänge zwischen divergenten Konstrukten (Hypothesen 4c bis 4e) angenommen. Die zugehörigen bivariaten Korrelationen der konvergenten und diskriminanten Maßen sind in der Tabelle 8-8 aufgeführt. Hierzu ist einschränkend anzumerken, dass die dargestellten Zusammenhänge nicht sauber zwischen

Methoden- und inhaltlichen Varianzen trennen. Im strengen Sinne erfüllt diese Analyse somit nicht die Anforderungen des MTMM-Modells von Campbell und Fiske (1959).

Tabelle 8-8: Korrelationen der MTMM-Analyse

Skala	Fleiß	Ordnung	GEW	NEU	OFF
Fleiß	(.86)				
Ordnung	.42**	(.78)			
GEW	.65**	.62**	(.80)		
NEU	-.47**	-.31**	-.48**	(.74)	
OFF	.26**	.36**	.35**	-.25**	(.65)
LPS 3+4	.06	-.06	.05	-.04	.01

Anmerkungen: $N_{Bew} = 740$; GEW= Gewissenhaftigkeit; NEU= Neurotizismus; OFF= Offenheit; ** $p < 0,01$ (zweiseitig); Werte in Klammern entsprechen Cronbachs α .

Die Interkorrelation der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung lagen im mittleren Bereich ($r_{F,O} = .42$, $p < 0,01$). Die vergleichsweise niedrige Korrelation deutete an, dass die beiden Subskalen inhaltlich nicht vollständig identisch sind, was dem Konstruktionsprinzip der AGS entspricht. Auffällig hoch waren demgegenüber die Korrelationen der beiden Gewissenhaftigkeitsaspekte mit der Gewissenhaftigkeitsskala aus dem NEO-FFI ($r_{F,G} = .65$, $p < 0,01$ bzw. $r_{O,G} = .62$, $p < 0,01$). Die hohen Zusammenhänge deuteten darauf hin, dass die inhaltliche Nähe im Sinne der konvergenten Validität der beiden Subskalen Fleiß und Ordnung zu einem etablierten Gewissenhaftigkeitsmaß gegeben war.

Bezüglich der diskriminanten Validität zeigten die beiden Gewissenhaftigkeitsaspekte der AGS negative Zusammenhänge mit der Neurotizismus-Skala des NEO-FFI. Fleiß zeigte einen mittleren Zusammenhang mit Neurotizismus ($r_{F,N} = -.47$, $p < 0,01$), während Ordnung eine niedrige, negative Korrelation mit Neurotizismus aufwies ($r_{O,N} = -.31$, $p < 0,01$). Das negative Vorzeichen der beiden Korrelationen war im Einklang mit Korrelationsstudien zu den fünf Faktoren des NEO-FFI, in denen der Zusammenhang der Gewissenhaftigkeits- und Neurotizismus-Skala ebenfalls negativ war (vgl. Borkenau & Ostendorf, 2008, S. 19 ff). Mit der Offenheitsskala des NEO-FFI zeigte sowohl der Fleißaspekt ($r_{F,OFF} = .26$, $p < 0,01$) als auch der Ordnungsaspekt ($r_{O,OFF} = .36$, $p < 0,01$) einen nur geringen Zusammenhang auf, was für eine Unterschiedlichkeit der Konstrukte sprach. Darüber hinaus wurde der Zusammenhang der beiden Gewissenhaftigkeitsskalen mit dem Summenwert der LPS-Subskalen 3 und 4 überprüft. Hier ergaben sich sowohl für den Fleiß- ($r_{F,LPS} = .06$, n.s.) als auch für den Ordnungsaspekt ($r_{O,LPS} = -.06$, n. s.) keine signifikanten Zusammenhänge, die nahe null lagen. Die gefundenen Ergebnisse sprachen somit insgesamt dafür, dass die Gewissenhaftigkeitsaspekte Fleiß und Ordnung zu konstruktfernen Kriterien keine bzw. nur wenig Übereinstimmung zeigten. Hypothese 4 kann somit als bestätigt angesehen werden.

Da die in Tabelle 8-8 berichteten Korrelationen keine statistische Überprüfung des Modells abbilden konnten und Messfehler nicht berücksichtigt wurden, wurden die Abbildung 6-3 dargestellten Modellannahmen über die MTMM-Analyse hinaus mittels konfirmatorischer Faktorenanalyse überprüft. Die Annahmen hierzu waren, dass sich das Modell der AGS im Vergleich mit einem weiteren Gewissenhaftigkeitsmaß als faktoriell valide erweisen würde (Hypothese 5a). Auch gegenüber divergenten Kriterien sollten sich die AGS als faktoriell valide erweisen (Hypothesen 5b und 5c).

Als Schätzmethode wurde die Maximum-Likelihood-Methode gewählt. Die Stichprobengröße war mit $N_{Bew} = 740$ Probanden für die untersuchten 42 Indikatoren ausreichend groß ($N_{Bew} > 420$). Die multivariate Normalverteilung wurde mit dem Mardia-Test überprüft. Die multivariate Kurtosis lag bei 256,057, der als z-Wert interpretierbare c.r.-Wert bei c.r. = 57,287, d.h. es lag mit hoher Wahrscheinlichkeit keine multivariate Normalverteilung der Daten vor, weshalb eine Korrektur des χ^2 -Wertes durch das Bollen-Stine-Bootstrap-Ver-

fahren angezeigt war. Die Überprüfung der Multikollinearität wurde anhand der Interitemkorrelationsmatrix vorgenommen. Da keine der Korrelationen größer als $r = .85$ war, konnte davon ausgegangen werden, dass keine Multikollinearität die Zusammenhänge der Daten beeinflusste.

Sämtliche Items der Subskala Ordnung luden auf dem eigenen Faktor ($\lambda_O = 0,13 - 0,75$). Auch die Items der Subskala Fleiß ($\lambda_F = 0,39 - 0,70$) sowie der Gewissenhaftigkeitsdimension luden auf den jeweils zugehörigen Faktor ($\lambda_{GH} = 0,33 - 0,66$).

Das Modell wurde durch den Modelltest abgelehnt, da $\chi^2(816) = 2564,675$ ($p = 0,000$). Da dieses nicht signifikante Ergebnis auch der großen Stichprobe geschuldet sein konnte, wurden zusätzlich die approximativen Fit-Indices *RMSEA* und *SRMR* zur Überprüfung des Modell-Fits herangezogen. Der *RMSEA* lag mit $RMSEA = 0,054$ unter dem in Bühner (2011) definierten Grenzwert für große Stichproben ($RMSEA \leq 0,06$). Der *SRMR* lag mit $SRMR = 0,0661$ unter der in Bühner (2011) genannten Grenze von $SRMR \leq 0,11$. Insgesamt deuteten die Fit-Indices somit auf die Gültigkeit des angenommenen Modells hin.

Unter den gegebenen strengen Cut-off-Werten kann daher unter Berücksichtigung der Fit-Indizes sowie der Faktorladungen davon ausgegangen werden, dass die Daten den theoretischen Modellannahmen entsprachen und man von der Konstruktvalidität der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung hinsichtlich ihrer Zugehörigkeit zur Gewissenhaftigkeitsdimension ausgehen kann. Hypothese 5a kann somit angenommen werden.

Die diskriminante Validität der AGS bzw. der Gewissenhaftigkeitsaspekte Fleiß und Ordnung wurde wiederum mittels konfirmatorischer Faktorenanalysen zur Überprüfung des Modell-Fits mit den beiden divergenten Variablen Neurotizismus sowie Offenheit aus dem NEO-FFI (Borkenau & Ostendorf, 2008) bestimmt. Sowohl die Neurotizismus- als auch die Offenheitsskala erfassen das zugehörige Konstrukt jeweils mit zwölf Items. Es wurde in zwei getrennten CFA-Analysen überprüft, inwiefern ein Zusammenhang mit den konstruktfernen Faktoren Neurotizismus und Offenheit existiert. Als Schätzmethode wurde jeweils die Maximum-Likelihood-Methode gewählt. Die Stichprobengröße war mit jeweils $N_{Bew} = 740$ Probanden für die in beiden Fällen untersuchten 42 Indikatoren ausreichend groß ($N_{Bew} > 420$). Die multivariate Normalverteilung wurde mit dem Mardia-Test überprüft.

Für beide Diskriminanzmodelle wiesen sämtliche Items der Subskala Ordnung gleiche Faktorladungen auf ($\lambda_O = 0,15 - 0,75$). Die Items der Subskala Fleiß luden ebenfalls auf den jeweils zugehörigen Faktor ($\lambda_{F;N} = 0,40 - 0,69$ bzw. $\lambda_{F;OFF} = 0,39 - 0,69$). Die Neurotizismus-items luden von $\lambda_N = 0,13$ bis $\lambda_N = 0,61$ auf die zugehörige Dimension des NEO-FFI. Auch die Items der Offenheitsskala luden ausschließlich auf den eigenen Faktor ($\lambda_{OFF} = 0,01 - 0,57$). Ein Item der Offenheitsskala wies eine sehr geringe negative Faktorladung auf ($\lambda_{OFF1} = -0,02$).

In beiden Modellen konnte man nicht von einer multivariaten Normalverteilung der Daten ausgehen, da die multivariate Kurtosis mit 187,525 (Neurotizismus) bzw. 174,131 (Offenheit) und die *c.r.*-Werte mit *c.r.* = 41,955 (Neurotizismus) und *c.r.* = 38,958 (Offenheit) sehr hoch waren. Die Überprüfung der Multikollinearität wurde anhand der Interitemkorrelationsmatrizen vorgenommen. Da keine der Korrelationen größer als $r = .85$ war, konnte davon ausgegangen werden, dass keine Multikollinearität die Zusammenhänge der Daten in den beiden Analysen beeinflusste.

Sowohl für Neurotizismus ($\chi^2(816) = 2144,431$, $p = 0,000$) als auch für Offenheit ($\chi^2(816) = 2414,067$, $p = 0,000$) wurde das Modell durch den Modelltest abgelehnt. Zusätzlich wurden die approximativen Fit-Indices *RMSEA* und *SRMR* interpretiert.

Der *RMSEA* lag für die Überprüfung der divergenten Validität mit der Neurotizismusskala mit $RMSEA = 0,047$ unter dem mit $RMSEA \leq 0,06$ definierten Grenzwert für große Stichproben.

Der *SRMR* befand sich mit $SRMR = 0,0599$ ebenfalls unterhalb der kritischen Grenze von $SRMR \leq 0,11$.

Für die Überprüfung der divergenten Validität mit der Offenheitsskala ergaben sich vergleichbare Fit-Indizes. So lag der *RMSEA* mit 0,051 ebenfalls unter dem mit $RMSEA \leq 0,06$ definierten Grenzwert. Der *SRMR* war mit 0,0689 ebenfalls niedriger als der Cut-off-Wert von $SRMR \leq 0,11$.

Unter den gegebenen strengen Cut-off-Werten kann daher unter Berücksichtigung der Fit-Indizes sowie der Faktorladungen davon ausgegangen werden, dass die Daten in der Bewerberstichprobe den theoretischen Modellannahmen entsprachen. Die Gewissenhaftigkeitsaspekte Fleiß und Ordnung unterscheiden sich inhaltlich und strukturell von den durch die NEO-FFI-Skalen erfassten Konstrukten Neurotizismus und Offenheit. Die Hypothesen 5b und 5c können somit angenommen werden.

8.3 Vorhersage von Schul- bzw. Berufsschulnoten

Im zweiten Analyseteil dieser Studie wurde die Vorhersagekraft der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie der Selbstwirksamkeit und wahrgenommenen eigenen Fähigkeiten in Bezug auf Schul- bzw. Berufsschulnoten untersucht. Zur Überprüfung der Zusammenhänge der einzelnen Variablen mit den Kriteriumsdaten und weiteren Variablen wurden Korrelationsanalysen (8.3.2) berechnet. Anschließend wurde mittels Moderationsanalysen überprüft, inwiefern die Konstrukte Selbstwirksamkeit und wahrgenommene eigene Fähigkeiten die Vorhersagegüte des Fleiß- bzw. Ordnungsaspektes in Bezug auf die Kriterien beeinflussen (Abschnitt 8.3.3). Im letzten Abschnitt (8.3.4) sind die Ergebnisse der Mediationsanalysen enthalten, die überprüfen, inwiefern Selbstwirksamkeit oder die wahrgenommenen eigenen Fähigkeiten den Zusammenhang zwischen den Gewissenhaftigkeitsaspekten Fleiß und Ordnung mit den verschiedenen Leistungskriterien vermittelt. Vorab werden jedoch noch die Voraussetzungen (8.3.1) für die Durchführung der o.g. Analysemethoden überprüft.

Als Grundlage für die im Folgenden beschriebenen Analysen wurde für die beiden Subskalen Fleiß und Ordnung jeweils der Skalenmittelwert verwendet.

8.3.1 Überprüfung der Voraussetzungen der Dateneignung

Um Korrelationsanalysen berechnen zu können, muss in den Daten stochastische Unabhängigkeit, Normalverteilung (8.3.1.1) sowie Varianzhomogenität (8.3.1.2) vorliegen. Außerdem muss eine ausreichende Stichprobengröße (8.3.1.3) vorhanden sein, um die statistische Aussagekraft der Ergebnisse sicherzustellen. Für Regressionsanalysen muss zudem die Multikollinearität (8.3.1.4) der Daten überprüft werden.

8.3.1.1 Normalverteilung der Daten

Die Überprüfung der Normalverteilung der Daten wurde bereits teilweise in Abschnitt 8.2.1.1 für die Gewissenhaftigkeitsaspekte Fleiß und Ordnung vorgenommen. In der kaufmännischen Subgruppe wurde aufgrund der kleinen Stichprobe ($N_{kfm} = 31$) der Saphiro-Wilk-Test verwendet. Den Ergebnissen zufolge waren die relevanten Variablen Fleiß ($p = .17$), Ordnung ($p = .67$), Selbstwirksamkeit ($p = .26$) sowie wahrgenommene eigene Fähigkeiten ($p = .60$) normalverteilt. Auch in der Subgruppe der technischen Auszubildenden ($N_{techn} = 32$) deuteten die Ergebnisse des Saphiro-Wilk-Tests auf Normalverteilung der Variablen Fleiß ($p = .42$), Ordnung ($p = .47$), Selbstwirksamkeit ($p = .12$) sowie wahrgenommene eigene Fähigkeiten ($p = .11$) hin. In der Kontrollgruppe ($N_{KG} = 68$) wurde zur Überprüfung der Normalverteilung der Variablen der Kolmogorov-Smirnov-Test verwendet. Dieser deutete für die Variablen Fleiß ($p = .10$), Ordnung ($p = .08$) und wahrgenommene eigene Fähigkeiten ($p = .10$) auf eine

Normalverteilung der Daten hin. Für die Variable Selbstwirksamkeit konnte eine Normalverteilung der Daten nicht bestätigt werden ($p < .05$).

8.3.1.2 Varianzhomogenität

Die Varianzhomogenität wurde mittels Levene-Tests überprüft. In der Stichprobe der Auszubildenden wiesen die zu T1 erhobenen Prädiktoren Fleiß ($F = 0,493$, $p = .49$), Ordnung ($F = 0,839$, $p = .36$) und Selbstwirksamkeit ($F = 0,394$, $p = .53$) Varianzhomogenität auf. Auch die zu T2 erhobenen Variablen Fleiß ($F = 0,684$, $p = .41$), Ordnung ($F = 0,265$, $p = .61$), Selbstwirksamkeit ($F = 0,245$, $p = .62$) und wahrgenommene eigene Fähigkeiten ($F = 1,039$, $p = .31$) wiesen gleiche Varianzen auf. Auch in der Kontrollgruppe erwiesen sich sämtliche Varianzen der Prädiktoren Fleiß ($F = 0,413$, $p = .52$), Ordnung ($F = 0,001$, $p = .97$), Selbstwirksamkeit ($F = 1,302$, $p = .26$) und wahrgenommene Fähigkeiten ($F = 0,593$, $p = .44$) als homogen. Insgesamt kann diese Voraussetzung somit für alle untersuchten Stichproben bzw. Variablen als gegeben angesehen werden.

8.3.1.3 Stichprobengröße

Laut Müller (2007) sollte eine Stichprobe zur Berechnung von Moderations- und Mediationsanalysen zudem mindestens fünf bis zehn Datensätze pro analysierter Prädiktorvariable nicht unterschreiten, um eine ausreichende statistische Aussagekraft der Ergebnisse zu erreichen. Dies wurde in allen drei Stichproben mit $N_{kfM} = 31$, $N_{techn} = 32$ und $N_{KG} = 68$ nur knapp sichergestellt. Je nach untersuchter Kriteriumsvariable schwankte die Stichprobengröße der technischen bzw. kaufmännischen Auszubildenden um einige wenige Probanden. Es ist möglich, dass es aufgrund der mäßigen Zellbesetzungen hinsichtlich der Aussagekraft der Ergebnisse zu Einschränkungen kommen kann.

8.3.1.4 Multikollinearität

Die Überprüfung der Multikollinearität ergab keine überhöhten Interkorrelationen der Prädiktorvariablen (s. Tabellen 8-9 bis 8-11). Lediglich der Zusammenhang des Fleißaspekts und der Selbstwirksamkeit (beides T1) in der Substichprobe der technischen Auszubildenden wies eine sehr hohe Interkorrelation auf ($r = .87$, $p < .01$), die als Hinweis auf das Vorliegen von Multikollinearität dienen könnte.

8.3.2 Ergebnisse der Korrelationsanalysen

Zur Überprüfung der Hypothesen 6a bis 6d zur Ermittlung der einfachen Zusammenhänge zwischen den dispositionellen Variablen Fleiß, Ordnung, Selbstwirksamkeit und den wahrgenommenen eigenen Fähigkeiten mit den Leistungsvariablen (Schul- bzw. Berufsschulnoten sowie Vorgesetztenbeurteilung) wurden für sowohl für die Bewerber- als auch für die Schülerkontrollgruppe Korrelationsanalysen (Abschnitt 6.3.1) berechnet. Die Ergebnisse sind in den Interkorrelationstabellen 8-9 bis 8-11 dargestellt.

Da in der Bewerberstichprobe einige Kriteriumsdaten (IHK-Noten) nur für die technischen Auszubildenden vorhanden waren und zudem aufgrund inhaltlicher Überlegungen (vgl. Abschnitt 3.2.2) eine Unterscheidung und somit getrennte Analyse kaufmännischer und technischer Auszubildenden angedacht wurde, werden im Folgenden alle Analyseschritte zur Vorhersage- und Modellüberprüfung getrennt für diese beiden Subgruppen vorgenommen. Im Anhang ist zudem eine Übersicht der Interkorrelationen der verschiedenen Variablen für die Gesamtstichprobe der eingestellten Bewerber (Anhang Tabelle A-16 Interkorrelationstabelle Retest-Stichprobe) enthalten.

Tabelle 8-9: Interkorrelationstabelle kaufmännische Auszubildende

	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10	11	12	13
Prädiktoren T1															
1 T1 Fleiß	3,13	.32													
2 T1 Ordnung	3,21	.34	.38*												
			(.14)												
3 T1 Selbstwirksamkeit	3,07	.30	.66**	.36**											
			(.43)	(.13)											
4 T1 GMA	62,93	3,66	-.24	.07	-.10										
			(.06)	(.00)	(.01)										
Prädiktoren T2															
5 T2 Fleiß	3,00	.29	.37*	-.06	.15	-.19									
			(.14)	(.00)	(.02)	(.03)									
6 T2 Ordnung	3,01	.34	.04	.48**	.04	.03	.04								
			(.00)	(.26)	(.00)	(.00)	(.00)								
7 T2 Selbstwirksamkeit	2,80	.31	.19	-.13	.36*	-.12	.61**	.03							
			(.04)	(.02)	(.13)	(.02)	(.37)	(.00)							
8 T2 PA	1,48	.28	-.32	-.11	-.28	-.06	-.02	.02	-.17						
			(.10)	(.01)	(.08)	(.00)	(.00)	(.00)	(.03)						
Kriterien															
9 Allg. Berufsschulnote ^a	1,93	.41	-.03	-.30	-.07	-.13	.21	-.13	.11	.32					
			(.00)	(.09)	(.00)	(.02)	(.04)	(.02)	(.01)	(.10)					
10 VG-Beurteilung	2,58	.77	-.11	.20	-.01	.02	.05	.06	-.07	.28	.58**				
			(.01)	(.04)	(.00)	(.00)	(.00)	(.00)	(.00)	(.08)	(.34)				
Kontrollvariablen															
11 Soziale Erwünschtheit	12,45	2,43	.03	.09	.03	-.16	.47**	.33	.38*	.11	.01	.14			
			(.00)	(.01)	(.00)	(.02)	(.22)	(.11)	(.14)	(.01)	(.00)	(.02)			
12 Alter	18,16	1,24	-.01	-.08	.22	.03	.07	.06	.24	.16	-.39*	-.28	-.05		
			(.00)	(.01)	(.05)	(.00)	(.01)	(.00)	(.06)	(.03)	(.15)	(.08)	(.00)		
13 Geschlecht			-.18	-.05	-.05	.19	-.35	-.08	-.01	-.03	.06	.06	-.17	.02	
			(.03)	(.00)	(.00)	(.04)	(.12)	(.01)	(.00)	(.00)	(.00)	(.00)	(.03)	(.00)	
14 Schulbildung			.03	-.01	.00	-.30	.02	-.03	.05	.08	.15	-.26	-.11	.20	-.18
			(.00)	(.00)	(.00)	(.09)	(.00)	(.00)	(.00)	(.01)	(.02)	(.07)	(.01)	(.04)	(.03)

Anmerkungen: $N_{kfM} = 31$; ^a $n = 27$ Berufsschulnoten; *M*= Mittelwert; *SD*= Standardabweichung; PA= wahrgenommene eigene Fähigkeiten; GMA= gemittelter T-Wert allgemeine kognitive Fähigkeiten; VG-Beurteilung= Vorgesetztenbeurteilung; ** $p < 0,01$ (zweiseitig); * $p < 0,05$ (zweiseitig); Werte in Klammern entsprechen r^2 .

Tabelle 8-10: Interkorrelationstabelle technische Auszubildende

	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Prädiktoren T1																		
1 T1 Fleiß	3,07	.38																
2 T1 Ordnung	3,07	.42	.52**															
			(.27)															
3 T1 SE	3,05	.38	.87**	.50**														
			(.75)	(.25)														
4 T1 GMA	60,55	4.94	-.10	-.18	-.01													
			(.01)	(.03)	(.00)													
Prädiktoren T2																		
5 T2 Fleiß	3,13	.33	.67**	.42*	.63**	.11												
			(.45)	(.17)	(.40)	(.01)												
6 T2 Ordnung	3,14	.35	.31	.54**	.41*	.13	.48**											
			(.10)	(.29)	(.17)	(.02)	(.22)											
7 T2 SE	2,92	.36	.49**	.17	.63**	.24	.65**	.33										
			(.24)	(.03)	(.39)	(.06)	(.43)	(.11)										
8 T2 PA	1,54	.37	-.46**	-.35	-.53**	-.26	-.55**	-.41*	-.65**									
			(.21)	(.12)	(.28)	(.07)	(.30)	(.16)	(.42)									
Kriterien																		
9 Schriftl. Prüfung ^a	81,69	8.72	.16	.07	.08	-.01	.09	-.23	.04	-.04								
			(.03)	(.00)	(.01)	(.00)	(.01)	(.05)	(.00)	(.00)								
10 Prakt. Prüfung ^a	79,77	11,69	.27	-.04	.04	-.29	.26	.18	.15	-.46*	.36							
			(.07)	(.00)	(.00)	(.08)	(.07)	(.03)	(.02)	(.21)	(.13)							
11 Gesamt- ergebnis ^b	81,07	8.60	.27	-.04	.07	-.30	.29	-.06	.16	-.27	.78**	.85**						
			(.07)	(.00)	(.00)	(.09)	(.08)	(.00)	(.03)	(.07)	(.60)	(.73)						
12 Allg. Berufs- schulnote ^c	2,17	.51	-.12	.13	-.16	-.02	-.22	-.06	-.30	.39*	-.66**	-.51**	-.69**					
			(.02)	(.02)	(.02)	(.00)	(.05)	(.00)	(.09)	(.15)	(.43)	(.26)	(.48)					
13 VG- Beurteilung ^d	3,26	.89	-.34	.00	-.31	-.45*	-.33	.02	-.42*	.45*	-.60**	-.56**	-.64**	.41*				
			(.12)	(.00)	(.09)	(.20)	(.11)	(.00)	(.18)	(.20)	(.36)	(.32)	(.41)	(.17)				
Kontrollvariablen																		
14 Soziale Erwünschtheit	12,47	3,06	.47**	.51**	.62**	-.14	.43*	.44*	.32	-.22	.24	-.07	-.02	-.06	.08			
			(.22)	(.26)	(.38)	(.02)	(.18)	(.19)	(.10)	(.05)	(.06)	(.00)	(.00)	(.00)	(.01)			
15 Alter	16,64	2,15	.18	.03	.31	.11	.15	-.04	.23	-.33	.25	-.25	.09	-.37*	.27	.19		
			(.03)	(.00)	(.09)	(.01)	(.02)	(.00)	(.05)	(.11)	(.06)	(.06)	(.01)	(.14)	(.07)	(.03)		
16 Geschlecht			-.12	-.32*	-.13	-.10	-.15	-.43*	-.24	.36*	-.10	-.07	.13	.02	.10	-.16	-.05	
			(.01)	(.10)	(.02)	(.01)	(.02)	(.19)	(.06)	(.13)	(.01)	(.00)	(.02)	(.00)	(.01)	(.03)	(.00)	
17 Schulbildung			.17	.10	.23	.41**	.19	.03	.34	-.51**	.29	-.06	.02	-.28	-.32	.10	.62**	-.43**
			(.03)	(.01)	(.05)	(.17)	(.04)	(.00)	(.11)	(.26)	(.09)	(.00)	(.00)	(.08)	(.10)	(.01)	(.38)	(.18)

Anmerkungen: $N_{\text{techn}} = 39$; $n_{\text{IHK}} = 26$ Zwischenprüfungsnoten; $n_{\text{IHK}} = 30$ Zwischenprüfungsgesamtnoten; $n_{\text{BS}} = 30$ Berufsschulnoten; $n_{\text{VG}} = 31$ Vorgesetztenbeurteilungen; *M*= Mittelwert; *SD*= Standardabweichung; SE= Selbstwirksamkeit; PA= wahrgenommene eigene Fähigkeiten; GMA= Allgemeine kognitive Fähigkeiten; VG-Beurteilung= Vorgesetztenbeurteilung; ** $p < 0,01$ (zweiseitig); * $p < 0,05$ (zweiseitig); Werte in Klammern entsprechen r^2 .

Tabelle 8-11: Interkorrelationstabelle Kontrollgruppe

	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9
Prädiktoren											
1 Fleiß	2,67	.41									
2 Ordnung	2,77	.47	.35** (.12)								
3 SE	2,63	.49	.72** (.51)	.36** (.13)							
4 PA	2,29	.36	.30** (.09)	.32** (.10)	.37** (.14)						
Kriterium											
5 Note ^a	2,83	.77	-.24* (.06)	-.30* (.09)	-.35** (.12)	-.21 (.05)					
Kontrollvariablen											
6 Note letztes Zeugnis	2,84	.54	-.26* (.07)	-.26* (.07)	-.27** (.07)	-.30** (.09)	.59** (.35)				
7 Soziale Erwünschtheit	10,31	3,15	.29** (.08)	.37** (.14)	.24* (.06)	.19 (.04)	-.33** (.11)	-.26* (.07)			
8 Alter	16,28	1,73	.22* (.05)	.01 (.00)	.17 (.03)	.35** (.12)	-.12 (.02)	-.05 (.00)	-.07 (.01)		
9 Schulart			.23* (.05)	-.02 (.00)	.13 (.02)	.40** (.16)	-.16 (.03)	-.20 (.04)	-.11 (.01)	.86** (.74)	
10 Berufl. Vorerfahrungen			.02 (.00)	.15 (.02)	-.11 (.01)	.13 (.02)	.07 (.00)	-.06 (.00)	.10 (.01)	.07 (.00)	.09 (.01)

Anmerkungen: $N_{KG} = 97$ gesamt; ^a $n_{KG, Note} = 69$; *M*= Mittelwert; *SD*= Standardabweichung; SE= Selbstwirksamkeit; PA= wahrgenommene eigene Fähigkeiten; Note= Schulabschlussnote; ** $p < 0,01$ (zweiseitig); * $p < 0,05$ (zweiseitig); Werte in Klammern entsprechen r^2 .

Hypothese 6a nahm einen Zusammenhang zwischen dem Fleißaspekt und den Leistungsdaten an. In der Subgruppe der kaufmännischen Auszubildenden ($n_{kfm} = 31$) korrelierte der Fleißaspekt, der im Rahmen der Einstellungstestung erhoben wurde, entgegen der Annahmen weder mit der allgemeinen Berufsschulnote ($r = -.03$, n.s.), noch mit der Vorgesetztenbeurteilung ($r = -.11$, n.s.). Auch die zweite Erhebung des Fleißaspekts im Rahmen der Retestung führte zu keinen signifikanten Zusammenhängen mit der allgemeinen Berufsschulnote ($r = .21$, n.s.) oder der Vorgesetztenbeurteilung ($r = .05$, n.s.). Der Zusammenhang mit der Berufsschulnote kann jedoch als Tendenz gewertet werden. Interessanterweise zeigten sich zu T1 negative Zusammenhänge, während sich zu T2 positive Zusammenhänge ergaben. Einschränkend muss hierzu angemerkt werden, dass zu T2 aufgrund der Auswahl der Auszubildenden nach T1 eine höhere Range Restriction vorlag. Hypothese 6a kann diesen Ergebnissen zufolge für die kaufmännischen Auszubildenden nicht bestätigt werden.

Für die Subgruppe der Techniker ($n_{techn} = 39$) ergaben sich ebenfalls keine signifikanten Zusammenhänge des Fleißaspekts mit den verschiedenen Leistungskriterien, allerdings zeigten sich im Vergleich zu den kaufmännischen Auszubildenden keine Vorzeichenunterschiede bei den verschiedenen Zusammenhängen zu T1 und T2. Die allgemeine Berufsschulnote der technischen Auszubildenden zeigte einen tendenziell negativen Zusammenhang mit Fleiß ($r_{T1} = -.12$ bzw. $r_{T2} = -.22$, n.s.). Das heißt, tendenziell waren die technischen Auszubildenden in der Berufsschule umso besser, je höher ihre Ausprägung im Fleißaspekt war. Bezüglich der Zusammenhänge mit den IHK-Zwischenprüfungsergebnissen ergeben sich positive Zusammenhänge, da hier nicht die Noten, sondern die erreichte Prozentzahl Grundlage der Korrelationsanalysen war. Hier zeigten sich geringe, nicht-signifikante Zusammenhänge des Fleißaspekts mit den Ergebnissen der schriftlichen Zwischenprüfung ($r_{T1} = .16$ bzw. $r_{T2} = .09$, n.s.). Die Zusammenhänge mit den Ergebnissen der mündlichen Zwischenprüfung waren ebenfalls nicht signifikant, zeigten aber deutliche Tendenzen ($r_{T1} = .27$ bzw. $r_{T2} = .26$, n.s.), was sich entsprechend auch in den Zusammenhängen mit dem Gesamtergebnis der Zwischenprüfung niederschlug ($r_{T1} = .27$ bzw. $r_{T2} = .29$, n.s.). Bezüglich des Zusammenhangs mit der Vorgesetztenbeurteilung zeigten sich in der Subgruppe der technischen Auszubildenden deutlich Effekte dahingehend, dass diese tendenziell durch den Fleißaspekt vorhergesagt werden konnte ($r_{T1} = -.34$ bzw. $r_{T2} = -.33$, n.s.). Das heißt, die Ausbildungsleiter bewerteten fleißigere Auszubildende tendenziell besser. Hypothese 6a kann somit nicht für diese Analysestichprobe bestätigt werden, sollte aufgrund der Tendenzen jedoch auch nicht verworfen werden.

In der Kontrollgruppe zeigte sich ein signifikanter Zusammenhang zwischen dem Fleißaspekt und der Abschlussnote ($r = -.24$, $p < .05$). Das heißt, je höher die Ausprägung im Fleißaspekt, desto geringer bzw. besser war die Abschlussnote eines Schülers. Wenn auch die Unreliabilität des Prädiktors sowie des Kriteriums mit berücksichtigt wurden, stieg die Korrelation auf $r_{xy} = -.30$. Für die Kontrollgruppe konnte die Hypothese 6a somit bestätigt werden.

Hypothese 6b nahm einen Zusammenhang des Ordnungsaspekts mit den Leistungsdaten an. Für die Subgruppe der kaufmännischen Auszubildenden ergaben sich mit dem Kriterium der allgemeinen Berufsschulnote keine signifikanten, aber dennoch bedeutsame negative Zusammenhänge für beide Erhebungsdaten ($r_{T1} = -.30$ bzw. $r_{T2} = -.13$, n.s.). Für die Vorgesetztenbeurteilung zeigten sich positive, ebenfalls nicht-signifikante Effekte ($r_{T1} = .20$ bzw. $r_{T2} = .06$, n.s.), was bedeutet, dass die Vorgesetzten sehr ordentliche Auszubildende tendenziell schlechter bewerteten. Hypothese 6b kann somit nicht für die Analysestichprobe der kaufmännischen Auszubildenden bestätigt werden, sollte aufgrund der positiven Tendenzen, v.a. zu T1 jedoch auch nicht verworfen werden.

Für die Subgruppe der technischen Auszubildenden zeigten sich in Bezug auf die Vorgesetztenbeurteilung keine Zusammenhänge ($r_{T1} = .00$ bzw. $r_{T2} = .02$, n.s.). Die allgemeine Berufsschulnote korrelierte ebenfalls nicht-signifikant mit dem Ordnungsaspekt ($r_{T1} = .13$ bzw.

$r_{T2} = -.06$, n.s.). Bezüglich der Zusammenhänge mit den Ergebnissen der IHK-Zwischenprüfung zeigten sich differenzierte Effekte. So zeigte der Ordnungsaspekt, der zu T1 erhoben wurde, kaum Zusammenhänge zu den schriftlichen oder mündlichen Prüfungsergebnissen ($r_{\text{schriftl.}} = .07$ bzw. $r_{\text{mündl.}} = -.04$, n.s.), was sich entsprechend auch im Zusammenhang mit dem Gesamtergebnis der Zwischenprüfung niederschlug ($r_{\text{ZWP}} = -.04$, n.s.). Der zu T2 erhobene Ordnungsaspekt zeigte demgegenüber höhere, nicht-signifikante Zusammenhänge mit den schriftlichen ($r_{\text{schriftl.}} = -.23$, n.s.) und mündlichen Prüfungsergebnissen ($r_{\text{mündl.}} = .18$, n.s.). Diese entgegengesetzt gepolten Zusammenhänge könnten den fehlenden Zusammenhang zwischen dem Ordnungsaspekt und dem Gesamtergebnis der IHK-Zwischenprüfung ($r_{T1} = -.06$, n.s.) erklären. Hypothese 6b kann somit für diese Analytestichprobe nicht bestätigt werden.

In der Kontrollgruppe korreliert der Ordnungsaspekt signifikant negativ mit den Abschlussnoten der Schüler ($r_{T1} = -.30$, $p < .05$). Die Korrelation erhöhte sich nach der Attenuitätskorrektur auf $r_{xy} = -.39$. Für diese Stichprobe konnte Hypothese 6b somit bestätigt werden.

Hypothese 6c postulierte einen Zusammenhang zwischen dem Selbstwirksamkeitskonstrukt und den verschiedenen Leistungskriterien.

In der Teilstichprobe der kaufmännischen Auszubildenden ergaben sich keine Zusammenhänge. Weder die allgemeine Berufsschulleistung ($r_{T1} = -.07$ bzw. $r_{T2} = .11$, n.s.) noch die Vorgesetztenbeurteilungen ($r_{T1} = -.01$ bzw. $r_{T2} = -.07$, n.s.) zeigten signifikante Zusammenhänge mit der Selbstwirksamkeit. Hypothese 6c kann somit für die Gruppe der kaufmännischen Auszubildenden nicht bestätigt werden.

Auch in der Teilstichprobe der technischen Auszubildenden zeigten sich keine signifikanten Zusammenhänge zwischen der Selbstwirksamkeit und der allgemeinen Berufsschulleistung. Allerdings waren die gefundenen Zusammenhänge deutlich höher als bei den kaufmännischen Auszubildenden ($r_{T1} = -.16$ bzw. $r_{T2} = -.30$, n.s.). Für die Vorgesetztenbeurteilung zeigte sich demgegenüber ein deutlicher Effekt dahingehend, dass Selbstwirksamkeit, die zu T2 erhoben wurde, mit dieser signifikant korrelierte ($r_{T1} = -.31$ bzw. $r_{T2} = -.42$, $p < .05$) und auch zu T1 einen positiven Effekt zeigt in dem Sinne, dass Auszubildende mit einer höheren Selbstwirksamkeit auch besser von den Ausbildungsleitern beurteilt wurden. In Bezug auf die IHK-Zwischenprüfung zeigten sich keine signifikanten Ergebnisse für das Gesamtergebnis ($r_{T1} = .07$ bzw. $r_{T2} = .16$, n.s.). Auch in Bezug auf die zwei Bestandteile der Zwischenprüfung zeigten sich keine differenzierten Effekte für die schriftlichen ($r_{T1} = .08$ bzw. $r_{T2} = .04$, n.s.) und mündlichen Prüfungsergebnisse ($r_{T1} = .04$ bzw. $r_{T2} = .15$, n.s.). Für diese Analytestichprobe kann die Hypothese 6c somit nur teilweise bestätigt werden.

In der Kontrollgruppe zeigte sich ein hochsignifikanter Zusammenhang des Selbstwirksamkeitskonstrukts mit den Abschlussnoten der Schüler ($r = -.30$, $p < .01$). Schüler mit hohen Ausprägungen im Merkmal Selbstwirksamkeit erhielten bessere Schulabschlussnoten als weniger selbstwirksame Schulabgänger. Dieser Effekt verstärkte sich nach der Attenuitätskorrektur auf $r_{xy} = -.41$. Hypothese 6c kann für die Kontrollgruppe somit bestätigt werden.

Hypothese 6d postulierte einen Zusammenhang zwischen den wahrgenommenen eigenen Fähigkeiten einer Person und den verschiedenen Leistungskriterien in dieser Untersuchung.

Für die Teilstichprobe der kaufmännischen Auszubildenden ergeben sich für die wahrgenommenen eigenen Fähigkeiten sowohl in Bezug auf die allgemeine Berufsschulleistung ($r = .32$, n.s.) als auch für die Vorgesetztenbeurteilung ($r = .28$, n.s.) positive Effekte, die jedoch nicht signifikant werden. In der Teilstichprobe der technischen Auszubildenden konnten demgegenüber signifikante Korrelationen beobachtet werden. Die Einschätzungen der wahrgenommenen eigenen Fähigkeiten korrelierten signifikant sowohl mit der allgemeinen Berufsschulleistung ($r = .39$, $p < .05$) als auch mit der Vorgesetztenbeurteilung ($r = .45$, $p < .05$).

Bezüglich der IHK-Zwischenprüfungsergebnisse zeigten sich unterschiedliche Ergebnisse. Hier korrelierten die wahrgenommenen eigenen Fähigkeiten lediglich mit den Ergebnissen der praktischen Prüfung signifikant ($r = -.46$, $p < .05$), nicht aber mit den schriftlichen Zwischenprüfungsergebnissen ($r = -.04$, n.s.). Dies führte vermutlich auch zu dem nicht-signifikanten mäßigen Zusammenhang des Konstrukts mit dem Gesamtergebnis der Zwischenprüfung ($r = -.27$, n.s.). Die Hypothese 6d konnte somit für die Auszubildenden nur teilweise bestätigt werden.

In der Kontrollgruppe zeigte sich kein signifikanter Zusammenhang zwischen den wahrgenommenen eigenen Fähigkeiten und der Schulabschlussnote ($r = -.21$, n.s.). Die Hypothese 6d muss somit für diese Stichprobe verworfen werden.

Wie in Abschnitt 6.3.2 beschrieben, unterlagen die Daten der Korrelationsanalysen einigen Einschränkungen, die möglicherweise zu einer Unterschätzung der Zusammenhänge geführt haben könnten. Zu diesen Einschränkungen zählten eine mögliche Range Restriction (s. Abschnitt 6.3.2.1), die aufgrund der Vorselektion der Bewerberstichprobe entstanden sein könnte, sowie eine Einschränkung der Varianz aufgrund der Unreliabilität der verwendeten Messverfahren, was mit Hilfe einer Minderungskorrektur (vgl. Abschnitt 6.3.2.2) kontrolliert wurde.

Tabelle 8-12: Ergebnisse der Korrekturberechnungen der Korrelationen in der Bewerberstichprobe

	Kaufmännische Azubis				Technische Azubis			
	r_{xy}	$r_{x'y'}$	σ_{xy}	$\sigma_{x'y'}$	r_{xy}	$r_{x'y'}$	σ_{xy}	$\sigma_{x'y'}$
Fleiß								
Berufsschulnote	.03 (.21)	.04 (.26)	.04	.04	-.12 (-.22)	-.14 (-.27)	-.12	-.14
Vorgesetztenbeurteilung ^a	-.11 (.05)	-.12 (.06)	-.11	-.12	-.34 (-.33)	-.36 (-.37)	-.34	-.36
IHK-Zwischenprüfung (mündl.)	-	-	-	-	.27 (.26)	.32 (.32)	.27	.32
IHK-Zwischenprüfung (schriftl.)	-	-	-	-	.16 (.09)	.19 (.11)	.16	.19
IHK-Zwischenprüfung (gesamt)	-	-	-	-	.27 (.29)	.32 (.36)	.27	.32
Ordnung								
Berufsschulnote	-.30 (-.13)	-.38 (-.17)	-.32	-.40	.13 (-.06)	.17 (-.08)	.11	.15
Vorgesetztenbeurteilung ^a	.20 (.06)	.23 (.07)	.21	.24	.00 (.02)	.00 (.02)	.00	.00
IHK-Zwischenprüfung (mündl.)	-	-	-	-	-.04 (.18)	-.05 (.24)	-.04	-.04
IHK-Zwischenprüfung (schriftl.)	-	-	-	-	.07 (-.23)	.09 (-.31)	.06	.08
IHK-Zwischenprüfung (gesamt)	-	-	-	-	-.04 (-.06)	-.05 (-.08)	-.04	-.04
Selbstwirksamkeit								
Berufsschulnote	-.07 (.11)	(-.09) (.13)	-.08	-.08	-.16 (-.30)	-.20 (-.37)	-.15	-.18
Vorgesetztenbeurteilung ^a	-.01 (-.07)	-.01 (-.08)	-.01	-.01	-.31 (-.42)	-.34 (-.71)	-.29	-.32
IHK-Zwischenprüfung (mündl.)	-	-	-	-	.04 (.15)	.05 (.18)	.04	.05
IHK-Zwischenprüfung (schriftl.)	-	-	-	-	.08 (.04)	.10 (.05)	.07	.09
IHK-Zwischenprüfung (gesamt)	-	-	-	-	.07 (.16)	.08 (.20)	.06	.08

Anmerkungen: Werte in Klammern entsprechen zu T2 erhobenen Merkmalsausprägungen; ^a bei der Vorgesetztenbeurteilung jeweils nur einfache Attenuitätskorrektur (Unreliabilität des Prädiktors).

Da in der Kontrollgruppe keine Kriteriumsinformationen über die ausgeschiedenen Schüler vorlagen, konnte die Range Restriction für diese Stichprobe nicht berechnet werden. Die Ergebnisse der Attenuitätskorrektur in der Kontrollgruppe wurden bereits zuvor genannt. Die Berechnung der Varianzeinschränkung innerhalb der Bewerber- bzw. Auszubildendenstichprobe konnte aufgrund der Vorselektion (Range Restriction) zudem nur für die Vorhersagekennwerte der bei T1 erhobenen Prädiktorvariablen berechnet werden. Die Ergebnisse dieser Korrekturberechnungen sind getrennt für die beiden Substichproben der kaufmännischen und technischen Auszubildenden bzw. Ausbildungsbewerber in der folgenden Tabelle 8-12 dargestellt.

Wie den zuvor in Tabelle 8-12 dargestellten korrigierten Korrelationswerten zu entnehmen ist, führten die Minderungskorrekturen jeweils zu keinen oder nur schwachen Verbesserungen der Korrelationskoeffizienten. Die Korrektur der Korrelationen der Vorhersagekennwerte, die zu T1 erhoben wurden, führte im Vergleich zur Attenuitätskorrektur ebenfalls teilweise zu marginalen Veränderungen der Korrelationen. Da die Varianz innerhalb der Gruppe der eingestellten Bewerber (= Retest-Stichprobe) teilweise auch größer war als die der Gesamtgruppe der Ausbildungsbewerber, ergaben sich für diese Merkmale teilweise auch geringere korrigierte Korrelationskoeffizienten. Dies betraf den Ordnungsaspekt sowie das Merkmal Selbstwirksamkeit in der Subgruppe der technischen Auszubildenden.

Aufgrund des heterogenen Antwortformats der Items zur Erfassung der wahrgenommenen eigenen Fähigkeiten konnten für dieses Kriterium keine korrigierten Korrelationen berechnet werden.

Eine Berechnung der Teststärke der Vorhersagekennwerte erfolgte mit Hilfe von G*Power 3 (Faul et al., 2007). Die Ergebnisse der Post-hoc Powerberechnungen zu den Korrelationskoeffizienten sind in Tabelle 8-13 dargestellt. Es zeigte sich, dass sich keine der Teststärken

Tabelle 8-13: Ergebnisse der Powerberechnungen für alle Stichproben

Kriterium	Teststärken							
	Kaufleute		Techniker					Schüler Note
	VG	BS	VG	BS	IHK mdl.	IHK schriftl.	IHK ges.	
Fleiß T1	.09	.05	.49	.10	.28	.12	.32	.53
Fleiß T2	.06	.19	.47	.22	.26	.07	.36	-
Ordnung T1	.20	.35	.05	.11	.05	.06	.06	.73
Ordnung T2	.06	.10	.05	.06	.15	.21	.06	-
SE T1	.05	.06	.42	.14	.05	.07	.07	.67
SE T2	.07	.09	.70	.38	.12	.05	.14	-
PA	.35	.39	.77	.61	.72	.05	.32	.83

Anmerkungen: SE= Selbstwirksamkeit; PA= wahrgenommene eigene Fähigkeiten; VG-Beurteilung= Vorgesetztenbeurteilung; BS= Berufsschulnote; IHK mdl.= IHK-Zwischenprüfungsnote mündlicher Teil; IHK prakt.= IHK-Zwischenprüfungsnote praktischer Teil; IHK ges.= IHK-Zwischenprüfungsnote gesamt.

in der Auszubildendenstichprobe über .80 befand. Dies war in Anbetracht der wenigen signifikanten Ergebnisse zu erwarten. Die geringen Teststärken deuteten an, dass die Studie „underpowered“ war, d.h. die untersuchten Stichproben waren zu klein, um die vermuteten Effekte zeigen zu können. Die gefundenen Korrelationskoeffizienten bewegten sich zwischen $r = .00$ und $r = .46$, wobei diese jedoch nur selten signifikant wurden (s. Tabellen 8-9 bis 8-11). Es handelte sich somit um kleine bis mittlere Effekte. Bortz und Döring (1995) geben an, dass für kleine Effekte ($r > .10$) mindestens $N = 618$ Versuchspersonen benötigt würden, um für $p < .05$ signifikante Ergebnisse zu erhalten. Für mittlere Effekte ($r > .30$) sind den Autoren zufolge mindestens $N = 68$ Versuchspersonen notwendig, um für $p < .05$ signifikante Ergebnisse zu erhalten. Mit $N = 26$ bis $N = 35$ werden diese optimalen Stichprobengrößen in der vor-

liegenden Untersuchung in den beiden Subgruppen der kaufmännischen und technischen Auszubildenden jedoch nicht erreicht.

Lediglich in der Kontrollgruppe ($N_{KG} = 69$) wird die optimale Stichprobengröße für mittlere Effekte knapp erreicht. Dies spiegelt sich auch in den Teststärken wider. Diese liegen mit $P = .53$ bis $.83$ unter bis über dem kritischen Wert.

Insgesamt waren die Korrelationskoeffizienten in den Auszubildendenstichproben nicht aussagekräftig genug, um zuverlässige Aussagen über die Vorhersagekraft der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie Selbstwirksamkeit und die wahrgenommenen eigenen Fähigkeiten tätigen zu können. Lediglich die Ergebnisse der Kontrollgruppe dürfen diesen Ergebnissen zufolge interpretiert werden.

8.3.3 Ergebnisse der Moderationsanalysen zur Modellüberprüfung

Zur Überprüfung des theoretischen Vorhersagemodells wurden Moderationsanalysen berechnet. Hypothese 7 postulierte, dass Selbstwirksamkeit bzw. wahrgenommene eigene Fähigkeiten sich aufgrund der Erfahrungen, die die Auszubildenden bzw. Schüler machen, verändern und somit die Vorhersagegüte des Fleiß- bzw. Ordnungsaspekts beeinflussen. Zuvor wurde jedoch auch überprüft, inwiefern eine Interaktion der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung in der Lage ist, mehr Varianz hinsichtlich der Kriterien aufzuklären als eine getrennte Betrachtung der beiden Faktoren.

Die Ergebnisse der Moderationsanalysen sind aufgrund der geringen Stichprobengrößen mit Vorbehalt zu betrachten.

Es zeigte sich in der Substichprobe der kaufmännischen Auszubildenden, dass der Interaktionsterm Fleiß x Ordnung zu keiner signifikanten Varianzaufklärung in Bezug auf die Kriteriumsvariable der Vorgesetztenbeurteilung beitrug ($R^2 = .08$, $F(1, 27) = 0,817$, $p = 0,50$), was auch Tabelle 8-14 zu entnehmen ist. Ähnliche Effekte ergaben sich auch für das Kriterium allgemeine Berufsschulnote. Mit $R^2 = .11$, $F(1, 23) = 0,974$, $p = .42$ und $\beta = -.920$, $t(28) = -0,253$, $p = .80$ trägt das Modell mit Interaktion nicht signifikant zur Varianzaufklärung bei.

Tabelle 8-14: Ergebnisse der Moderationsanalyse für die Gewissenhaftigkeitsaspekte Fleiß und Ordnung zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-0,505	.462	-0,213	-1,092	-1,677	4,521	-0,709	-0,371
Ordnung	0,631	.434	0,285	1,456	-0,474	4,264	-0,214	-0,111
Fleiß x Ordnung					0,342	1,312	0,826	0,261
R^2		.08				.08		
R^2 adj.		.02				-.02		
ΔR^2		.08				.00		
ΔF		1,232				0,068		

Anmerkungen: $N_{kf} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; $^\dagger p < .10$; * $p < .05$; ** $p < .01$.

Vergleichbare nicht-signifikante Effekte zeigten sich in der Subgruppe der technischen Auszubildenden. Mit $R^2 = .18$, $F(1, 27) = 1,990$, $p = 0,139$ ergab sich kein signifikanter Effekt für den Interaktionsterm Fleiß x Ordnung hinsichtlich des Kriteriums der Vorgesetztenbeurteilung (s. Tabelle 8-15). Für das Kriterium der allgemeinen Berufsschulnote ergab sich mit $R^2 = .08$, $F(1, 31) = 0,888$, $p = .46$ und $\beta = -.041$, $t(29) = 0,020$, $p = .98$ ebenfalls kein signifikanter Effekt für den Interaktionsterm. Dieser Trend setzt sich auch für das Kriterium der IHK-

Zwischennoten fort. Hier ergab sich mit $R^2 = .03$, $F(1, 22) = 0,232$, $p = .87$ und $\beta = -.437$, $t(23) = -0,179$, $p = .86$ für die schriftlichen Prüfungsergebnisse sowie $R^2 = .15$, $F(1, 22) = 1,286$,

Tabelle 8-15: Ergebnisse der Moderationsanalyse für die Gewissenhaftigkeitsaspekte Fleiß und Ordnung zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-1,050	0,465	-0,451	-2,260*	-3,352	2,495	-1,440	-1,344
Ordnung	0,450	0,396	0,227	1,135	-1,989	2,627	-1,003	-0,757
Fleiß x Ordnung					0,783	0,834	1,927	0,939
R^2		.15				.18		
R^2 adj.		.09				.09		
ΔR^2		.15				.03		
ΔF		2,555 [†]				0,882		

Anmerkungen: $N_{techn} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; [†] $p < .10$; * $p < .05$; ** $p < .01$.

$p = .30$ und $\beta = .306$, $t(25) = 0,134$, $p = .90$ für die mündlichen bzw. praktischen Zwischenprüfungsergebnisse kein signifikanter Effekt hinsichtlich einer erhöhten Varianzaufklärung. Für das Gesamtergebnis der Zwischenprüfung spiegelte sich diese Tendenz allerdings nicht wider. Diesbezüglich klärte die Interaktion Fleiß x Ordnung mit $R^2 = .19$, $F(1, 22) = 1,685$, $p = .20$ tendenziell mehr Varianz auf als die beiden Gewissenhaftigkeitsaspekte allein ($R^2 = .05$, $F(1, 22) = 0,608$, $p = .55$) und das Modell passte mit $\beta = -4,228$, $t(23) = -1,922$, $p < .10$ signifikant besser auf die Daten als eine getrennte Betrachtung der beiden Aspekte Fleiß und Ordnung.

Für die Kontrollgruppe der Schüler ergab sich ein signifikanter Effekt mit $R^2 = .18$, $F(1, 65) = 4,401$, $p < .05$. In dieser Stichprobe trug der Interaktionsterm Fleiß x Ordnung somit signifikant zu einer Erhöhung der Gesamtvarianzaufklärung der beiden Prädiktoren Fleiß und Ordnung in Bezug auf die Kriteriumsvariable Abschlussnote bei (s. Tabelle 8-16).

Tabelle 8-16: Ergebnisse der Moderationsanalyse für die Gewissenhaftigkeitsaspekte Fleiß und Ordnung zur Vorhersage von Schulleistung in der Kontrollgruppe

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-0,285	.210	-.166	-1,360	-2,150	.913	-1,252	-2,355*
Ordnung	-0,393	.197	-.244	-2,000 [†]	-2,317	.938	-1,439	-2,471*
Fleiß x Ordnung					0,705	.336	1,865	2,096*
R^2		.11				.17		
R^2 adj.		.09				.13		
ΔR^2		.11				.06		
ΔF		4,190*				4,392*		

Anmerkungen: $N_{KG} = 68$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Hypothese 7a postulierte, dass Personen mit hohen Ausprägungen im Fleißaspekt und hohen Ausprägungen im Selbstwirksamkeitskonstrukt tendenziell erfolgreicher in der Ausbildung sind. In der Subgruppe der kaufmännischen Auszubildenden musste diese Hypothese für das Kriterium Vorgesetztenbeurteilung verworfen werden, da die Interaktion Fleiß x Selbstwirksamkeit zu keiner erhöhten Varianzaufklärung beitrug, was den Angaben in Tabelle 8-17 entnommen werden kann. Auch für das Kriterium allgemeine Berufsschulnote erwies sich das Interaktionsmodell mit $R^2 = .06$, $F(1, 23) = 0,060$, $p = .70$ im Vergleich mit einer getrennten Betrachtung der beiden Faktoren Fleiß und Selbstwirksamkeit ($R^2 = .01$, $F(2, 24) = 0,477$,

$p = .94$) als nicht besser. Hypothese 7a musste somit für die kaufmännischen Auszubildenden abgelehnt werden.

Tabelle 8-17: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und das Selbstwirksamkeitskonstrukt zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-0,405	0,588	-0,171	-0,689	-1,977	4,135	-0,836	-0,478
Selbstwirksamkeit	0,254	0,640	0,099	0,397	-1,336	4,190	-0,519	-0,319
Fleiß x Selbstwirksamkeit					0,497	1,294	1,170	0,384
R^2		.02				.02		
R^2 adj.		-.05				-.09		
ΔR^2		.02				.01		
ΔF		0,240				0,148		

Anmerkungen: $N_{kf} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; $^\dagger p < .10$; * $p < .05$; ** $p < .01$.

In der Subgruppe der technischen Auszubildenden ergaben sich vergleichbare Ergebnisse. Für das Kriterium der Vorgesetztenbeurteilung führte die Analyse der Interaktion der Variablen Fleiß und Selbstwirksamkeit zu keiner Verbesserung der Varianzaufklärung, was den Ergebnissen in Tabelle 8-18 entnommen werden kann. Auch für das Kriterium der allgemeinen Berufsschulnote ergab sich kein signifikanter Effekt, der für eine höhere Varianzaufklärung der

Tabelle 8-18: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und das Selbstwirksamkeitskonstrukt zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-0,745	.899	-0,320	-0,829	-2,545	2,326	-1,094	-1,094
Selbstwirksamkeit	-0,047	.832	-0,022	-0,057	-1,731	2,172	-0,803	-0,797
Fleiß x Selbstwirksamkeit					0,590	0,703	1,518	0,840
R^2		.11				.14		
R^2 adj.		.05				.04		
ΔR^2		.12				.02		
ΔF		1,828				0,705		

Anmerkungen: $N_{techn} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; $^\dagger p < .10$; * $p < .05$; ** $p < .01$.

Interaktion Fleiß x Selbstwirksamkeit ($R^2 = .05$, $F(1, 31) = 0,538$, $p = .66$) im Vergleich mit einer getrennten Betrachtung der beiden Variablen ($R^2 = .03$, $F(2, 32) = 0,405$, $p = .67$) sprach. Das Interaktionsmodell erwies sich nicht als signifikanter Prädiktor des Zusammenhangs ($\beta = -1,616$, $t(32) = -0,900$, $p = .38$). In Bezug auf das Kriterium IHK-Zwischenprüfungen ergab sich ein differenziertes Bild. In Bezug auf das Ergebnis der mündlichen Prüfungen zeigte sich eine signifikant erhöhte Varianzaufklärung der Interaktion Fleiß x Selbstwirksamkeit ($R^2 = .30$, $F(1, 22) = 4,147$, $p < .05$) im Vergleich zur getrennten Betrachtung der beiden Prädiktorvariablen ($R^2 = .27$, $F(2, 23) = 3,093$, $p < .05$). Ein signifikanter Moderationseffekt konnte allerdings nicht gefunden werden ($\beta = -2,033$, $t(23) = -0,994$, $p = .33$). Für den schriftlichen Prüfungsteil war das Interaktionsmodell Fleiß x Selbstwirksamkeit ($R^2 = .19$, $F(1, 22) = 1,685$, $p = .20$) in der Lage mehr Varianz aufzuklären als beide Faktoren getrennt ($R^2 = .05$, $F(2, 23) = 0,608$, $p = .55$), allerdings wurde dieser Effekt nicht signifikant. Das Interaktionsmodell erwies sich jedoch der getrennten Betrachtung der Variablen Fleiß und Selbstwirksamkeit gegenüber tendenziell als überlegen ($\beta = -4,228$, $t(23) = -1,922$, $p < .10$). Inse-

samt muss aufgrund der meist nicht signifikanten Ergebnisse Hypothese 7a für die Stichprobe der technischen Auszubildenden jedoch verworfen werden. Eine eindeutige Überlegenheit des Interaktionsmodells Fleiß x Selbstwirksamkeit konnte für diese Substichprobe nicht bestätigt werden.

In der Kontrollgruppe passt das ursprüngliche Modell mit getrennten Faktoren besser als das Modell mit Interaktion Fleiß x Selbstwirksamkeit, was auch in Tabelle 8-19 dargestellt ist. Allerdings klärt das Modell mit getrennten Faktoren ($R^2 = .12$, $F(2, 66) = 4,647$, $p < .05$) etwas weniger Varianz auf als das Interaktionsmodell ($R^2 = .14$, $F(1, 65) = 3,551$, $p < .05$). Hypothese 7a muss für diese Stichprobe somit abgelehnt werden.

Tabelle 8-19: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und das Selbstwirksamkeitskonstrukt zur Vorhersage von Schulleistung in der Kontrollgruppe

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	0,022	.280	0,013	0,078	-0,920	.867	-0,536	-1,061
Selbstwirksamkeit	-0,515	.234	-0,360	-2,204*	-1,424	.826	-0,996	-1,724
Fleiß x Selbstwirksamkeit					0,359	.313	1,103	1,147
R^2			.12				.14	
R^2 adj.			.10				.10	
ΔR^2			.12				.02	
ΔF			4,647*				1,316	

Anmerkungen: $N_{KG} = 68$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; $^\dagger p < .10$; * $p < .05$; ** $p < .01$.

Insgesamt ergaben sich somit bezüglich Hypothese 7a nur wenige signifikante Ergebnisse. Eine Interaktion des Fleißaspektes und des Selbstwirksamkeitskonstruktes führte somit nicht zu einer erhöhten Varianzaufklärung in den vorliegenden Datensätzen. Hypothese 7a muss verworfen werden.

Hypothese 7b nahm an, dass Personen mit hohen Ausprägungen im Fleißaspekt, die gleichzeitig auch die wahrgenommenen eigenen Fähigkeiten höher einschätzen, in der Ausbildung tendenziell erfolgreicher sind.

Für die Substichgruppe der kaufmännischen Auszubildenden ergaben sich für das Kriterium der allgemeinen Berufsschulleistung keine signifikanten Effekte. Weder war das Interaktionsmodell Fleiß x wahrgenommene eigene Fähigkeiten in der Lage, mehr Varianz aufzuklären ($R^2 = .14$, $F(1, 23) = 1,236$, $p = .32$) als eine getrennte Analyse der beiden Faktoren ($R^2 = .11$, $F(2, 24) = 1,502$, $p = .24$), noch erwies sich das Interaktionsmodell als besser geeig-

Tabelle 8-20: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und die wahrgenommenen eigenen Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung in der Subgruppe der kaufmännischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-0,050	0,453	-0,021	-0,110	1,912	2,695	0,809	0,710
wahrg. eig. Fähigkeiten	0,726	0,518	0,268	1,401	5,041	5,864	1,864	0,860
Fleiß x wahrg.eig. Fähigk.					-1,376	1,862	-1,553	-0,739
R^2		.08				.09		
R^2 adj.		.01				-.01		
ΔR^2		.08				.02		
ΔF		1,153				0,546		

Anmerkungen: $N_{kfM} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; wahrg. eig. Fähigk.= wahrgenommene eigene Fähigkeiten; $^\dagger p < .10$; * $p < .05$; ** $p < .01$.

net, das Kriterium vorherzusagen ($\beta = -1,877$, $t(24) = -0,858$, $p = .40$). Vergleichbare Ergebnisse ergaben sich für das Kriterium der Vorgesetztenbeurteilung (s. Tabelle 8-20). Für die kaufmännischen Auszubildenden konnte Hypothese 7b somit nicht bestätigt werden.

In der Stichprobe der technischen Auszubildenden ergab sich für das Kriterium der allgemeinen Berufsschulleistung ein signifikanter Effekt dahingehend, dass die Interaktion Fleiß x wahrgenommene eigene Fähigkeiten tendenziell nicht in der Lage war, mehr Varianz aufzuklären ($R^2 = .22$, $F(1, 24) = 2,279$, $p = .11$) als eine getrennte Analyse der beiden Faktoren ($R^2 = .20$, $F(2, 25) = 3,163$, $p < .10$). Analog hierzu ergab sich keine signifikante Modellpassung für den Interaktionsterm ($\beta = .853$, $t(25) = 0,781$, $p = .44$). Für das Kriterium der Vorgesetztenbeurteilung zeigte sich ein vergleichbarer Effekt (s. Tabelle 8-21).

Tabelle 8-21: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und die wahrgenommenen eigenen Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung in der Subgruppe der technischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-0,369	0,444	-.159	-0,830	-0,468	1,543	-.201	-0,303
wahrg. eig. Fähigkeiten	0,893	0,458	.373	1,952 [†]	0,701	2,908	.292	0,241
Fleiß x wahrg.eig. Fähigk.					0,067	1,006	.071	0,067
R^2		.22				.22		
R^2 adj.		.17				.14		
ΔR^2		.22				.00		
ΔF		3,979*				0,004		

Anmerkungen: $N_{\text{techn}} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; wahrg. eig. Fähigk.= wahrgenommene eigene Fähigkeiten; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Das Interaktionsmodell war nicht in der Lage, mehr Varianz aufzuklären als beide Faktoren allein. Model 1 schien darüber hinaus besser in der Lage zu sein, die Daten zu repräsentieren als das Interaktionsmodell. In Bezug auf die IHK-Zwischenprüfungsergebnisse ergab sich für den schriftlichen Prüfungsteil, dass die Interaktion Fleiß x wahrgenommene eigene Fähigkeiten ($R^2 = .02$, $F(1, 18) = 0,048$, $p = .95$) hinsichtlich der Varianzaufklärung ebenfalls nicht einer einzelnen Analyse der beiden Faktoren ($R^2 = .01$, $F(2, 19) = 0,144$, $p = .93$) gegenüber überlegen war und auch kein besseres Modell darstellte ($\beta = .859$, $t(19) = 0,583$, $p = .57$). Hinsichtlich des praktischen Prüfungsteils klärten die beiden Variablen Fleiß und wahrgenommene eigene Fähigkeiten Varianz auf ($R^2 = .23$, $F(2, 19) = 2,833$, $p < .10$), allerdings war die Interaktion der Variablen nicht in der Lage, mehr Varianz aufzuklären ($R^2 = .23$, $F(1, 18) = 1,810$, $p = .18$). Eine Überlegenheit des Interaktionsmodells konnte nicht nachgewiesen werden ($\beta = .283$, $t(19) = 0,217$, $p = .83$). Hypothese 7b musste somit für die Subgruppe der technischen Auszubildenden verworfen werden. Die wahrgenommenen eigenen Fähigkeiten moderieren nicht den Einfluss der wahrgenommenen eigenen Fähigkeiten hinsichtlich der verschiedenen Leistungskriterien.

In der Kontrollgruppe der Schüler ergab sich kein signifikanter Effekt bezüglich der Vorhersage der Schulabschlussnote. Das Interaktionsmodell war nicht in der Lage, mehr Varianz aufzuklären als beide Variablen Fleiß und wahrgenommene eigene Fähigkeiten gemeinsam. Es zeigte sich eher, dass das Interaktionsmodell Model 1 unterlegen war (vgl. Tabelle 8-22). Hypothese 7b musste somit auch für die Kontrollgruppe verworfen werden.

Insgesamt deuteten die Ergebnisse an, dass wahrgenommene eigene Fähigkeiten nicht als Moderator zwischen dem Gewissenhaftigkeitsaspekt "Fleiß" und den verschiedenen Leistungskriterien fungierte. Dieser Effekt zeigte sich in allen untersuchten Stichproben, weshalb Hypothese 7b verworfen werden musste.

Tabelle 8-22: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Fleiß und die wahrgenommenen eigenen Fähigkeiten zur Vorhersage von Schulleistung in der Kontrollgruppe

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-0,333	.214	-.194	-1,554	-1,192	1,199	-.694	-0,994
wahrg. eig. Fähigkeiten	-0,302	.252	-.150	-1,199	-1,334	1,440	-.660	-0,926
Fleiß x wahrg.eig. Fähigk.					0,378	0,520	.827	0,728
R^2		.08				.09		
R^2 adj.		.05				.04		
ΔR^2		.08				.01		
ΔF		2,828 [†]				0,530		

Anmerkungen: N_{KG} = 68; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; wahrg. eig. Fähigk.= wahrgenommene eigene Fähigkeiten; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Hypothese 7c postulierte, dass Personen mit hohen Ausprägungen im Ordnungsaspekt und hohen Ausprägungen im Selbstwirksamkeitskonstrukt in der Ausbildung erfolgreicher sind.

In der Stichprobe der kaufmännischen Auszubildenden zeigte sich, dass in Bezug auf das Kriterium der Vorgesetztenbeurteilung die Interaktion Ordnung x Selbstwirksamkeit zu keiner Erhöhung der Varianzaufklärung beitrug (s. Tabelle 8-23). Für das Kriterium der allgemeinen

Tabelle 8-23: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und das Selbstwirksamkeitskonstrukt zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	0,535	0,438	0,241	1,221	4,524	4,116	2,040	1,099
Selbstwirksamkeit	-0,261	0,508	-0,101	-0,514	3,936	4,336	1,530	0,908
Ord. x Selbstwirksamkeit					-1,259	1,292	-2,839	-0,975
R^2		.05				.08		
R^2 adj.		-.02				-.02		
ΔR^2		.05				.03		
ΔF		0,748				0,950		

Anmerkungen: N_{kf} = 30; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; Ord.= Ordnung; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Berufsschulleistung zeigte sich ebenfalls kein signifikanter Effekt. Mit $R^2 = .09$, $F(1, 23) = 1,223$, $p = .52$ trug die Interaktion Ordnung x Selbstwirksamkeit zu keiner erhöhten Varianzaufklärung bei als beide Variablen allein ($R^2 = .09$, $F(2, 24) = 0,784$, $p = .31$). Das Interaktionsmodell erwies sich nicht als besser gegenüber dem Alternativmodell 1 ($\beta = -.263$, $t(24) = -0,081$, $p = .94$). Hypothese 7c musste somit für die kaufmännischen Auszubildenden zurückgewiesen werden. Selbstwirksamkeit moderiert nicht den Zusammenhang zwischen Ordnung und den Leistungskriterien.

Für die Subgruppe der technischen Auszubildenden ergab sich hinsichtlich des Kriteriums der Vorgesetztenbeurteilung kein signifikanter Effekt, der eine Überlegenheit des Interaktionsmodells andeutete (s. Tabelle 8-24). Hinsichtlich des Kriteriums der allgemeinen Berufsschulnote deuteten die Ergebnisse mit $R^2 = .10$, $F(1, 31) = 1,112$, $p = .36$ und $\beta = .042$, $t(32) = 0,023$, $p = .98$ ebenfalls nicht auf eine Überlegenheit des Interaktionsmodells hin. Für die IHK-Zwischenprüfungsergebnisse ergaben sich mit $R^2 = .04$, $F(1, 22) = 0,288$, $p = .83$ und $\beta = -1,792$, $t(23) = -0,811$, $p = .43$ für die schriftlichen Prüfungsergebnisse sowie $R^2 = .06$, $F(1, 22) = 0,438$, $p = .73$ und $\beta = -2,362$, $t(23) = -1,079$, $p = .29$ für die praktischen Prüfungsergeb-

nisse vergleichbare Resultate. Auch hier erwies sich die Interaktion Ordnung x Selbstwirksamkeit als nicht besser hinsichtlich der Varianzaufklärung und Modellpassung, weshalb auch für die Stichprobe der technischen Auszubildenden Hypothese 7c verworfen werden musste.

Tabelle 8-24: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und das Selbstwirksamkeitskonstrukt zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	0,394	0,401	0,198	0,982	-2,646	2,362	-1,334	-1,120
Selbstwirksamkeit	-0,867	0,435	-0,402	-1,991 [†]	-3,706	2,218	-1,719	-1,671
Ord. x Selbstwirksamkeit					0,988	0,757	2,467	1,305
R^2		.12				.18		
R^2 adj.		.06				.08		
ΔR^2		.12				.05		
ΔF		1,981				1,704		

Anmerkungen: $N_{techn} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; Ord.= Ordnung; [†] $p < .10$; * $p < .05$; ** $p < .01$.

In der Kontrollgruppe zeigten sich gegensätzliche Effekte dahingehend, dass die Interaktion der Variablen Ordnung und Selbstwirksamkeit signifikant zu einer Erhöhung der Varianzaufklärung beitrug und das Interaktionsmodell die Zusammenhänge der Variablen besser erklären konnte als eine getrennte Betrachtung der beiden Prädiktoren (s. Tabelle 8-25). Hypothese 7c konnte somit für die Kontrollgruppe bestätigt werden.

Tabelle 8-25: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und das Selbstwirksamkeitskonstrukt zur Vorhersage von Schulleistung in der Kontrollgruppe

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	-0,336	.190	-0,209	-1,768 [†]	-1,626	.631	-1,010	-2,575*
Selbstwirksamkeit	-0,411	.169	-0,288	-2,435*	-1,740	.643	-1,217	-2,705*
Ord. x Selbstwirksamkeit					0,502	.235	1,418	2,136*
R^2		.16				.22		
R^2 adj.		.14				.18		
ΔR^2		.16				.06		
ΔF		6,426*				4,564*		

Anmerkungen: $N_{KG} = 68$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; Ord.= Ordnung; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Insgesamt ergab die Überprüfung des moderierenden Effektes des Selbstwirksamkeitskonstrukts in Bezug auf den Gewissenhaftigkeitsaspekt Ordnung in der Feldstichprobe keine signifikanten Ergebnisse bzw. Hinweise auf eine Interaktion der Variablen hinsichtlich der verschiedenen Leistungskriterien. In der Kontrollgruppe konnte ein solcher Effekt jedoch nachgewiesen werden. Hypothese 7c kann somit auf Basis der vorliegenden Ergebnisse derzeit weder eindeutig bestätigt noch verworfen werden.

Hypothese 7d nahm an, dass Personen mit hohen Ausprägungen im Ordnungsaspekt, die ihre eigenen Fähigkeiten als hoch einschätzen, erfolgreicher in der Ausbildung sind.

Für die Subgruppe der kaufmännischen Auszubildenden ergab sich hinsichtlich der Varianzaufklärung in Bezug auf das Kriterium Vorgesetztenbeurteilung keine Steigerung. Auch das Interaktionsmodell Ordnung x wahrgenommene eigene Fähigkeiten erwies sich gegenüber

Model 1 nicht als besser (vgl. Tabelle 8-26). Für das Kriterium der allgemeinen Berufsschulnote zeigte sich mit $R^2 = .17$, $F(1, 23) = 1,580$, $p = .22$ und $\beta = -1,286$, $t(24) = -0,496$, $p = .63$ ein vergleichbares Ergebnis. Für die Subgruppe der kaufmännischen Auszubildenden musste Hypothese 7d somit zurückgewiesen werden. Wahrgenommene eigene Fähigkeiten moderierten nicht den Zusammenhang zwischen Ordnung und Vorgesetztenbeurteilung bzw. allgemeiner Berufsschulleistung.

Tabelle 8-26: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und die wahrgenommenen eigenen Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung in der Subgruppe der kaufmännischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	0,523	0,393	0,236	1,331	1,958	2,599	0,883	0,753
wahrg. eig. Fähigkeiten	0,811	0,479	0,300	1,693	3,986	5,704	1,474	0,699
Ord. x wahrg. eig. Fähigk.					-0,982	1,758	-1,283	-0,559
R^2		.13				.14		
R^2 adj.		.07				.05		
ΔR^2		.13				.01		
ΔF		2,106				0,312		

Anmerkungen: $N_{kfm} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; Ord.= Ordnung; wahrg. eig. Fähigk.= wahrgenommene eigene Fähigkeiten; $^\dagger p < .10$; * $p < .05$; ** $p < .01$.

In der Substichprobe der technischen Auszubildenden zeigte sich hinsichtlich der Vorgesetztenbeurteilung tendenziell eine etwas höhere Varianzaufklärung der Interaktion Ordnung x wahrgenommene eigene Fähigkeiten ($R^2 = .24$, $F(1, 27) = 2,818$, $p < .10$). Allerdings konnte keine Überlegenheit des Interaktionsmodells aufgefunden werden (vgl. Tabelle 8-27). Bezüg-

Tabelle 8-27: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und die wahrgenommenen eigenen Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung in der Subgruppe der technischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	0,378	0,352	.190	1,074	0,913	1,351	.461	0,676
wahrg. eig. Fähigkeiten	1,242	0,425	.518	2,922	2,260	2,512	.942	0,900
Ord. x wahrg. eig. Fähigk.					-0,341	0,829	-.419	-0,411
R^2		.23				.24		
R^2 adj.		.18				.15		
ΔR^2		.23				.01		
ΔF		4,270*				0,169		

Anmerkungen: $N_{techn} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; Ord.= Ordnung; wahrg. eig. Fähigk.= wahrgenommene eigene Fähigkeiten; $^\dagger p < .10$; * $p < .05$; ** $p < .01$.

lich der Kriteriumsvariable allgemeine Berufsschulleistung ergab sich mit $R^2 = .37$, $F(1, 24) = 4,746$, $p < .01$ zwar eine deutliche Varianzsteigerung durch den Interaktionsterm, allerdings konnte mit $\beta = -.463$, $t(25) = -0,492$, $p = .63$ auch für dieses Kriterium nicht die Überlegenheit des Interaktionsmodells belegt werden. Bezüglich der IHK-Zwischenprüfung ergab sich für die schriftliche Prüfung mit $R^2 = .09$, $F(1, 18) = 0,053$, $p = .98$ keine Varianzsteigerung durch den Interaktionsterm Ordnung x wahrgenommene eigene Fähigkeiten, der mit $\beta = .446$, $t(19) = 0,353$, $p = .73$ zudem nicht Model 1 überlegen war. Bezüglich der praktischen Prüfung ergab sich mit $R^2 = .25$, $F(1, 18) = 1,947$, $p = .16$ und $\beta = .486$, $t(19) = 0,440$, $p = .67$ ebenfalls keine signifikante Varianzsteigerung durch die Moderation. Diesbezüglich erwies sich

Modell 1 mit $R^2 = .24$, $F(2, 19) = 2,949$, $p < .10$ als tendenziell besser zur Vorhersage des Leistungskriteriums. Hypothese 7d konnte somit auch für die Substichprobe der technischen Auszubildenden nicht bestätigt werden.

In der Kontrollgruppe ergab sich mit $R^2 = .10$, $F(1, 65) = 2,503$, $p < .10$ eine Tendenz dahingehend, dass das Interaktionsmodell etwas mehr Varianz aufklärte als Model 1 (s. Tabelle 8-28). Allerdings erwies sich das Interaktionsmodell mit $\beta = .290$, $t(66) = 0,275$, $p = .78$ dem Alternativmodell gegenüber als nicht besser. Hypothese 7d musste somit für die Kontrollgruppe ebenfalls verworfen werden.

Tabelle 8-28: Ergebnisse der Moderationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und wahrgenommene eigene Fähigkeiten zur Vorhersage von Schulleistung in der Kontrollgruppe

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	-.409	0,199	-.254	-2,052*	-.683	1,016	-.424	-0,672
wahrg.eig. Fähigkeiten	-.260	0,250	-.129	-1,040	-.630	1,368	-.312	-0,460
Ord. x wahrg. eig. Fähigk.					.129	0,468	.290	0,275
R^2		.10				.10		
$R^2_{adj.}$		.08				.06		
ΔR^2		.10				.00		
ΔF		3,769*				0,076		

Anmerkungen: $N_{KG} = 68$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; R^2 = Gesamtvarianzaufklärung; Ord.= Ordnung; wahrg. eig. Fähigk.= wahrgenommene eigene Fähigkeiten; $^\dagger p < .10$; * $p < .05$; ** $p < .01$.

Insgesamt konnte in keiner der untersuchten Stichproben ein moderierender Effekt der wahrgenommenen eigenen Fähigkeiten zwischen dem Ordnungsaspekt und den verschiedenen Leistungsvariablen signifikant nachgewiesen werden. Hypothese 7d musste somit verworfen werden.

Die angenommenen moderierenden Effekte (Hypothesen 7a bis 7d) konnten insgesamt nur wenig bestätigt werden. Die folgenden Mediationsanalysen dienen dazu, den Zusammenhang der verschiedenen Selbstbeschreibungskonstrukte in Bezug auf die verschiedenen Leistungskriterien weiter zu erforschen.

8.3.4 Ergebnisse der Mediationsanalysen zur Modellüberprüfung

Hypothese 8 postulierte, dass die Variablen Selbstwirksamkeit bzw. wahrgenommene eigene Fähigkeiten als Mediatoren den Zusammenhang der Gewissenhaftigkeitsaspekte Fleiß bzw. Ordnung mit den verschiedenen Kriterien vermitteln.

Die Ergebnisse Mediationsanalysen sind aufgrund der geringen Stichprobengrößen mit Vorbehalt zu betrachten.

Hypothese 8a nahm an, dass Selbstwirksamkeit den Zusammenhang zwischen Fleiß und der verschiedenen Leistungskriterien mediiert.

In der Subgruppe der kaufmännischen Auszubildenden korrelierte der Prädiktor Fleiß signifikant mit dem Mediator Selbstwirksamkeit ($r = .66$, $p < .01$). Bezüglich der Kriterien der Vorgesetztenbeurteilung ($r = -.11$, n.s.) als auch der Berufsschulnote ($r = -.03$, n.s.) erwies sich Fleiß jedoch nicht als signifikanter Prädiktor. Dennoch wurden die Mediationsanalysen berechnet, um zu überprüfen, inwiefern das Mediationsmodell zu einer höheren Varianzaufklärung führt als das Alternativmodell.

Mit $\beta = -.106$, $t(28) = -0,576$, $p = .57$ erwies sich Fleiß allein nicht als signifikanter Prädiktor für die Vorgesetztenbeurteilung (s. Tabelle 8-29). Allerdings wurde auch das Mediationsmodell (Model 2) mit $\beta = .099$, $t(28) = 0,397$, $p = .69$ nicht signifikant. Entsprechend deutete auch das

Tabelle 8-29: Ergebnisse der multiplen Regression für Fleiß und Selbstwirksamkeit zur Vorhersage der Vorgesetztenbeurteilung bei den kaufmännischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-.252	.437	-.106	-0,576	-.405	.588	-.171	-0,689
Selbstwirksamkeit					.254	.640	.099	0,397

Anmerkungen: $N_{kf\ddot{m}} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; $^{\dagger} p < .10$; * $p < .05$; ** $p < .01$.

Ergebnis des Sobel-Tests mit $z = -0,33$ an, dass der indirekte Vorhersagepfad nicht signifikant wurde. Für das Kriterium der allgemeinen Berufsschulnote ergab sich mit $\beta = -.025$, $t(24) = -0,124$, $p = .90$ ebenfalls keine signifikante Vorhersagekraft für den Prädiktor Fleiß allein. Das Mediationsmodell zeigte mit $\beta = -.089$, $t(24) = -0,323$, $p = .75$ ebenfalls keine gute Modellpassung, was auch durch den Sobel-Test mit $z = 0,12$ bestätigt wurde. Für die Stichprobe der kaufmännischen Auszubildenden konnte ein medierender Effekt des Selbstwirksamkeitskonstrukts in Bezug auf den Prädiktor Fleiß somit nicht aufgefunden werden. Hypothese 8a konnte in dieser Substichprobe daher nicht bestätigt werden.

In der Subgruppe der technischen Auszubildenden korrelierte die Prädiktorvariable Fleiß signifikant mit der medierenden Variable Selbstwirksamkeit ($r = .87$, $p < .01$). Dieser sehr hohe Zusammenhang könnte auch ein Hinweis auf Multikollinearität sein. Mit den Kriterien der Vorgesetztenbeurteilung ($r = -.34$, n.s.), allgemeinen Berufsschulnoten ($r = -.12$, n.s.) sowie den IHK-Zwischenprüfungsnoten sowohl schriftlich ($r = .16$, n.s.) als auch praktisch ($r = .27$, n.s.) ergaben sich keine signifikanten Zusammenhänge. Aufgrund der tendenziell mäßigen bis mittleren Zusammenhänge wurden die Mediationsanalysen dennoch durchgeführt. Bezüglich der Vorgesetztenbeurteilung erwies sich Fleiß allein in der Regression entgegen dem Ergebnis der Korrelationsanalyse mit $\beta = -.340$, $t(28) = -1,945$, $p < .10$ tendenziell als signifikanter Prädiktor des Kriteriums (s. Tabelle 8-30). Demgegenüber zeigte das Mediationsmodell mit

Tabelle 8-30: Ergebnisse der multiplen Regression für Fleiß und Selbstwirksamkeit zur Vorhersage der Vorgesetztenbeurteilung bei den technischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-.790	.406	-.340	-1,945 [†]	-.745	.899	-.320	-0,829
Selbstwirksamkeit					-.047	.832	-.022	-0,057

Anmerkungen: $N_{techn} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; $^{\dagger} p < .10$; * $p < .05$; ** $p < .01$.

$\beta = -.022$, $t(28) = -0,057$, $p = .96$ Model 1 gegenüber nicht als signifikant besser. Dies wurde auch durch das Ergebnis des Sobel-Tests mit $z = 0,07$ bestätigt. Dieser deutete ebenfalls nicht auf eine signifikante Mediation hin. Für das Kriterium der allgemeinen Berufsschulnote zeigte Fleiß allein mit $\beta = -.124$, $t(32) = -0,720$, $p = .48$ keine signifikante Modellpassung. Das Mediationsmodell wies mit $\beta = -.197$, $t(32) = -0,549$, $p = .59$ keine bessere Modellpassung auf als das Alternativmodell. Auch das Ergebnis des Sobel-Tests ergab mit $z = 0,48$ keinen Hinweis auf eine signifikante Mediation. In Bezug auf die praktischen IHK-Zwischenprüfungsergebnisse wies Fleiß allein mit $\beta = .266$, $t(23) = 1,354$, $p = .19$ keine Modellpassung auf. Demgegenüber zeigte das Mediationsmodell mit $\beta = -.977$, $t(23) = -2,465$, $p < .05$ eine signifikante Modellpassung. Da die Regressionsgewichte ungleich null waren, deutete dies auf eine partielle Me-

diation des Selbstwirksamkeitskonstrukts hin. Allerdings wurde der Sobel-Test mit $z = -1,19$ nicht signifikant. Für die schriftlichen Zwischenprüfungsergebnisse zeigte Fleiß allein mit $\beta = .163$, $t(23) = 0,810$, $p = .43$ keine gute Modellpassung. Für Model 2 zeigte sich mit $\beta = -.341$, $t(23) = -0,756$, $p = .20$ zwar eine leichte Verbesserung, allerdings ergab sich kein signifikanter Hinweis auf einen mediiierenden Effekt des Selbstwirksamkeitskonstrukts. Auch der Sobel-Test deutete mit $z = -0,55$ nicht auf eine signifikante Mediation hin. Insgesamt konnte in der Subgruppe der technischen Auszubildenden zu Hypothese 8a somit nur für das Kriterium der praktischen Zwischenprüfungsergebnisse ein Hinweis auf eine partielle Mediation des Selbstwirksamkeitskonstrukts gefunden werden. Für alle weiteren Kriterien musste Hypothese 8a verworfen werden.

In der Kontrollgruppe der Schüler korrelierte Fleiß signifikant mit dem Mediator Selbstwirksamkeit ($r = .72$, $p < .01$) sowie dem Kriterium der Anschlussnote ($r = -.24$, $p < .05$). Fleiß allein sagte mit $\beta = .243$, $t(66) = -2,048$, $p < .05$ signifikant die Abschlussnoten der Schüler voraus (vgl. Tabelle 8-31). Bei gemeinsamer Betrachtung der Variablen Fleiß und Selbstwirksamkeit wies das Mediationsmodell mit $\beta = -.360$, $t(66) = -2,204$, $p < .05$ eine gute Passung

Tabelle 8-31: Ergebnisse der multiplen Regression für Fleiß und Selbstwirksamkeit zur Vorhersage von Schulleistung in der Kontrollgruppe

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-.417	.203	-.243	-2,048*	.022	.280	.013	0,078
Selbstwirksamkeit					-.515	.234	-.360	-2,204*

Anmerkungen: $N_{KG} = 68$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; $^{\dagger} p < .10$; * $p < .05$; ** $p < .01$.

auf. Der Sobel-Test deutete mit $z = 1,50$ jedoch nicht auf eine signifikante Mediation hin. Hypothese 8a konnte für die Kontrollgruppe der Schüler tendenziell bestätigt werden. Da das Regressionsgewicht ungleich null war, handelte es sich hierbei um eine partielle Mediation des Selbstwirksamkeitsaspekts.

Insgesamt sprachen nur wenige Ergebnisse für eine partielle Mediation des Selbstwirksamkeitskonstrukts des Zusammenhangs zwischen Fleiß und den verschiedenen Leistungskriterien. Die Sobel-Tests wurden durchweg nicht signifikant. Hypothese 8a konnte somit nur teilweise bestätigt werden.

Hypothese 8b postulierte einen mediiierenden Effekt des Selbstwirksamkeitskonstrukts zwischen dem Ordnungsaspekt und den schulischen Leistungskriterien. Die Annahme war, dass Ordnung sich vermittelt über Selbstwirksamkeit auf Erfolg in der Ausbildung bzw. Schule auswirkt.

In der Subgruppe der kaufmännischen Auszubildenden korrelierte der Ordnungsaspekt signifikant mit dem Selbstwirksamkeitskonstrukt ($r = .36$, $p < .01$). Mit den Kriterien der Vorgesetztenbeurteilung ($r = .20$, n.s.) sowie der allgemeinen Berufsschulleistung ($r = -.30$, n.s.) zeigte der Ordnungsaspekt keine signifikanten Zusammenhänge. Da die Korrelationskoeffizienten allerdings recht hoch waren und als Tendenzen gewertet werden konnten, wurden die Mediationsanalysen dennoch durchgeführt. Die Ergebnisse der Mediationsanalyse für das Kriterium der Vorgesetztenbeurteilung sind in Tabelle 8-32 dargestellt. Es zeigte sich, dass Ordnung allein kein guter Prädiktor der Vorgesetztenbeurteilung war ($\beta = .204$, $t(28) = 1,124$, $p = .27$). Ein mediiender Effekt durch die Hinzunahme von Selbstwirksamkeit konnte aber auch nicht gefunden werden ($\beta = -.101$, $t(28) = -0,514$, $p = .61$). Der Sobel-Test wies mit $z = -0,47$ ebenfalls nicht auf eine signifikante Mediation hin. Bezüglich des Kriteriums der allgemeinen Berufsschulnote ergaben sich vergleichbare Ergebnisse. Hier erwies sich das

Mediationsmodell mit $\beta = .067$, $t(24) = 0,316$, $p = .76$ dem Alternativmodell ($\beta = -.298$, $t(24) = -1,560$, $p = .13$) gegenüber nicht als überlegen. Auch der Sobel-Test bestätigte die Mediationsannahmen nicht ($z = -0,31$). Hypothese 8b konnte für die Stichprobe der kaufmännischen Auszubildenden somit nicht bestätigt werden.

Tabelle 8-32: Ergebnisse der multiplen Regression für Ordnung und Selbstwirksamkeit zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	.453	.403	.204	1,124	.535	.438	.241	1,221
Selbstwirksamkeit					-.261	.508	-.101	-0,514

Anmerkungen: $N_{kfm} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; $^{\dagger} p < .10$; * $p < .05$; ** $p < .01$.

In der Subpopulation der technischen Auszubildenden korrelierte das Selbstwirksamkeitskonstrukt signifikant mit dem Ordnungsaspekt ($r = .50$, $p < .01$). Bezüglich der Kriterien Vorgesetztenbeurteilung ($r = .00$, n.s.), allgemeine Berufsschulnote ($r = .13$, n.s.) sowie der Ergebnisse der praktischen ($r = -.04$, n.s.) und schriftlichen IHK-Zwischenprüfungen ($r = .07$, n.s.) ergaben sich keine signifikanten Zusammenhänge. Dennoch wurden die Mediationsanalysen berechnet. Es ergab sich hinsichtlich der Mediation des Selbstwirksamkeitskonstrukts für die Vorgesetztenbeurteilung gegenüber den kaufmännischen Auszubildenden eine Tendenz hin zu einer signifikanten Modellpassung (s. Tabelle 8-33). Mit $\beta = -.402$, $t(28) = -1,991$, $p < .10$

Tabelle 8-33: Ergebnisse der multiplen Regression für Ordnung und Selbstwirksamkeit zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	.008	.368	.004	0,023	.394	.401	.198	0,982
Selbstwirksamkeit					-.867	.435	-.402	1,991 [†]

Anmerkungen: $N_{techn} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; $^{\dagger} p < .10$; * $p < .05$; ** $p < .01$.

wies das Mediationsmodell tendenziell einen guten Model-Fit auf. Ordnung allein erwies sich nicht als guter Prädiktor ($\beta = .004$, $t(28) = -0,023$, $p = .89$). Der Sobel-Test auf Mediation wurde mit $z = -0,02$ allerdings nicht signifikant. Für das Kriterium der allgemeinen Berufsschulnote ergab sich für die technischen Auszubildenden mit $\beta = -.354$, $t(32) = -1,699$, $p < .10$ eine leichte Überlegenheit des Mediationsmodells gegenüber dem Alternativmodell ($\beta = .125$, $t(32) = 0,725$, $p = .47$). Der Sobel-Test wurde allerdings nicht signifikant ($z = -0,67$). Bezüglich der IHK-Zwischenprüfung ergaben sich für den praktischen Prüfungsteil keine signifikanten Ergebnisse die Modellpassung betreffend. Weder das Mediationsmodell ($\beta = .108$, $t(23) = 0,408$, $p = .69$) noch das Alternativmodell ($\beta = -.043$, $t(23) = -0,213$, $p = .83$) wiesen eine gute Modellpassung auf. Der Sobel-Test wurde ebenfalls nicht signifikant ($z = -0,19$). Diese Tendenz spiegelte sich auch im schriftlichen Prüfungsteil wider. Diesbezüglich konnte wiederum keine Überlegenheit des Mediationsmodells ($\beta = .059$, $t(23) = 0,221$, $p = .83$) gegenüber dem Alternativmodell ($\beta = .065$, $t(23) = 0,319$, $p = .75$) gefunden werden. Der Sobel-Test bestätigte diese Ergebnisse ($z = 0,18$). Insgesamt ergaben sich in der Stichprobe der technischen Auszubildenden keine signifikanten Hinweise, die für einen mediiierenden Einfluss des Selbstwirksamkeitsaspekts zwischen Ordnung und den verschiedenen Leistungskriterien sprachen. Allerdings ergaben sich einige Hinweise auf eine leichte Überlegenheit des Mediationsmodells für zwei abhängige Variablen. Hypothese 8b konnte in dieser Subpopulation somit zwar nicht bestätigt werden, sollte aufgrund der Tendenzen hin zu einer Mediation aber auch nicht verworfen werden.

In der Kontrollgruppe zeigten das Selbstwirksamkeitskonstrukt und der Ordnungsaspekt einen signifikanten Zusammenhang ($r = .36$, $p < .01$). Auch mit dem Kriterium der Abschlussnote korrelierte der Ordnungsaspekt signifikant ($r = -.30$, $p < .05$). Die Ergebnisse der Mediationsanalyse sind in Tabelle 8-34 dargestellt.

Tabelle 8-34: Ergebnisse der multiplen Regression für Ordnung und Selbstwirksamkeit zur Vorhersage der Abschlussnote in der Kontrollgruppe

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	-.447	.188	-.296	-2,539*	-.336	.190	-.209	-1,768 [†]
Selbstwirksamkeit					-.411	.169	-.288	-2,435*

Anmerkungen: $N_{KG} = 68$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Mit $\beta = -.288$, $t(66) = -2,435$, $p < .05$ war das Mediationsmodell signifikant besser in der Lage, den Zusammenhang der Variablen zu erklären als Model 1. Der Sobel-Test wurde mit $z = 1,76$ knapp nicht signifikant. In der Kontrollgruppe konnte Hypothese 8b somit dahingehend bestätigt werden, dass es durch das Selbstwirksamkeitskonstrukt zu einer partiellen Mediation kam.

Insgesamt ergaben sich widersprüchliche Ergebnisse hinsichtlich Hypothese 8b. In den Feldstichproben ergaben sich kaum signifikante Ergebnisse, die für einen vermittelnden Effekt des Selbstwirksamkeitskonstrukts zwischen dem Ordnungsaspekt und den verschiedenen Leistungskriterien sprachen. In der Kontrollgruppe hingegen erwies sich das Mediationsmodell als signifikant besser. Hypothese 8b konnte somit nur teilweise bestätigt werden.

Hypothese 8c nahm einen mediierenden Effekt der wahrgenommenen eigenen Fähigkeiten zwischen dem Fleißaspekt und den Leistungskriterien an.

Die wahrgenommenen eigenen Fähigkeiten und der Fleißaspekt zeigten in der Subgruppe der kaufmännischen Auszubildenden keinen signifikanten Zusammenhang ($r = -.32$, n.s.). Allerdings deutete dies auf eine Tendenz hin. Bezüglich der Kriterien der Vorgesetztenbeurteilungen ($r = -.11$, n.s.) und allgemeinen Berufsschulnote ($r = -.03$, n.s.) ergaben sich ebenfalls keine signifikanten Zusammenhänge. Die Mediationsanalysen wurden dennoch berechnet. Die Ergebnisse für das Kriterium der Vorgesetztenbeurteilung sind in Tabelle 8-35 dargestellt. Für die Vorgesetztenbeurteilung zeigte sich kein mediierender Effekt ($\beta = .268$,

Tabelle 8-35: Ergebnisse der multiplen Regression für Fleiß und wahrgenommene eigene Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-.252	.437	-.106	-0,576	-.050	.453	-.021	-0,110
wahrgen. eig. Fähigkeiten					.726	.518	.268	1,401

Anmerkungen: $N_{kf} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; wahrgen. eig. Fähigkeiten= wahrgenommene eigene Fähigkeiten; [†] $p < .10$; * $p < .05$; ** $p < .01$.

$t(28) = 1,401$, $p = .17$). Der Sobel-Test wurde ebenfalls nicht signifikant ($z = -0,53$). Allerdings erwies sich das Alternativmodell tendenziell als noch weniger passend ($\beta = -.106$, $t(28) = -0,576$, $p = .57$). Für das Kriterium der allgemeinen Berufsschulleistung ergaben sich vergleichbare Resultate. Hier zeichnete sich tendenziell eine Modellpassung des Mediationsmodells ($\beta = .353$, $t(24) = 1,729$, $p < .10$) im Vergleich zum Alternativmodell ($\beta = -.025$, $t(24) = -0,124$, $p = .90$) ab. Der Sobel-Test konnte jedoch keinen Mediationseffekt bestätigen

($z = -0,12$). Insgesamt waren die Ergebnisse in der Subgruppe der kaufmännischen Auszubildenden nicht signifikant, deuteten tendenziell jedoch auf eine partielle Mediation der wahrgenommenen eigenen Fähigkeiten hin. Hypothese 8c konnte für diese Subgruppe somit weder bestätigt noch verworfen werden.

In der Subpopulation der technischen Auszubildenden korrelierte der Fleißaspekt signifikant mit den wahrgenommenen eigenen Fähigkeiten ($r = -.46$, $p < .1$). Wie bereits oben beschrieben wurde, zeigten die Kriterien Vorgesetztenbeurteilung ($r = -.34$, n.s.), allgemeine Berufsschulnoten ($r = -.12$, n.s.) sowie die schriftlichen ($r = .16$, n.s.) und praktischen IHK-Zwischenprüfungsnoten ($r = .27$, n.s.) keine signifikanten Zusammenhänge mit dem Prädiktor Fleiß. Aufgrund der tendenziell mäßigen bis mittleren Zusammenhänge wurden die Mediationsanalysen trotzdem berechnet. Für das Kriterium der Vorgesetztenbeurteilung sind die Ergebnisse der Mediationsanalyse in Tabelle 8-36 dargestellt. Es zeigte sich, dass die Hinzunahme der wahrgenommenen eigenen Fähigkeiten als Mediatorvariable die Zusammenhänge

Tabelle 8-36: Ergebnisse der multiplen Regression für Fleiß und wahrgenommene eigene Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-.790	.406	-.340	-1,945 [†]	-.369	.444	-.159	-0,830
wahrgen. eig. Fähigkeiten					.893	.458	.373	1,952 [†]

Anmerkungen: $N_{\text{techn}} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; wahrgen. eig. Fähigkeiten= wahrgenommene eigene Fähigkeiten; [†] $p < .10$; * $p < .05$; ** $p < .01$.

tendenziell erklären konnte ($\beta = .373$, $t(28) = 1,952$, $p < .10$). Der Sobel-Test deutete jedoch nicht auf eine signifikante Mediation hin ($z = -1,38$). Für das Kriterium der allgemeinen Berufsschulleistung zeigte sich mit $\beta = .520$, $t(25) = 2,515$, $p < .05$ ein deutlicher Effekt, der für eine Mediation der Variable im Vergleich mit Model 1 ($\beta = -.007$, $t(25) = -0,037$, $p = .97$) sprach. Dieses Ergebnis wurde allerdings nicht durch den Sobel-Test gestützt ($z = -0,04$). Für die Ergebnisse der IHK-Zwischenprüfung zeigte sich für den schriftlichen Prüfungsteil mit $\beta = -.011$, $t(19) = -0,043$, $p = .97$ keine Überlegenheit des Mediationsmodells gegenüber dem Alternativmodell ($\beta = .070$, $t(19) = 0,314$, $p = .76$). Auch der Sobel-Test deutete mit $z = -0,04$ nicht auf eine signifikante Mediation hin. Für den praktischen Prüfungsteil wies das Mediationsmodell mit $\beta = -.389$, $t(19) = -1,704$, $p = .11$ ebenfalls keinen signifikanten Model-Fit auf. Auch das Alternativmodell ($\beta = .335$, $t(19) = 1,588$, $p = .13$) wurde nicht signifikant. Der Sobel-Test bestätigte diese Ergebnisse ($z = -1,16$). Zusammenfassend kann für die Subpopulation der technischen Auszubildenden teilweise von einer partiellen Mediation der wahrgenommenen eigenen Fähigkeiten ausgegangen werden. Hypothese 8c kann in dieser Stichprobe als teilweise bestätigt angesehen werden.

In der Kontrollgruppe korrelierten die wahrgenommenen eigenen Fähigkeiten signifikant mit dem Fleißaspekt ($r = .30$, $p < .01$), der auch mit dem Kriterium der Abschlussnote signifikant zusammenhing ($r = -.24$, $p < .05$). Hier erwies sich das Mediationsmodell ($\beta = -.150$, $t(66) = -1,199$, $p = .24$) dem Alternativmodell ($\beta = -.243$, $t(66) = -2,048$, $p < .05$) gegenüber als unterlegen. Der Sobel-Test bestätigte dieses Ergebnis ($z = 0,04$). Hypothese 8c kann für diese Stichprobe somit nicht als bestätigt angesehen werden.

Insgesamt lieferten die Ergebnisse zur Überprüfung von Hypothese 8c keine eindeutigen Befunde. Einige Ergebnisse deuteten auf eine partielle Mediation der wahrgenommenen eigenen Fähigkeiten zwischen dem Fleißaspekt und den verschiedenen Leistungskriterien hin, andere Ergebnisse ergaben keine signifikanten Mediationen. Hypothese 8c kann somit weder eindeutig bestätigt noch verworfen werden.

Hypothese 8d postulierte einen mediiierenden Effekt der wahrgenommenen eigenen Fähigkeiten in dem Sinne, dass sich der Ordnungsaspekt vermittelt über die eigenen Fähigkeiten auf den Schul- bzw. Ausbildungserfolg der Schüler bzw. Auszubildenden auswirkt.

In der Subpopulation der kaufmännischen Auszubildenden korrelierten die wahrgenommenen eigenen Fähigkeiten nicht signifikant mit dem Ordnungsaspekt ($r = -.11$, n.s.). Auch die Kriterien Vorgesetztenbeurteilung ($r = -.30$, n.s.) und allgemeine Berufsschulleistung ($r = .20$, n.s.) zeigten mit dem Ordnungsaspekt keine signifikanten Zusammenhänge, aber Tendenzen. Daher wurden die Mediationsanalysen trotzdem durchgeführt. Tabelle 8-37 gibt einen Überblick

Tabelle 8-37: Ergebnisse der multiplen Regression für Ordnung und wahrgenommene eigene Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	.453	.403	.204	1,124	.523	.393	.236	1,331
wahrgen. eig. Fähigkeiten					.811	.479	.300	1,693

Anmerkungen: $N_{kfm} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; wahrgen. eig. Fähigkeiten= wahrgenommene eigene Fähigkeiten; $^{\dagger} p < .10$; * $p < .05$; ** $p < .01$.

über die Ergebnisse für das Kriterium der Vorgesetztenbeurteilung. Das Mediationsmodell ($\beta = .300$, $t(28) = 1,693$, $p = .10$) wies gegenüber Model 1 ($\beta = .204$, $t(28) = 1,124$, $p = .27$) eine etwas bessere Modellpassung auf, allerdings wurde dieser Effekt nicht signifikant ($z = 0,94$). Bezüglich der allgemeinen Berufsschulnote ergab sich mit $\beta = .276$, $t(24) = 1,449$, $p = .16$ kein signifikanter Model-Fit, was durch den Sobel-Test gestützt wurde ($z = -1,06$). Auch das Alternativmodell ($\beta = -.298$, $t(24) = -1,560$, $p = .13$) wurde allerdings nicht signifikant. Trotz einiger Tendenzen hinsichtlich der Überlegenheit des Mediationsmodells musste Hypothese 8d somit für die Subpopulation der kaufmännischen Auszubildenden verworfen werden.

In der Teilstichprobe der technischen Auszubildenden zeigte sich ebenfalls kein signifikanter Zusammenhang zwischen dem Ordnungsaspekt und den wahrgenommenen eigenen Fähigkeiten ($r = -.35$, n.s.). Auch bezüglich der Kriterien Vorgesetztenbeurteilung ($r = .00$, n.s.), allgemeine Berufsschulnote ($r = .13$, n.s.) sowie der Ergebnisse der praktischen ($r = -.04$, n.s.) und schriftlichen IHK-Zwischenprüfungen ($r = .07$, n.s.) ergaben sich keine signifikanten Zusammenhänge. Dennoch wurden die Mediationsanalysen durchgeführt. Tabelle 8-38 gibt einen Überblick über die Ergebnisse für das Kriterium der Vorgesetztenbeurteilung. Die Ergebnisse sprachen für eine signifikante Überlegenheit des Mediationsmodells ($\beta = .518$, $t(28) = 2,922$, $p < .01$) gegenüber Model 1 ($\beta = .004$, $t(28) = 0,023$, $p = .98$). Allerdings bestätigte der Sobel-Test nicht den Mediationseffekt ($z = 0,02$). Für das Kriterium der allgemeinen

Tabelle 8-38: Ergebnisse der Regressionsanalysen für Ordnung und wahrgenommene eigene Fähigkeiten zur Vorhersage der Vorgesetztenbeurteilung für die technischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	0,008	.368	.004	0,023	0,378	.352	.190	1,074
wahrgen. eig. Fähigkeiten					1,242	.452	.518	2,922**

Anmerkungen: $N_{techn} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; wahrgen. eig. Fähigkeiten= wahrgenommene eigene Fähigkeiten; $^{\dagger} p < .10$; * $p < .05$; ** $p < .01$.

allgemeinen Berufsschulnoten ergab sich mit $\beta = .584$, $t(25) = 3,386$, $p < .01$ wiederum ein besserer Model-Fit für das Mediationsmodell gegenüber dem Alternativmodell ($\beta = .274$, $t(25) = 1,454$, $p = .16$). Der Sobel-Test konnte dieses Ergebnis mit $z = 1,33$ allerdings nicht

bestätigen. Mediationseffekte konnten auch teilweise für die Ergebnisse der IHK-Zwischenprüfungen gefunden werden. Hinsichtlich des praktischen Prüfungsteils wies das Mediationsmodell einen signifikant besseren Model-Fit auf ($\beta = -.515$, $t(19) = -2,428$, $p < .05$) als das Alternativmodell ($\beta = .001$, $t(19) = 0,002$, $p = .99$). Der Sobel-Test deutete allerdings nicht auf eine signifikante Mediation hin ($z = 0,00$). Für den schriftlichen Prüfungsteil konnte dieser Effekt für das Mediationsmodell demgegenüber nicht bestätigt werden ($\beta = -.046$, $t(19) = -0,188$, $p = .85$). Auch das Alternativmodell wies keine gute Modellpassung auf ($\beta = .003$, $t(19) = 0,013$, $p = .99$). Der zugehörige Sobel-Test wurde ebenfalls nicht signifikant ($z = -0,01$). Mit Ausnahme des schriftlichen Prüfungsteils erwiesen sich die wahrgenommenen eigenen Fähigkeiten somit bei den technischen Auszubildenden als partieller Mediator des Zusammenhangs zwischen Ordnung und den Leistungskriterien, weshalb Hypothese 8d teilweise bestätigt werden konnte.

In der Kontrollgruppe der Schüler korrelierten die wahrgenommenen eigenen Fähigkeiten und der Ordnungsaspekt signifikant miteinander ($r = .32$, $p < .01$). Auch mit dem Kriterium zeigte der Ordnungsaspekt einen signifikanten Zusammenhang ($r = -.30$, $p < .05$). Es zeigte sich allerdings keine Überlegenheit des Mediationsmodells ($\beta = -.129$, $t(66) = -1,040$, $p = .30$) gegenüber dem Alternativmodell ($\beta = -.296$, $t(66) = -2,539$, $p < .05$). Dies wird auch durch den nicht-signifikanten Sobel-Test gestützt ($z = 0,96$). Die Ergebnisse sind auch in Tabelle 8-39 dargestellt. Hypothese 8d konnte somit für die Kontrollgruppe nicht bestätigt werden.

Tabelle 8-39: Ergebnisse der Mediationsanalyse für den Gewissenhaftigkeitsaspekt Ordnung und wahrgenommene eigene Fähigkeiten zur Vorhersage von Schulleistung in der Kontrollgruppe

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Ordnung	-.477	.188	-.296	-2,539*	-.409	.199	-.254	-2,052*
wahrg.eig. Fähigkeiten					-.260	.250	-.129	-1,040

Anmerkungen: $N_{KG} = 68$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; wahrgen. eig. Fähigkeiten= wahrgenommene eigene Fähigkeiten; $^{\dagger} p < .10$; * $p < .05$; ** $p < .01$.

Insgesamt ergaben sich widersprüchliche Ergebnisse hinsichtlich des mediierenden Effektes der wahrgenommenen eigenen Fähigkeiten. Je nach untersuchter Stichprobe sprachen die Daten für bzw. gegen einen Effekt, weshalb Hypothese 8d weder eindeutig bestätigt noch verworfen werden konnte.

Im Anhang A6 sind zusätzlich Mediationsanalysen enthalten, die den Einfluss des Ordnungsaspektes als mediierender Variable des Zusammenhangs zwischen Fleiß und den verschiedenen Kriterien in den drei Stichproben untersuchen. Die Ergebnisse sind uneindeutig. Abhängig von der jeweils untersuchten Stichprobe und dem Kriterium ist es möglich, dass Ordnung als Mediator des Zusammenhanges fungiert.

8.4 Inkrementelle Validität

Der folgende Abschnitt enthält die Ergebnisse der Überprüfung der inkrementellen Validität der AGS. Zum einen soll die Vorhersagekraft der Gewissenhaftigkeitsaspekte Fleiß und Ordnung im Vergleich mit der Vorhersagekraft der Gewissenhaftigkeitsskala des NEO-FFI (Borkenau & Ostendorf, 1993, 2008) verglichen werden (Abschnitt 8.4.1). Zum anderen soll die inkrementelle Validität der AGS bzw. der Gewissenhaftigkeitsaspekte über die Gewissenhaftigkeitsskala des NEO-FFI hinaus überprüft werden (Abschnitt 8.4.2). Zuletzt sollen die inter- und intrapersonellen Antwortvariabilitäten der AGS und der Gewissenhaftigkeitsskala des NEO-FFI miteinander verglichen werden (Abschnitt 8.4.3). Da in der Kontrollgruppe der

NEO-FFI nicht miterhoben werden konnte, beziehen sich die Ergebnisse der folgenden Abschnitte lediglich auf die Retest-Stichprobe der Auszubildenden.

8.4.1 Vergleich der Vorhersagekraft der AGS mit dem NEO-FFI

Hypothese 9 nahm sich des Vergleichs der verschiedenen Gewissenhaftigkeitsmaße an. Entsprechend postulierte Hypothese 9a, dass Fleiß Ausbildungserfolg präziser vorhersagt als Gewissenhaftigkeit allgemein, während Hypothese 9b dies entsprechend für den Ordnungsaspekt annahm. Die Ergebnisse zu diesen Hypothesen sind in Tabelle 8-40 dargestellt.

Tabelle 8-40: Korrelationen der AGS bzw. Gewissenhaftigkeitsskala (NEO-FFI) mit den Leistungskriterien in der Auszubildendenstichprobe

Kriterium	kaufmännische Auszubildende		technische Auszubildende			
	VG ^a	BS ^b	VG ^c	BS ^d	IHK mdl. ^e	IHK schriftl. ^e
GEW T1	-.31 [†]	-.36 [†]	-.42*	-.01	.23	.21
Fleiß T1	-.11	-.03	-.34	-.12	.27	.16
Fleiß T2	.05	.21	-.33 [†]	-.22	.26	.09
Ordnung T1	.20	-.30	.00	.13	-.04	.07
Ordnung T2	.06	-.13	.02	-.06	.18	-.23

Anmerkungen: ^a $N_{kfm} = 27$; ^b $N_{kfm} = 31$; ^c $N_{techn} = 31$; ^d $N_{techn} = 35$; ^e $N_{techn} = 26$; VG= Vorgesetztenbeurteilung; BS= Berufsschulnote; IHK mdl.= Ergebnis mündliche IHK-Zwischenprüfung; IHK schriftl.= Ergebnis schriftliche IHK-Zwischenprüfung; GEW= Gewissenhaftigkeit (NEO-FFI).
[†] $p < .10$; * $p < .05$; ** $p < .01$.

Entgegen der Annahme aus Hypothese 9a korrelierte der Fleißaspekt weder in der kaufmännischen noch in der technischen Subpopulation zu T1 oder T2 signifikant mit den verschiedenen Ausbildungsleistungskriterien. Lediglich die Korrelation der Vorgesetztenbeurteilung in der Stichprobe der technischen Auszubildenden zeigte tendenziell einen Zusammenhang mit dem zu T2 erhobenen Fleißaspekt ($r = -.33$, $p < .10$). Hypothese 9a musste somit verworfen werden.

Der Ordnungsaspekt zeigte in keiner Stichprobe mit einem der erhobenen Leistungskriterien einen signifikanten Zusammenhang. Hypothese 9b musste somit ebenfalls verworfen werden.

Demgegenüber korrelierte allgemeine Gewissenhaftigkeit signifikant mit der Vorgesetztenbeurteilung für die technischen Auszubildenden ($r = -.42$, $p < .05$). Mit den weiteren Leistungskriterien in der technischen Teilstichprobe ergaben sich allerdings auch für die allgemeine Gewissenhaftigkeitsskala keine signifikanten Zusammenhänge. In der kaufmännischen Subpopulation zeigte die allgemeine Gewissenhaftigkeitsskala wiederum tendenziell Zusammenhänge mit der Vorgesetztenbeurteilung ($r = -.31$, $p < .10$) und der allgemeinen Berufsschulnote ($r = -.36$, $p < .10$). Da die Zusammenhänge der Gewissenhaftigkeitsaspekte Fleiß und Ordnung gegenüber der allgemeinen Gewissenhaftigkeitsskala nicht größer und signifikant wurden, musste Hypothese 9 verworfen werden. Die Gewissenhaftigkeitsaspekte Fleiß und Ordnung zeigten in den beiden Auszubildendenstichproben keine höhere Vorhersagekraft.

Die Ergebnisse der Überprüfung der Vorhersagekraft der verschiedenen Gewissenhaftigkeitsmaße sind aufgrund der geringen Stichprobengrößen ($N = 26$ bis 35) und der damit einhergehenden geringen Power allerdings mit Vorbehalt zu betrachten.

8.4.2 Inkrementelle Validität der AGS

Hypothese 10 nahm an, dass die Gewissenhaftigkeitsaspekte Fleiß und Ordnung inkrementelle Validität über ein allgemeines Gewissenhaftigkeitsmaß hinaus besitzen. Dies wurde so-

wohl für die beiden Aspekte Fleiß (Hypothese 10a) und Ordnung (Hypothese 10b) allein als auch für eine gemeinsame Analyse der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung (Hypothese 10c) angenommen.

Die Ergebnisse der Überprüfung der inkrementellen Validität der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sind in den folgenden Tabellen 8-41 bis 8-46 dargestellt.

Die Ergebnisse der Überprüfung der inkrementellen Validität sind aufgrund der geringen Stichprobengrößen ($N=25$ bis 34) mit Vorbehalt zu betrachten.

Tabelle 8-41: Regressionsmodelle zur Vorhersage des Kriteriums der Vorgesetztenbeurteilung für die kaufmännischen Auszubildenden

	R^2	$R^2 \text{ adj.}$	ΔR^2	F	B	SE	β
Modell 1	.09	.06	.09	2,958 [†]			
GEW					-0,894	.520	-.304 [†]
Modell 2a	.10	.03	.01	1,518			
GEW					-1,034	.631	-.352
Fleiß					0,206	.508	.087
Modell 2b	.33	.28	.24	6,953**			
GEW					-1,990	.571	-.677**
Ordnung					1,365	.431	.615**
Modell 3	.33	.26	.24	4,504*			
GEW					-2,063	.645	-.702*
Fleiß					0,116	.446	.049
Ordnung					1,357	-.439	.612*
Modell 4	.08	.02	.08	1,232			
Fleiß					-0,505	.462	-.213
Ordnung					0,631	.434	.258

Anmerkungen: $N=30$; GEW= allgemeine Gewissenhaftigkeit (NEO-FFI); R^2 = Determinationskoeffizient; $R^2 \text{ adj.}$ = korrigierter Determinationskoeffizient; ΔR^2 = inkrementelle Validität; B = unstandardisierter Koeffizient; SE = Standardfehler, β = standardisierter Koeffizient; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Tabelle 8-42: Regressionsmodelle zur Vorhersage des Kriteriums der allgemeinen Berufsschulleistung für die kaufmännischen Auszubildenden

	R^2	$R^2 \text{ adj.}$	ΔR^2	F	B	SE	β
Modell 1	.12	.09	.12	3,467 [†]			
GEW					-0,535	.287	-.349 [†]
Modell 2a	.16	.09	.04	2,347			
GEW					-0,746	.345	-.486*
Fleiß					0,297	.271	.246
Modell 2b	.13	.06	.01	1,841			
GEW					-0,409	.369	-.267
Ordnung					-0,179	.321	-.134
Modell 3	.19	.09	.03	1,800			
GEW					-0,589	.391	-.384
Fleiß					0,358	.281	.297
Ordnung					-.284	.328	-.213
Modell 4	.11	.04	.11	1,486			
Fleiß					0,250	.269	.170
Ordnung					-0,513	.298	-.384 [†]

Anmerkungen: $N=26$; GEW= allgemeine Gewissenhaftigkeit (NEO-FFI); R^2 = Determinationskoeffizient; $R^2 \text{ adj.}$ = korrigierter Determinationskoeffizient; ΔR^2 = inkrementelle Validität; B = unstandardisierter Koeffizient; SE = Standardfehler, β = standardisierter Koeffizient; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Tabelle 8-43: Regressionsmodelle zur Vorhersage des Kriteriums der Vorgesetztenbeurteilung für die technischen Auszubildenden

	R^2	$R^2 \text{ adj.}$	ΔR^2	F	B	SE	β
Modell 1	.18	.15	.18	6,315*			
GEW					-0,872	.347	-.423*
Modell 2a	.18	.12	.00	3,049 [†]			
GEW					-0,866	.589	-.420
Fleiß					-0,010	.664	-.004
Modell 2b	.35	.31	.17	7,619**			
GEW					-1,707	.437	-.828**
Ordnung					1,152	.421	.581*
Modell 3	.36	.29	.18	4,977**			
GEW					-1,894	.651	-.918**
Fleiß					0,237	.605	.102
Ordnung					1,177	.432	.594*
Modell 4	.15	.09	.15	2,555 [†]			
Fleiß					-1,050	.465	-.451*
Ordnung					0,450	.396	.227

Anmerkungen: $N=30$; GEW= allgemeine Gewissenhaftigkeit (NEO-FFI); R^2 = Determinationskoeffizient; $R^2 \text{ adj.}$ = korrigierter Determinationskoeffizient; ΔR^2 = inkrementelle Validität; B = unstandardisierter Koeffizient; SE = Standardfehler, β = standardisierter Koeffizient; [†] $p < .10$; * $p < .05$; ** $p < .01$.

In der technischen Substichprobe können für die Kriterien der allgemeinen Berufsschulleistung sowie der schriftlichen und praktischen IHK-Zwischenprüfungsergebnisse keine Aussagen gemacht werden, da keine der berechneten Regressionsanalysen signifikante Ergebnisse zeigte (vgl. Tabellen 8-44 bis 8-46).

Tabelle 8-44: Regressionsmodelle zur Vorhersage des Kriteriums der allgemeinen Berufsschulleistung für die technischen Auszubildenden

	R^2	$R^2 \text{ adj.}$	ΔR^2	F	B	SE	β
Modell 1	.00	-.03	.00	0,001			
GEW					-0,008	.210	-.006
Modell 2a	.04	-.02	.04	0,722			
GEW					0,349	.362	.290
Fleiß					-0,474	.395	-.362
Modell 2b	.04	-.02	.04	0,601			
GEW					-0,252	.306	-.210
Ordnung					0,361	.330	.278
Modell 3	.08	-.01	.04	0,910			
GEW					0,105	.420	.088
Fleiß					-0,483	.393	-.369
Ordnung					0,369	.327	.285
Modell 4	.08	.02	.08	1,374			
Fleiß					-0,415	.280	-.317
Ordnung					0,412	.277	.317

Anmerkungen: $N=34$; GEW= allgemeine Gewissenhaftigkeit (NEO-FFI); R^2 = Determinationskoeffizient; $R^2 \text{ adj.}$ = korrigierter Determinationskoeffizient; ΔR^2 = inkrementelle Validität; B = unstandardisierter Koeffizient; SE = Standardfehler, β = standardisierter Koeffizient; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Tabelle 8-45: Regressionsmodelle zur Vorhersage des Kriteriums der praktischen IHK-Zwischenprüfungsergebnisse für die technischen Auszubildenden

	R^2	$R^2 \text{ adj.}$	ΔR^2	F	B	SE	β
Modell 1 GEW	.02	-.03	.02	0,375	3,225	5,265	.124
Modell 2a GEW Fleiß	.12	.04	.10	1,534	-11,338 17,728	10,268 10,854	-.436 .645
Modell 2b GEW Ordnung	.05	-.03	.04	0,653	8,660 -7,787	7,713 8,065	.333 -.286
Modell 3 GEW Fleiß Ordnung	.16	.04	.04	1,382	-5,907 17,950 -8,043	11,519 10,840 7,778	-.227 .653 -.296
Modell 4 Fleiß Ordnung	.15	.08	.15	2,006	13,705 -9,861	6,885 6,810	.499 [†] -.363

Anmerkungen: $N=25$; GEW= allgemeine Gewissenhaftigkeit (NEO-FFI); R^2 = Determinationskoeffizient; $R^2 \text{ adj.}$ = korrigierter Determinationskoeffizient; ΔR^2 = inkrementelle Validität; B = unstandardisierter Koeffizient; SE = Standardfehler, β = standardisierter Koeffizient; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Tabelle 8-46: Regressionsmodelle zur Vorhersage des Kriteriums der schriftlichen IHK-Zwischenprüfungsergebnisse für die technischen Auszubildenden

	R^2	$R^2 \text{ adj.}$	ΔR^2	F	B	SE	β
Modell 1 GEW	.06	.02	.06	1,404	4,558	3,847	.235
Modell 2a GEW Fleiß	.06	-.02	.01	0,762	7,368 -3,421	7,895 8,346	.380 -.167
Modell 2b GEW Ordnung	.08	.00	.02	0,995	7,788 -4,628	5,673 5,932	.402 -.228
Modell 3 GEW Fleiß Ordnung	.09	-.04	.02	0,690	10,461 -3,294 -4,581	8,954 8,427 6,046	.540 -.161 -.226
Modell 4 Fleiß Ordnung	.03	-.06	.03	0,347	4,224 -1,360	5,183 5,423	.770 -.251

Anmerkungen: $N=25$; GEW= allgemeine Gewissenhaftigkeit (NEO-FFI); R^2 = Determinationskoeffizient; $R^2 \text{ adj.}$ = korrigierter Determinationskoeffizient; ΔR^2 = inkrementelle Validität; B = unstandardisierter Koeffizient; SE = Standardfehler, β = standardisierter Koeffizient; [†] $p < .10$; * $p < .05$; ** $p < .01$.

Hypothese 10a postulierte, dass Fleiß gegenüber einem allgemeinen Gewissenhaftigkeitsmaß bei der Vorhersage von Ausbildungsleistung inkrementelle Validität besitzt. Diese Annahme konnte nicht bestätigt werden. Zwar klärte der Fleißaspekt für das Kriterium der Vorgesetztenbeurteilung in der technischen Substichprobe tendenziell Varianz auf ($R^2 \text{ adj.}$ = .12; $F(2,28)=3,049$, $p < .10$; s. Tab. 8-43). Mit ΔR^2 = .00 wies der Fleißaspekt allerdings keine inkrementelle Validität auf. Für die weiteren Kriterien in der technischen sowie in der kaufmännischen Substichprobe ergaben sich keine signifikanten Effekte, die für eine Überlegenheit des Fleißaspektes gegenüber dem allgemeinen Gewissenhaftigkeitsmaßes sprachen. Hy-

pothese 10a musste somit verworfen werden. Fleiß besitzt keine inkrementelle Validität gegenüber Gewissenhaftigkeit.

Hypothese 10b nahm an, dass Ordnung gegenüber dem allgemeinen Gewissenhaftigkeitsmaß inkrementelle Validität besitzt. Diese Annahme konnte in der kaufmännischen Stichprobe für das Kriterium der Vorgesetztenbeurteilung bestätigt werden. Ordnung klärte nicht nur 28 % der Gesamtvarianz auf ($R^2 \text{ adj.} = .18$; $F(2,28) = 6,953$, $p < .01$), sondern wies mit $\Delta R^2 = .24$ auch inkrementelle Validität gegenüber der allgemeinen Gewissenhaftigkeit auf (s. Tab. 8-41). Dieses Ergebnis spiegelte sich auch in der technischen Substichprobe im entsprechenden Kriterium wider ($R^2 \text{ adj.} = .31$; $F(2,28) = 7,619$, $p < .01$). Hier klärte der Ordnungsaspekt mit $\Delta R^2 = .17$ ebenfalls inkrementelle Validität gegenüber dem allgemeinen Gewissenhaftigkeitsmaß auf (s. Tab. 8-43). Für das Kriterium der allgemeinen Berufsschulleistung sowie bezüglich der Ergebnisse aus den schriftlichen bzw. praktischen IHK-Zwischenprüfungen konnten allerdings keine inkrementellen Effekte des Ordnungsaspektes gefunden werden. Hypothese 10b konnte somit nur teilweise bestätigt werden.

Hypothese 10c nahm an, dass Fleiß und Ordnung gemeinsam gegenüber dem allgemeinen Gewissenhaftigkeitsmaß inkrementelle Validität besitzen. Diese Annahme konnte für das Kriterium der Vorgesetztenbeurteilung in beiden Substichproben bestätigt werden. Bei den kaufmännischen Auszubildenden klärten die beiden Gewissenhaftigkeitsaspekte signifikant Varianz auf ($R^2 \text{ adj.} = .26$; $F(3,27) = 4,504$, $p < .05$) und besaßen mit $\Delta R^2 = .24$ inkrementelle Validität gegenüber dem allgemeinen Gewissenhaftigkeitsmaß (s. Tab. 8-41). Ähnliche Ergebnisse zeigten sich in der technischen Substichprobe. Hier klärten die Gewissenhaftigkeitsaspekte gemeinsam 29 % der Varianz auf ($R^2 \text{ adj.} = .29$; $F(3,27) = 4,977$, $p < .01$) und wiesen mit $\Delta R^2 = .18$ inkrementelle Validität gegenüber dem allgemeinen Gewissenhaftigkeitsmaß auf (s. Tab. 8-43). Für die weiteren untersuchten Kriterien (allgemeinen Berufsschulleistung, praktische und schriftliche Prüfungsergebnisse) konnte die Überlegenheit der Gewissenhaftigkeitsaspekte allerdings nicht nachgewiesen werden. Hypothese 10c kann somit als teilweise bestätigt angesehen werden.

Bezüglich des Kriteriums der allgemeinen Berufsschulleistung in der kaufmännischen Substichprobe erwies sich das allgemeine Gewissenhaftigkeitsmaß tendenziell als das bessere Modell zur Erklärung der Varianz ($R^2 \text{ adj.} = .09$; $F(1,29) = 3,467$, $p < .10$; s. Tab. 8-42). Für dieses Kriterium musste Hypothese 10 somit eher verworfen werden.

Zudem wurde für alle Kriterien auch die Varianzaufklärung der AGS allein berechnet (Modell 4). Die Gewissenhaftigkeitsaspekte Fleiß und Ordnung wiesen jedoch in keinem der ermittelten Modelle eine höhere Varianzaufklärung auf als die anderen Modelle.

8.4.3 Intrapersonelle Antwortvariabilität

Hypothese 11 postulierte eine Abnahme der intra- und interpersonellen Antwortvariabilität durch die Kontextualisierung der AGS, die sich in höheren Reliabilitäten bzw. geringeren Standardabweichungen der Skala im Vergleich zu einem allgemeinen Gewissenhaftigkeitsmaß ausdrücken sollte. Eine Übersicht über die Ergebnisse ist in Tabelle 8-47 enthalten.

Tabelle 8-47: Vergleich der Reliabilität und Standardabweichungen der Gewissenhaftigkeitsmaße

	Kaufmännische Auszubildende		Technische Auszubildende	
	<i>Cronbachs α</i>	<i>SD</i>	<i>Cronbachs α</i>	<i>SD</i>
Gewissenhaftigkeit (NEO-FFI)	.76 (.65)	.32 (.26)	.82 (.82)	.41 (.41)
Fleiß (AGS)	.85 (.83)	.33 (.32)	.87 (.87)	.38 (.38)
Ordnung (AGS)	.78 (.80)	.36 (.34)	.77 (.82)	.37 (.42)

Anmerkungen: $N_{kfm} = 397$; $N_{techn} = 343$; Werte in Klammern entsprechen den Ergebnissen der Retest-Stichprobe zu T1; $N_{R,kfm} = 31$; $N_{R,techn} = 39$; SD= Standardabweichung.

Hypothese 11a nahm an, dass die Fleißskala reliabler ist als das allgemeine Gewissenhaftigkeitsmaß. Dies traf sowohl für die kaufmännischen als auch technischen Auszubildenden zu. In der kaufmännischen Teilstichprobe betrug die Reliabilität der allgemeinen Gewissenhaftigkeitsskala $\alpha = .76$ (bzw. $\alpha = .65$ in der Retest-Stichprobe) und war somit geringer als die interne Konsistenz des Fleißaspektes mit $\alpha = .85$ ($\alpha = .83$). Auch in der technischen Teilpopulation war die interne Konsistenz der allgemeinen Gewissenhaftigkeitsskala mit $\alpha = .82$ geringer als die des Fleißaspektes mit $\alpha = .87$. Hier gab es keine Unterschiede zwischen der Grundgesamtheit der Ausbildungsbewerber und der ausgewählten Auszubildenden. Hypothese 11a konnte somit bestätigt werden.

Hypothese 11b postulierte, dass die Items der Fleißskala eine geringere mittlere Streuung aufweisen als die Items der allgemeinen Gewissenhaftigkeitsskala. Die Streuung der Items der Fleißskala in der kaufmännischen Teilstichprobe war mit $SD = .33$ ($SD_R = .32$) größer als die Streuung der Items des allgemeinen Gewissenhaftigkeitsmaßes $SD = .32$ ($SD_R = .26$). Die Differenz war in der Retest-Stichprobe vergleichsweise größer. In der Subpopulation der technischen Auszubildenden ergab sich mit $SD = .38$ eine geringere Streuung der Fleiß-Items im Vergleich mit dem allgemeinen Gewissenhaftigkeitsmaß ($SD = .41$). Auch hier unterschieden sich die Ausprägungen zwischen Grundgesamtheit und Retest-Stichprobe nicht voneinander. Hypothese 11b konnte somit ebenfalls für beide Substichproben bestätigt werden.

Hypothese 11c ging von einer höheren Reliabilität des Ordnungsaspektes im Vergleich mit der allgemeinen Gewissenhaftigkeitsskala aus. Dies konnte in der kaufmännischen Teilstichprobe mit $\alpha = .78$ ($\alpha_R = .80$) für den Ordnungsaspekt im Vergleich zu $\alpha = .76$ ($\alpha_R = .65$) für die allgemeine Gewissenhaftigkeitsskala bestätigt werden. Der Effekt trat in der Retest-Teilstichprobe allerdings deutlicher zutage. In der technischen Teilstichprobe war die interne Konsistenz des Ordnungsaspektes mit $\alpha = .77$ kleiner als die des allgemeinen Gewissenhaftigkeitsmaßes ($\alpha = .82$). In der Retest-Stichprobe gab es keine Unterschiede zwischen den beiden Skalen (jew. $\alpha = .82$). Hypothese 11c konnte daher nur für die kaufmännischen Auszubildenden bestätigt werden.

Hypothese 11d postulierte eine geringere Streuung der arbeitsbezogenen Items des Ordnungsaspektes im Vergleich mit der allgemeinen Gewissenhaftigkeitsskala. Diese Annahme konnte für die kaufmännische Subpopulation mit $SD = .36$ ($SD_R = .34$) für den Ordnungsaspekt und $SD = .32$ ($SD_R = .26$) für die allgemeine Gewissenhaftigkeitsskala nicht bestätigt werden. In der Grundgesamtheit der technischen Auszubildenden zeigte der Ordnungsaspekt mit $SD = .37$ eine geringere Streuung als der allgemeine Gewissenhaftigkeitsaspekt mit $SD_R = .41$. In der Retest-Stichprobe ergab sich mit $SD_R = .38$ eine geringere Streuung der Fleiß-Items im Vergleich mit dem allgemeinen Gewissenhaftigkeitsmaß ($SD_R = .41$). Hypothese 11d konnte somit außer für die Grundgesamtheit der technischen Auszubildenden nicht bestätigt werden.

Zusammenfassend lässt sich sagen, dass die Annahmen zur Skalenhomogenität für den Fleißaspekt bestätigt werden konnten. Dieser zeigte eine höhere interne Konsistenz und geringere Streuung auf als das allgemeine Gewissenhaftigkeitsmaß. Die Annahmen zur Skalenhomogenität konnten für den Ordnungsaspekt demgegenüber nur teilweise bestätigt werden.

8.5 Ergebnisse der Überprüfung der Kontrollvariablen

Der folgende Abschnitt enthält die Ergebnisse zur Überprüfung der in Abschnitt 3.2 beschriebenen Kontrollvariablen Alter (8.5.1), Geschlecht (8.5.2), Schulbildung (8.5.3), berufliche Erfahrungen (8.5.4) sowie sozial erwünschtes Antwortverhalten (8.5.5).

8.5.1 Alter

Zur Überprüfung, inwiefern das Alter der untersuchten Personen einen Einfluss auf die Ergebnisse der Untersuchung hatte, wurden Korrelationsanalysen berechnet. Die Ergebnisse zur Überprüfung des Einflusses des Alters auf die Untersuchungsvariablen sind in den Korrelationstabellen in Abschnitt 8.3.1 enthalten (s. Tabellen 8-9 bis 8-11).

Die Überprüfung der Korrelationen, inwiefern ein Zusammenhang der untersuchten Variablen mit dem Alter der Probanden besteht, ergab in beiden Subgruppen der Auszubildendenstichprobe keine signifikanten Zusammenhänge, wie den Ergebnissen der Interkorrelationstabellen 8-9 und 8-10 entnommen werden kann. Sowohl die Ausprägung des Fleißaspekts ($r = -.01$ bis $.18$, n.s.) als auch des Ordnungsaspekts ($r = -.08$ bis $.06$, n.s.) waren somit unabhängig vom Alter der Probanden. Dieser Effekt zeigte sich sowohl für die Ergebnisse des Testtags als auch für die Retestung. Auch für das Selbstwirksamkeitskonstrukt ergaben sich in beiden Substichproben für beide Erhebungsdaten keine signifikanten Effekte ($r = .22$ bis $.31$, n.s.), wenn auch hier von Tendenzen gesprochen werden kann. In Bezug auf die wahrgenommenen eigenen Fähigkeiten ergaben sich ebenfalls sowohl in der Subgruppe der kaufmännischen Auszubildenden ($r = .16$, n.s.) als auch bei den technischen Auszubildenden ($r = -.33$, n.s.) keine signifikanten Zusammenhänge. Allerdings kann der Zusammenhang bei den technischen Auszubildenden als eine Tendenz gewertet werden. Bei gemeinsamer Betrachtung der beiden Stichproben (s. Appendix Tabelle A-16) ergaben sich keine signifikanten Zusammenhänge des Alters mit den beiden Gewissenhaftigkeitsaspekten, weder zu T1 ($r_{FIT1} = .14$ und $r_{OrdT1} = .07$, n.s.), noch zu T2 ($r_{FIT2} = .01$ und $r_{OrdT2} = -.08$, n.s.). Das Selbstwirksamkeitskonstrukt korrelierte nur zu T1 signifikant mit dem Alter ($r_{SET1} = .29$, $p < .05$). Zu T2 ergab sich diesbezüglich kein signifikanter Zusammenhang ($r_{SET2} = .12$, n.s.) für die Gesamtstichprobe der Auszubildenden. Die wahrgenommenen eigenen Fähigkeiten zeigten bei gemeinsamer Analyse der beiden Substichproben ebenfalls keinen signifikanten Zusammenhang mit dem Alter der Probanden ($r_{PA} = -.20$, n.s.).

In Bezug auf die Kriterien ergab sich in beiden Substichproben ein signifikanter Zusammenhang der allgemeinen Berufsschulnoten mit dem Alter ($r_{kfm} = -.39$ und $r_{techn} = -.37$, $p < .05$), was bedeutet, dass ältere Auszubildende im allgemeinen Berufsschulalltag bessere Noten erhielten bzw. erreichten als jüngere Auszubildende. Bei gemeinsamer Betrachtung der beiden Subgruppen zeigt sich diesbezüglich ein starker Zusammenhang ($r_{ges} = -.43$, $p < .01$). Dieser Trend spiegelte sich nur teilweise in den Vorgesetztenbeurteilungen wider, wo die Zusammenhänge in den Subgruppen nicht signifikant wurden ($r_{kfm} = -.28$ und $r_{techn} = .27$, n.s.), hier aber von einem Effekt gesprochen werden kann. Bei gemeinsamer Betrachtung der Auszubildendenstichprobe zeigte sich wiederum ein signifikanter Effekt ($r_{ges} = -.39$, $p < .01$) dahingehend, dass ältere Auszubildende auch seitens der Ausbildungsleiter tendenziell besser bewertet wurden. Für die Ergebnisse der IHK-Zwischenprüfung konnten keine signifikanten Zusammenhänge mit dem Alter gefunden werden ($r_{ZnPges} = .09$). Allerdings zeigten sich differenzierte Tendenzen bei separierter Betrachtung der mündlichen und schriftlichen Prüfungsergebnisse ($r_{mündl.} = -.25$ bzw. $r_{schriftl.} = -.25$, n.s.). Diese deuteten an, dass jüngere Auszubildende in schriftlichen Prüfungen besser sind, während ältere in mündlichen Prüfungen bessere Ergebnisse erreichten.

Für die Schülerkontrollgruppe ergab sich ein signifikanter Zusammenhang des Fleißaspekts mit dem Alter ($r = .21$, $p < .05$) dahingehend, dass ältere Schüler sich tendenziell als fleißiger beschrieben als jüngere. Für den Ordnungsaspekt konnte kein Effekt gefunden werden ($r = .01$, n.s.). Auch das Selbstwirksamkeitskonstrukt zeigte keine Zusammenhänge mit dem Alter der Schüler ($r = .17$, n.s.). Die wahrgenommenen eigenen Fähigkeiten korrelierten wiederum signifikant mit dem Alter ($r = .35$, $p < .01$). Mit dem Kriterium der Abschlussnote hing das Alter ebenfalls nicht zusammen ($r = -.12$, n.s.).

Insgesamt ist nicht davon auszugehen, dass das Alter der untersuchten Probanden einen Einfluss auf die Merkmalsausprägungen in dieser Untersuchung hatte. Allerdings gab es Hinweise darauf, dass das Alter mit den Vorgesetztenbeurteilungen und teilweise den Berufsschulnoten interagierte.

8.5.2 Geschlecht

Zur Überprüfung, inwiefern das Geschlecht der Probanden einen Einfluss auf die Merkmalsausprägungen und die Kriteriumsdaten hatte, wurden Korrelationsanalysen und Mittelwertsvergleiche berechnet.

Für die Subgruppe der kaufmännischen Auszubildenden ergaben sich keine signifikanten Korrelationen des Geschlechts der Untersuchungsteilnehmer mit den Untersuchungsvariablen Fleiß, Ordnung, Selbstwirksamkeit und wahrgenommene eigene Fähigkeiten sowie den Kriterien der allgemeinen Berufsschulnote und der Vorgesetztenbeurteilung, was der Korrelationstabelle 8-9 (s.o.) entnommen werden kann. Dies bestätigte sich auch in den *t*-Tests zum Vergleich weiblicher und männlicher kaufmännischer Auszubildender. Zu beiden Testzeitpunkten unterschieden sich diese Studienteilnehmer weder hinsichtlich der Gewissenhaftigkeitsaspekte Fleiß ($t_{T1} = 0,960$, $df = 29$, $p = .345$ bzw. $t_{T2} = 1,983$, $df = 29$, $p = .057$) und Ordnung ($t_{T1} = 0,243$, $df = 29$, $p = .810$ bzw. $t_{T2} = 0,418$, $df = 29$, $p = .679$), noch bezüglich ihrer Selbstwirksamkeit ($t_{T1} = 0,261$, $df = 29$, $p = .796$ bzw. $t_{T2} = 0,063$, $df = 29$, $p = .950$) und wahrgenommenen eigenen Fähigkeiten ($t = 0,173$, $df = 29$, $p = .864$). Auch für die Kriterien allgemeine Berufsschulleistung ($t = -0,309$, $df = 25$, $p = .760$) sowie Vorgesetztenbeurteilung ($t = -0,328$, $df = 29$, $p = .745$) konnten keine signifikanten Unterschiede gefunden werden.

Die Ergebnisse der Korrelationen der Subgruppe der technischen Auszubildenden sind in Tabelle 8-10 mit dargestellt. Für diese Teilstichprobe ergaben sich für den Gewissenhaftigkeitsaspekt Fleiß ($r = -.12$ bzw. $-.15$, n.s.) sowie das Selbstwirksamkeitskonstrukt ($r = -.13$ bzw. $-.24$, n.s.) keine signifikanten Zusammenhänge mit dem Geschlecht der Probanden. Auch die Kriteriumsdaten wurden den Ergebnissen zufolge nicht durch das Geschlecht der Auszubildenden beeinflusst ($r = -.16$ bis $.13$, n.s.). Allerdings korrelierte der Ordnungsaspekt zu beiden Testzeitpunkten signifikant mit dem Geschlecht der Auszubildenden ($r = -.32$ bzw. $-.43$, jew. $p < .01$). Das heißt, dass sich weibliche Ausbildungsbewerberinnen bzw. Auszubildende eher als ordentlich beschrieben als männliche Auszubildende. Auch bezüglich der wahrgenommenen eigenen Fähigkeiten ergab sich ein signifikanter Zusammenhang ($r = .36$, $p < .01$), der andeutete, dass weibliche Auszubildende ihre eigenen Fähigkeiten tendenziell besser einschätzten als männliche. Dies spiegelte sich auch in den Ergebnissen der *t*-Tests wider, die auf einen überzufällig großen Unterschied zwischen den männlichen und weiblichen Auszubildenden hinsichtlich ihrer Ordnung ($t_{T1} = 2,053$, $df = 37$, $p = .047$ bzw. $t_{T2} = 2,613$, $df = 30$, $p = .014$) sowie wahrgenommenen eigenen Fähigkeiten ($t = -2,129$, $df = 30$, $p = .042$) hindeuteten. Bezüglich des Gewissenhaftigkeitsaspektes Fleiß ($t_{T1} = 0,718$, $df = 37$, $p = .477$ bzw. $t_{T2} = 0,853$, $df = 30$, $p = .400$) und der Selbstwirksamkeit ($t_{T1} = 0,824$, $df = 37$, $p = .415$ bzw. $t_{T2} = 1,346$, $df = 30$, $p = .188$) zeigten sich keine signifikanten Mittelwertunterschiede. Auch für die Kriterien ergaben sich keine signifikanten Unterschiede hinsichtlich der Vorgesetztenbeurteilung ($t = -0,520$, $df = 29$, $p = .607$) oder allgemeinen Berufsschulleistung ($t = -0,133$, $df = 33$, $p = .895$) sowie der IHK-Zwischenprüfungsergebnisse ($t = -0,686$, $df = 28$, $p = .498$). Auch die differenzierte Betrachtung der mündlichen ($t = 0,276$, $df = 24$, $p = .785$) und schriftlichen Prüfungsergebnisse ($t = 0,496$, $df = 24$, $p = .624$) ergab keine signifikanten Geschlechtsunterschiede.

Für die Kontrollgruppe wurden *t*-Tests berechnet, um weibliche und männliche Schüler hinsichtlich der Ausprägungen in den Gewissenhaftigkeitsaspekten Fleiß und Ordnung und den Selbsteinschätzungskonstrukten Selbstwirksamkeit und wahrgenommene eigene Fähigkeiten

miteinander zu vergleichen. Die Ergebnisse deuten an, dass sich die beiden Gruppen nicht hinsichtlich ihres Fleißes ($t = -0,362$, $df = 95$, $p = .718$) und ihrer Ordnung ($t = 1,973$, $df = 95$, $p = .051$) unterscheiden. Auch im Merkmal Selbstwirksamkeit ($t = -0,441$, $df = 95$, $p = .660$) ergaben sich keine signifikanten Unterschiede zwischen weiblichen und männlichen Schülern. Bezüglich der wahrgenommenen eigenen Fähigkeiten ($t = 2,567$, $df = 95$, $p = .012$) deutete das Ergebnis allerdings auf eine überzufällige hohe Abweichung zwischen den Geschlechtern hin. Die Schülerinnen schätzten ihre eigenen Fähigkeiten signifikant höher ein als die Schüler. Auch das Kriterium der Abschlussnote wurde einem t -Test zur Ermittlung der Geschlechterdifferenzen unterzogen. Hier zeigten sich keine signifikanten Unterschiede zwischen Schülerinnen und Schülern ($t = 0,780$, $df = 67$, $p = .780$).

Insgesamt ist nicht davon auszugehen, dass das Geschlecht der Probanden einen Einfluss auf die Vorhersagekraft der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie der Selbstwirksamkeit hatte. Die unterschiedlichen Ergebnisse bezüglich der wahrgenommenen eigenen Fähigkeiten könnten auf einen Einfluss dieser Variablen durch das Geschlecht der Probanden hindeuten. In Bezug auf die verschiedenen in den Studien verwendeten Kriteriumsvariablen scheint das Geschlecht jedoch wiederum keinen Einfluss gehabt zu haben.

8.5.3 Schulbildung

Die Ergebnisse der Zusammenhangsanalysen zur Schulbildung der kaufmännischen und technischen Auszubildenden sowie der Kontrollgruppe sind in den o.g. Tabellen 8-9 bis 8-11 bereits enthalten.

Für die kaufmännischen Auszubildenden zeigten sich keine signifikanten Korrelationen der Gewissenhaftigkeitsaspekte Fleiß ($r_{T1} = .03$ bzw. $r_{T2} = .02$, n.s.) und Ordnung ($r_{T1} = -.01$ bzw. $r_{T2} = -.03$, n.s.) sowie der Selbstwirksamkeit ($r_{T1} = .00$ bzw. $r_{T2} = .05$, n.s.) mit der Schulbildung der eingestellten Bewerber. Auch die wahrgenommenen eigenen Fähigkeiten zeigten keine Zusammenhänge mit der Schulbildung der untersuchten Auszubildenden ($r = .08$, n.s.). Auf Seiten der Kriterien hingen korrelierten die allgemeinen Berufsschulleistung ($r = .15$, n.s.) und Vorgesetztenbeurteilungen ($r = -.26$, n.s.) ebenfalls nicht mit der schulischen Vorbildung der Auszubildenden.

Für die Subgruppe der technischen Auszubildenden zeigten sich vergleichbare Effekte. So korrelierten die Gewissenhaftigkeitsaspekte Fleiß ($r_{T1} = .17$ bzw. $r_{T2} = .19$, n.s.) und Ordnung ($r_{T1} = .10$ bzw. $r_{T2} = .03$, n.s.) sowie die Selbstwirksamkeit ($r_{T1} = .23$ bzw. $r_{T2} = .34$, n.s.) nicht signifikant mit der Schulbildung der eingestellten Bewerber. Letzteres könnte jedoch als Tendenz gewertet werden. Die wahrgenommenen eigenen Fähigkeiten zeigten einen starken signifikanten Zusammenhang mit der Schulbildung der Auszubildenden dahingehend, dass sich Auszubildende mit einer höheren Schulbildung tendenziell auch höhere Eigenschaftsausprägungen zuschrieben ($r = -.51$, $p < .01$). In Bezug auf die verschiedenen Kriterien konnten wiederum keine signifikanten Zusammenhänge der schulischen Vorbildung mit der allgemeinen Berufsschulleistung ($r = -.28$, n.s.) und der Vorgesetztenbeurteilung ($r = -.32$, n.s.) gefunden werden, die jedoch als Tendenzen interpretiert werden könnten. Bezüglich der IHK-Zwischenprüfungsergebnisse zeigten sich wiederum keine statistisch bedeutsamen Zusammenhänge ($r = -.06$ bis $r = .29$, jeweils n.s.).

In der Kontrollgruppe zeigten sich ähnliche Ergebnisse. Auch bezüglich der schulischen Vorbildung ergaben sich bei den Schülern keine signifikanten Zusammenhänge zwischen dem Ordnungsaspekt ($r = -.02$, n.s.) und der Selbstwirksamkeit ($r = .13$, n.s.). Allerdings zeigte der Fleißaspekt einen Zusammenhang mit der Schulbildung dahingehend, dass sich Schüler auf höheren Schulen tendenziell als fleißiger einschätzten ($r = .23$, $p < .05$). Bezüglich der wahrgenommenen eigenen Fähigkeiten zeigte sich ein starker Zusammenhang ($r = .40$, $p < .01$), der andeutete, dass sich Schüler mit höherem Bildungsniveau auch besser bezüglich ihrer

eigenen schulischen Fähigkeiten einschätzten. In Bezug auf das Kriterium der Abschlussnote zeigte sich wiederum kein signifikanter Zusammenhang mit dem Bildungsniveau der Schüler ($r = -.16$, n.s.).

Insgesamt ist somit davon auszugehen, dass die schulische Vorbildung keinen besonderen Einfluss auf die in dieser Studie erhobenen Untersuchungsvariablen hatte. Lediglich in Bezug auf die wahrgenommenen eigenen Fähigkeiten ist es möglich, dass die schulische Vorbildung die Einschätzung dieser beeinflusst hat.

8.5.4 Berufliche Erfahrungen

Die beruflichen Vorerfahrungen wurden in beiden Stichproben mit Hilfe einer bivariaten Variablen erhoben. In der Auszubildendenstichprobe gaben 21 (67,7 %) der 31 untersuchten kaufmännischen Probanden an, bereits über erste berufliche Erfahrungen zu verfügen. In der technischen Subgruppe gaben 16 (55,2 %) der 29 untersuchten Personen an, bereits erste berufliche Erfahrungen zu haben. Die Ergebnisse der Überprüfung, inwiefern sich Auszubildende mit und ohne berufliche Erfahrungen hinsichtlich der Prädiktor- sowie Kriteriumsvariablen unterschieden, sind in Tabelle 8-48 dargestellt.

Tabelle 8-48: Ergebnisse der t-Tests zur Überprüfung der Gruppenunterschiede von Auszubildenden mit und ohne berufliche Vorerfahrungen

	Kaufmännische Auszubildende		Technische Auszubildende	
	df	t	df	t
T1 Fleiß	29	1,508	27	2,696*
T1 Ordnung	29	-0,660	27	1,998
T1 Selbstwirksamkeit	29	0,935	27	1,292
T2 Fleiß	29	0,698	27	3,248*
T2 Ordnung	29	-1,192	27	1,620
T2 Selbstwirksamkeit	29	0,879	27	0,846
Wahrgenommene eigene Fähigkeiten	29	-0,909	27	-0,967
Allgemeine Berufsschulleistung	25	1,389	23	0,230
Vorgesetztenbeurteilung	29	-0,399	26	0,818
IHK-Zwischenprüfung schriftlich	-	-	17	-1,260
IHK-Zwischenprüfung praktisch	-	-	17	1,117
IHK-Zwischenprüfung gesamt	-	-	19	0,659

Anmerkungen: $N_{kfm} = 31$; $N_{techn} = 29$; * $p < .05$.

In der Subgruppe der kaufmännischen Auszubildenden ergaben sich keine signifikanten Unterschiede. In der technischen Subgruppe unterschieden sich die Auszubildenden mit und ohne Berufserfahrungen in ihrer Selbsteinschätzung des Fleißaspekts. Dieser Effekt zeigte sich sowohl zu T1 ($t = 2,696$, $df = 27$, $p < .05$) als auch T2 ($t = 3,248$, $df = 27$, $p < .05$). Zu beiden Testzeitpunkten schrieben sich Bewerber bzw. Auszubildende ohne Berufserfahrung ($M_{T1} = 3,26$ bzw. $M_{T1} = 3,34$) signifikant höhere Ausprägungen im Fleißaspekt zu als Testpersonen, die bereits über erste Berufserfahrungen verfügten ($M_{T1} = 2,89$ bzw. $M_{T1} = 2,99$).

In der Kontrollgruppe gaben 89 (91,8 %) der untersuchten 97 Schüler an, bereits erste berufliche Erfahrungen zu besitzen. Es ergaben sich keine signifikanten Unterschiede zwischen Schülern mit und ohne erste berufliche Erfahrungen. Die Mittelwerte der beiden Gewissenhaftigkeitsaspekte Fleiß ($t = -0,172$, $df = 95$, $p = .86$) und Ordnung ($t = -1,432$, $df = 95$, $p = .16$) unterschieden sich nicht signifikant voneinander. Die Selbstwirksamkeit ($t = 1,055$, $df = 95$, $p = .29$) und die wahrgenommenen eigenen Fähigkeiten ($t = -1,318$, $df = 95$, $p = .19$) wurden von beiden Schülergruppen ebenfalls gleich eingeschätzt. Auch hinsichtlich der schulischen

Leistungen ergaben sich keine signifikanten Unterschiede der beiden Gruppen bezüglich der durchschnittlichen Schulabschlussnote ($t = -0,578$, $df = 67$, $p = .57$).

Insgesamt ist somit davon auszugehen, dass berufliche Vorerfahrungen keinen Einfluss auf die Ausprägungen der Prädiktor- und Kriteriumsvariablen hatten und somit die Ergebnisse dieser Studie nicht durch etwaige berufliche Vorerfahrungen beeinflusst wurden. Eine weitere Berücksichtigung dieser Variablen im Rahmen dieser Untersuchung war somit nicht notwendig.

8.5.5 Sozial erwünschtes Antwortverhalten

Zur Überprüfung, inwiefern sozial erwünschtes Antwortverhalten einen Einfluss auf die Ergebnisse der Untersuchung hatte, wurden Korrelationsanalysen berechnet. Die Ergebnisse zur Überprüfung des Einflusses sozial erwünschter Antworttendenzen auf die Untersuchungsvariablen sind bereits in den Korrelationstabellen in Abschnitt 8.3.1 enthalten (s. Tabellen 8-9 bis 8-11).

In der Stichprobe der kaufmännischen Auszubildenden ergaben sich für die Erhebung der Variablen Fleiß ($r = .03$, n.s.), Ordnung ($r = .09$, n.s.) und Selbstwirksamkeit ($r = .03$, n.s.) zum Testtag (T1) keine signifikanten Zusammenhänge mit der Skala zur Erfassung des sozial erwünschten Antworten (SES-17). Zu T2 zeigten die Variablen Fleiß ($r = .47$, $p < .01$) sowie Selbstwirksamkeit ($r = .38$, $p < .05$) signifikante Korrelationen mit der SES-17. Ordnung ($r = .33$, n.s.) sowie wahrgenommene eigene Fähigkeiten ($r = .11$, n.s.) zeigten keine signifikanten Zusammenhänge, wenn auch für den Ordnungsaspekt von einer Tendenz gesprochen werden kann. In Bezug auf die Kriterien allgemeine Berufsschulleistung ($r = .01$, n.s.) und die Vorgesetztenbeurteilung ($r = .14$, n.s.) erwies sich die SES-17 nicht als prädiktiv.

In der Substichprobe der technischen Auszubildenden korrelierte die SES-17 signifikant mit allen zu T1 erhobenen SelbstbeschreibungsvARIABLEN Fleiß ($r = .47$, $p < .01$), Ordnung ($r = .51$, $p < .01$) und Selbstwirksamkeit ($r = .62$, $p < .01$). Zu T2 zeigten in dieser Substichprobe nur die Gewissenhaftigkeitsaspekte Fleiß ($r = .43$, $p < .05$) und Ordnung ($r = .44$, $p < .05$) signifikanten Zusammenhänge mit der SES-17 auf, Selbstwirksamkeit ($r = .32$, n.s.) und die wahrgenommenen eigenen Fähigkeiten ($r = -.22$, n.s.) korrelierten nicht signifikant mit der Skala, zeigten allerdings tendenziell Zusammenhänge. In Bezug auf die Kriterien erwies sich auch in dieser Stichprobe die SES-17 nicht als prädiktiv für die allgemeine Berufsschulnote ($r = -.06$, n.s.) bzw. die IHK-Zwischenprüfungsergebnisse ($r = -.02$, n.s.). Auch bei differenzierter Betrachtung der praktischen ($r = -.07$, n.s.) und schriftlichen Prüfungsergebnisse ($r = .24$, n.s.) zeigten sich keine signifikanten Zusammenhänge, wobei letzterer als Tendenz gewertet werden kann. Die Vorgesetztenbeurteilung ($r = .08$, n.s.) korrelierte ebenfalls nicht signifikant mit den Kennwerten der SES-17.

In der Kontrollgruppe zeigten sich für die Gewissenhaftigkeitsaspekte Fleiß ($r = .29$, $p < .01$) und Ordnung ($r = .37$, $p < .01$) sowie die Selbstwirksamkeit ($r = .24$, $p < .05$) signifikante Korrelationen mit der sozialen Erwünschtheit. Die wahrgenommenen eigenen Fähigkeiten ($r = .19$, n.s.) korrelierten nicht mit der SES-17. Die Noten der Abschlusszeugnisse zeigten wiederum signifikante Zusammenhänge mit sozialer Erwünschtheit ($r = -.33$, $p < .01$) dahingehend, dass in der Schule besser beurteilte Schüler in der SES-17 höhere Werte zeigten, also eine höhere Tendenz hinsichtlich sozial erwünschter Antworttendenzen aufwiesen.

Insgesamt zeigen die Daten keine eindeutige Tendenz auf. In der Feldstichprobe waren sozial erwünschte Antworttendenzen nicht in der Lage, die Leistungsdaten vorherzusagen, zeigten aber teilweise hohe Zusammenhänge mit den Prädiktoren. Dies unterschied sich jedoch je nach Stichprobe. So ergaben sich in der technischen Auszubildendengruppe vermehrt höhere und signifikante Korrelationen, während in der Stichprobe der kaufmännischen Auszubilden-

den nur partiell signifikante Zusammenhänge gefunden werden konnten. Da keine systematischen Zusammenhänge gefunden werden konnten, wird daher davon ausgegangen, dass sozial erwünschtes Antwortverhalten die Ergebnisse zwar beeinflusst haben könnte, Ein systematischer Einfluss konnte jedoch nicht aufgefunden werden. Eine weitere Berücksichtigung dieser Variable erfolgte daher nicht.

Eine Überprüfung der Mittelwertsunterschiede (s. Abschnitt 8.2.1.1) mittels *t*-Tests ergab, dass sich die Mittelwerte der Auszubildenden und der Kontrollgruppe signifikant voneinander unterschieden. Signifikante Unterschiede wurden für Fleiß ($t(165) = 7,109, p < .01$), Ordnung ($t(165) = 5,348, p < .01$) und Selbstwirksamkeit ($t(165) = 6,666, p < .01$) gefunden.

8.6 Zusammenfassung der Befunde

Zusammenfassend lässt sich festhalten, dass die AGS der zuerst durchgeführten Überprüfung der Skaleneigenschaften auf Ebene der Items (Hypothese 1) standhalten. Nach den Kriterien der Itemtrennschärfe und -schwierigkeit bildeten die AGS zwei voneinander unterscheidbare Faktoren ab, die den vorab definierten Aspekten Fleiß und Ordnung entsprechen. Die beiden Skalen Fleiß und Ordnung erwiesen sich über die verschiedenen Stichproben hinweg als reliabel auf der Basis einer hohen internen Konsistenz.

Darüber hinaus erwiesen sich die beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung als zeitlich relativ stabil (Hypothese 2).

Die Überprüfung des theoretischen Modells der AGS (Hypothese 3) bestätigte sowohl für den Ordnungs- als auch den Fleißaspekt die Eindimensionalität innerhalb der beiden Gewissenhaftigkeitsaspekte. Darüber hinaus luden die Items einer Skala nur auf die korrespondierende latente Variable. Keines der Items lud substantiell auf die latente Variable der anderen Skala.

Auch die Überprüfung der Konstruktvalidität der AGS (Hypothesen 4 und 5) führte zu signifikanten Ergebnissen. Die Ergebnisse der MTMM-Analyse bestätigten die Homogenität der beiden Subskalen der AGS. Sowohl der Fleiß- als auch der Ordnungsaspekt zeigten mittlere, positive Korrelationen zur konstruktkonvergenten Gewissenhaftigkeitsskala. Die konstruktdivergenten Korrelationen zu den Skalen Neurotizismus und Offenheit für Erfahrungen sowie den LPS-Skalen 3 und 4 fielen niedriger als die Zusammenhänge zur konstruktkonvergenten Skala der allgemeinen Gewissenhaftigkeit aus. Zudem erwiesen sich die AGS bzw. der Fleiß- und Ordnungsaspekt in den konfirmatorischen Faktorenanalysen gegenüber dem allgemeinen Gewissenhaftigkeitsmaß als faktoriell valide in dem Sinne, dass keine substantiellen Ladungen auf die latente Variable der anderen Skala vorlagen. Auch gegenüber der Neurotizismus- und Offenheitsskala erwiesen sich die AGS als faktoriell valide.

Die Vorhersagekraft von Ordnung, Fleiß und Selbstwirksamkeit sowie den wahrgenommenen eigenen Fähigkeiten in Bezug auf verschiedene Ausbildungs- und Schulleistungskriterien wurde ebenfalls ermittelt (Hypothese 6). Fleiß und Ordnung waren in den Feldstichproben nicht in der Lage, die Leistungskriterien vorherzusagen, wohl aber in der Kontrollgruppe. Selbstwirksamkeit zeigte bei den kaufmännischen Auszubildenden keine Zusammenhänge mit den Leistungskriterien, war allerdings in der technischen Subpopulation teilweise in der Lage, die Leistungskriterien vorherzusagen. In der Kontrollgruppe war Selbstwirksamkeit ein signifikanter Prädiktor der Schulabschlussnote. Die wahrgenommenen eigenen Fähigkeiten zeigten in den Auszubildendenstichproben teilweise Zusammenhänge mit den Leistungskriterien, erwiesen sich in der Kontrollgruppe demgegenüber allerdings nicht als guter Prädiktor der Schulabschlussnote.

Die Moderations- und Mediationsanalysen (Hypothesen 7 und 8) ergaben uneinheitliche Ergebnisse. In den Auszubildendenstichproben erwiesen sich die Modellannahmen meist nicht als zutreffend. In der Kontrollgruppe kam es dagegen häufiger zu signifikanten Ergebnissen

dahingehend, dass das Selbstwirksamkeitskonstrukt oder die wahrgenommenen eigenen Fähigkeiten als Moderator bzw. Mediator den Zusammenhang der AGS mit der Schulabschlussnote beeinflussten. Allerdings gab es bezüglich dieser Fragestellungen keine eindeutigen Ergebnisse, die für oder gegen ein bestimmtes Modell sprachen.

Die Überprüfung der inkrementellen Validität der AGS (Hypothesen 9 und 10) ergab ebenfalls keine eindeutigen Ergebnisse hinsichtlich einer Überlegenheit der berufsbezogenen Skala gegenüber einem allgemeinen Gewissenhaftigkeitsmaß. Die Ergebnisse zur intra- und interpersonellen Variabilität sprachen für eine Überlegenheit der AGS (Hypothese 11), dennoch konnte der Fleißaspekt Ausbildungserfolg bzw. -leistung nicht signifikant besser vorhersagen als das allgemeine Gewissenhaftigkeitsmaß. Für den Ordnungsaspekt konnte – abhängig vom untersuchten Kriterium – eine Überlegenheit des Aspektes gegenüber dem allgemeinen Gewissenhaftigkeitsmaß aufgefunden werden. Eine gemeinsame Analyse der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung sprach ebenfalls teilweise für eine Überlegenheit der AGS.

9. Diskussion

Der Diskussion wird zunächst eine kritische Bewertung des methodischen Vorgehens vorangestellt. Im Rahmen dieses Abschnittes (9.1) erfolgt eine Analyse eventueller Einschränkungen der Interpretierbarkeit der Ergebnisse. Die darauf aufbauende Diskussion der Ergebnisse besteht aus einer allgemeinen Bewertung und Einordnung der Befunde dieser Studie in die aktuelle Forschung (Abschnitt 9.2). Dabei werden zunächst die Befunde selbst diskutiert, bevor anschließend eine zusammenfassende Bewertung der Ergebnisse vorgenommen wird (9.3). Der letzte Abschnitt (9.4) enthält Implikationen sowohl für zukünftige Forschungsarbeiten in diesem Bereich als auch Hinweise für die Praxis der beruflichen Eignungsdiagnostik.

9.1 Bewertung des methodischen Vorgehens

Die Bewertung des methodischen Vorgehens betrifft verschiedene Bereiche. Hierzu zählen eine Bewertung der internen (9.1.1) und externen Validität (9.1.2) der Untersuchung sowie eine kritische Auseinandersetzung mit der statistischen Validität der Daten (9.1.3) sowie eine kurze zusammenfassende kritische Bewertung des Vorgehens in dieser Studie (9.1.4).

9.1.1 Interne Validität der Untersuchung

Eine Untersuchung ist *intern valide*, wenn die Ergebnisse aus dieser eindeutig interpretierbar sind in dem Sinne, dass die „Veränderungen in den abhängigen Variablen eindeutig auf den Einfluss der unabhängigen Variablen zurückzuführen sind bzw. wenn es neben der Untersuchungshypothese keine besseren Alternativerklärungen gibt“ (Bortz & Döring, 1995, S. 53). Alternativerklärungen für das Vorliegen oder die Höhe der gefundenen Effekte sollten weitestgehend ausgeschlossen werden können. Um dies zu gewährleisten, müssen Störvariablen kontrolliert bzw. durch verschiedene Methoden wie z.B. Elimination, Konstanthaltung und Parallelisierung ausgeschaltet werden. Die interne Validität ist somit ein wichtiges Gütekriterium einer empirischen Untersuchung. Mit Zunahme alternativer Erklärungen für die gefundenen Ergebnisse nimmt die interne Validität ab, weshalb mögliche störende Einflüsse hinreichend kontrolliert werden müssen. Zur Sicherstellung der internen Validität wurden in dieser Studie einige Maßnahmen ergriffen. Diese sollen im Folgenden beschrieben und bewertet werden.

Es gibt verschiedenste Störvariablen, die die Ergebnisse von Untersuchungen beeinflussen können. Man kann diese allgemein in systematische und zufällige Einflüsse unterteilen. An dieser Stelle soll nur auf mögliche systematische Einflussfaktoren Bezug genommen werden, die in der vorliegenden Untersuchung eine Verzerrung der Ergebnisse verursacht haben könnten und auf welche Weise diese Effekte kontrolliert wurden. Eine Darstellung und Diskussion zufälliger Einflüsse wäre zu umfangreich.

9.1.1.1 Drop-out

Unter *Drop-out* oder auch *experimenteller Mortalität* versteht man einen Ausfall von Versuchspersonen, sodass die Gruppen in einer zweiten Messung nicht mehr vergleichbar sind. In der vorliegenden Untersuchung gab es von T1 zu T2 in der Bewerberstichprobe aufgrund der Selektion durch das Unternehmen einen großen Ausfall von Versuchspersonen. Dieser war jedoch vorab bekannt und wurde beispielsweise durch Berechnungen der Range Restriction neutralisiert (Bühner, 2011; Schmidt et al., 2008). Zudem kam es innerhalb der Retest-Stichprobe der Auszubildenden aufgrund einiger weniger Ausbildungsabbrüche sowie einem Todesfall zu einer weiteren Verringerung der Stichprobe. Auch in der Kontrollgruppe haben

nicht alle Schüler das Ziel des Schulabschlusses erreicht. Vor allem in den niedrigeren Bildungsgraden kam es vermehrt zu Schulabbrüchen.

Im Rahmen von Längsschnittuntersuchungen treten allerdings immer Drop-out-Effekte auf. Dies ist ein natürlich auftretender Ausfall von Versuchspersonen, der die Ergebnisse der vorliegenden Studie nicht systematisch beeinflusst haben sollte. Es wird somit davon ausgegangen, dass der Ausfall einiger Versuchspersonen keinen systematischen Einfluss auf die Ergebnisse hatte.

Die systematische Varianzeinschränkung aufgrund der Bewerberselektion wurde mittels Berechnung der Range Restriction berücksichtigt. Dies geschieht allerdings nicht in vielen vergleichbaren Untersuchungen (Lievens et al., 2009) und ist daher als Vorteil dieser Studie zu bewerten. Unberücksichtigt bleiben musste die Varianzeinschränkung, die bereits aufgrund der Vorselektion durch das Unternehmen sowie das dreigliedrige Schulsystem in Deutschland besteht. So wurden die Schüler bzw. Bewerber bereits vorab hinsichtlich ihrer allgemeinen kognitiven Fähigkeiten und implizit auch ihrer Gewissenhaftigkeit vorselektiert (Chamorro-Premuzic & Furnham, 2008). Diese Effekte konnten nicht berechnet werden. Neuere Forschungsarbeiten behaupten sogar, dass Range Restriction kein ernstzunehmendes Problem für Persönlichkeitsmaße sei, da dieses sich eher in einer geringen Power auswirke (Schmidt et al., 2008).

9.1.1.2 Reifung

Unter *Reifung* versteht man intrapersonelle Veränderungen der Versuchspersonen zwischen zwei Messzeitpunkten, die die untersuchten Variablen ebenfalls beeinflussen können. Der Begriff fasst sämtliche physiologischen und psychologischen Veränderungen innerhalb einer Person zusammen, die aufgrund natürlicher Reifungsprozesse geschehen, aber nicht auf eine der unabhängigen Variablen zurückzuführen sind.

Da der Zeitraum zwischen T1 und T2 in der Auszubildendenstichprobe mit ca. einem Jahr sehr groß war, ist nicht auszuschließen, dass in diesem Zeitraum bei den zumeist Jugendlichen verschiedene Reifungsprozesse gewirkt haben. So ist aufgrund des Alters der Probanden, die sich größtenteils in der späten Adoleszenz befanden, mit großen inneren Reifungsprozessen, wie z.B. hormonellen Veränderungen, zu rechnen. Auch Veränderungen der Persönlichkeit könnten in dieser Phase dadurch begründet sein. Das genaue Ausmaß des Einflusses innerer Reifungsprozesse auf die Veränderung von Persönlichkeitseigenschaften ist allerdings bis dato noch unbekannt. Eine Kontrolle dieser Variablen konnte daher nicht implementiert werden. Es ist daher möglich, dass Reifung die interne Validität der Untersuchung beeinflusst hat.

9.1.1.3 Zeiteinflüsse

Unter *Zeiteinflüssen* versteht man – im Gegensatz zu intrapersonellen Reifungsprozessen – Veränderungen in der Umwelt der Probanden, die ebenfalls einen Einfluss auf die Ausprägung der abhängigen Variablen haben können. Bei Zeiteinflüssen handelt es sich um äußere bzw. situationsbedingte Einflüsse. Jedes geplante oder ungeplante Ereignis zwischen zwei Messungen kann einen ungewollten Einfluss auf die Probanden haben. Je größer der Abstand zwischen zwei Messzeitpunkten bzw. zwischen Erhebung der Prädiktoren und Kriterien, desto wahrscheinlicher ist es, dass es zu solchen Einflüssen kommt. Das heißt, dass die gefundenen Zusammenhänge zwischen unabhängigen und abhängigen Variablen auch auf andere Einflüsse zurückgeführt werden könnten bzw. der Zusammenhang zwischen Prädiktoren und Kriterien aufgrund von Zeiteinflüssen verringert sein kann. In beiden Fällen verringert sich die Validität einer Untersuchung.

Aufgrund der sich stark verändernden äußeren Umstände im Leben der Auszubildenden, wozu konkret das Bestehen der Schulabschlussprüfungen und der Beginn der Ausbildung

bzw. die Aufnahme der Arbeitstätigkeit gehören, ist es wahrscheinlich, dass es zwischen T1 und T2 sowie der Erhebung der Kriteriumsvariablen zu zeitlich bedingten Veränderungen gekommen ist, die sich sowohl auf die Persönlichkeitseigenschaften als auch auf die Leistungskriterien der Auszubildenden ausgewirkt haben. Wie der Einfluss der Reifungsprozesse ist auch das Ausmaß der Zeiteinflüsse allerdings nicht bestimmbar, da dieser sich individuell unterschiedlich auswirken kann. Die Retest-Reliabilität der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie des Selbstwirksamkeitsmaßes können hierfür lediglich teilweise als Indikatoren dienen. Die interne Validität dieser Untersuchung kann somit aufgrund von Zeiteinflüssen verringert sein.

9.1.1.4 Testeffekte

Testeffekte können auftreten, wenn die Situation, in der die Daten erhoben werden, reaktiv ist. Das heißt, dass der Messvorgang selbst das Ergebnis beeinflusst. Testeffekte können vor allem in neuen und motivierenden Situationen, wie beispielsweise Personalauswahlverfahren, sowie bei Messwiederholungen auftreten. Es ist möglich, dass die Testpersonen für die Inhalte des Tests sensibilisiert sind und aufgrund dessen anders antworten als unter normalen bzw. neutralen Bedingungen. Dadurch können die Messergebnisse nicht eindeutig auf die unabhängigen Variablen zurückgeführt werden, was zu einer Verringerung der internen Validität der Untersuchung führt und die Eindeutigkeit der gezogenen Schlussfolgerungen gefährden kann. Es ist möglich, dass die Ergebnisse der vorliegenden Untersuchung durch Testeffekte beeinflusst wurden. Durch den Einsatz der Testverfahren in einer Kontrollgruppe und dem Vergleich der Werte wurde in dieser Untersuchung versucht, die Testeffekte zu ermitteln. Allerdings war das Ziel dieser Untersuchung die Vorhersage von Leistungskriterien in einer Feldstichprobe. Das Auftreten von Testeffekten ist kein Spezifikum, das nur in dieser Stichprobe wirkt, sondern auch in vergleichbaren Auswahl-situationen auftreten kann. Die hier gefundenen Ergebnisse können somit als realistisches Abbild von Personalauswahl-situationen interpretiert werden. Zwar ist es möglich, dass zwischen unabhängigen und abhängigen Variablen weniger starke Effekte auftraten, als diese in Laborstudien gefunden wurden, allerdings sind diese aufgrund der Untersuchungseigenschaften nicht weniger bedeutungsvoll.

9.1.1.5 Statistische Regression

Statistische Regression bzw. *Regression zur Mitte* beschreibt ein statistisches Artefakt. Es beschreibt das Problem, dass Extremwerte (T1) in einer zweiten Messung (T2) zur Mitte tendieren (Bühner, 2011). Vor allem bei der Untersuchung von Extremgruppen tritt dieser Effekt häufig auf. Da die in dieser Studie untersuchten Ausbildungsbewerber auch nach den Ausprägungen ihrer Persönlichkeit ausgewählt wurden und hier u.a. hohe Gewissenhaftigkeitsausprägungen wichtig waren, ist es möglich, dass die Ergebnisse der vorliegenden Studie durch die statistische Regression überlagert wurden und es aufgrund dessen zu weniger signifikanten Ergebnissen kam. Die Veränderung der Testwerte der Probanden geht somit nicht nur auf den Einfluss der unabhängigen Variablen zurück, sondern basiert auch auf Wahrscheinlichkeitstheoretischen Überlegungen. Den Annahmen der Standardnormalverteilung der Daten zufolge ist ein durchschnittliches Ergebnis wahrscheinlicher als ein (weiteres) extremes Ergebnis. Je extremer die Werteverteilung in der ausgewählten Gruppe der Bewerber zu T1 war, desto größer ist bei T2 die Regression zur Mitte. Die Ergebnisse der Untersuchung sind in diesem Falle konfundiert, was die interne Validität der Untersuchung verringert.

9.1.2 Externe Validität der Untersuchung

„*Externe Validität* liegt vor, wenn das in einer Stichprobenuntersuchung gefundene Ergebnis auf andere Personen, Situationen oder Zeitpunkte generalisiert werden kann“ (Bortz & Döring, 1995, S. 53). Die externe Validität wird auch als *Allgemeingültigkeit* oder *ökologi-*

sche Validität bezeichnet und betrifft die Frage, inwiefern sich die Ergebnisse einer Untersuchung auf eine bestimmte Praxissituation oder Vergleichsstichprobe übertragen lassen. Sie nimmt mit zunehmender Unnatürlichkeit der Untersuchungsbedingungen und abnehmender Repräsentativität der untersuchten Probanden ab. Allerdings ist die externe Validität keine absolute Größe. Sie bezieht sich auf die Aussagen und Schlussfolgerungen, die mit den Ergebnissen einer Untersuchung für eine bestimmte Zielgruppe getätigt werden. Um die Aussagekraft der Ergebnisse einer Studie bewerten zu können, müssen ausreichend Angaben zur Planung und Durchführung der Studie vorliegen, um einschätzen zu können, ob sich die Ergebnisse auf eine allgemeine Grundgesamtheit verallgemeinern lassen und darüber hinaus auch auf andere Settings übertragen lassen.

Im Folgenden werden einige Überlegungen zur externen Validität dieser Studie getätigt.

9.1.2.1 Repräsentativität der Stichprobe

Bezüglich der Frage der Generalisierbarkeit der Ergebnisse ist oftmals vor allem die personenbezogene externe Validität infrage zu stellen. Dies betrifft v.a. die Art und Weise der Rekrutierung der Versuchspersonen, die Freiwilligkeit der Teilnahme der Probanden an der jeweiligen Untersuchung sowie die Motivation der Testpersonen, die diese zur Teilnahme an der Untersuchung bewegt hat. Da die Mehrheit der psychologischen Experimente im Labor durchgeführt wird, ist die Generalisierbarkeit der darin gefundenen Ergebnisse häufig eingeschränkt.

Da es sich bei der vorliegenden Untersuchung jedoch um ein Feldexperiment mit Vergleichsgruppe handelte, ist grundsätzlich von einer hohen Generalisierbarkeit der Ergebnisse für die Praxis der Personalauswahl auszugehen. Die Probanden nahmen freiwillig am Auswahlverfahren teil.

Untersucht wurde eine große Anzahl von Ausbildungsbewerbern, die sich für verschiedenste Berufe im kaufmännischen und technischen Bereich interessierten. Hinsichtlich ihrer Alters- und Geschlechterverteilung sowie weiteren sozio-ökonomischer Merkmale ist davon auszugehen, dass die Feldstichprobe in vollem Umfang den in der Praxis vorkommenden Bewerbergruppen bei größeren Industrieunternehmen entsprach. Nicht untersucht wurden Interessenten für handwerkliche Berufe sowie ältere Arbeitnehmer, die sich um eine allgemeine Anstellung bewarben. Die Ergebnisse der vorliegenden Studie sind somit für Ausbildungsbewerber im kaufmännischen und technischen Bereich der IHK gültig.

Auch die Kontrollgruppe wurde nicht künstlich hinsichtlich ihrer Zusammensetzung beeinflusst. Es handelte sich um Schüler in Abschlussklassen allgemeinbildender Schulen in den drei gängigen Schulformen in Deutschland. Zwar war die Kontrollgruppe mit den Auszubildenden aufgrund einer Überrepräsentation von Hauptschülern in der Kontrollgruppe nur bedingt mit der Feldstichprobe vergleichbar, dennoch kann aufgrund der sonst ähnlichen Eigenschaften innerhalb der untersuchten Stichproben von einer ausreichenden Vergleichbarkeit der Schüler mit den Auszubildenden ausgegangen werden. Zudem haben Forscher bemängelt, dass viele vergleichbare Experimente überproportional häufig v.a. mit weißen Probanden aus der Mittel- und Oberschicht durchgeführt wurden (Jackson et al., 2009). In diesem Sinne ist die Kontrollgruppe, die zu einem Großteil aus sozial-ökonomisch benachteiligten Jugendlichen bestand, positiv zu bewerten. Allerdings ist es auch möglich, dass es aus diesem Grund zu unterschiedlichen Ergebnissen gekommen ist. Im Nachhinein ist eine Überprüfung dieses Einflusses allerdings nicht möglich. Allgemein weisen Persönlichkeitstests von allen beruflichen Auswahlverfahren aber die geringste soziale Ungerechtigkeit auf (Risavy & Hausdorf, 2011), weshalb davon ausgegangen wird, dass die Unterschiede zu keiner signifikanten Beeinträchtigung der Ergebnisse in der vorliegenden Studie geführt haben.

9.1.2.2 Operationalisierung

Wie in Kapitel 5 beschrieben ist, handelt es sich bei der vorliegenden Untersuchung vor allem um eine Feldstudie, die im Rahmen eines realen Auswahlverfahrens durchgeführt wurde. Der Realitätsbezug der Studie ist somit maximal gegeben. Die angewendeten Testverfahren entsprachen den Untersuchungsmethoden, wie sie standardmäßig bei dem Unternehmen zur Bewerberauswahl im Bereich der beruflichen Ausbildung eingesetzt werden. Auch in der Kontrollgruppe der Schüler wurde keine künstliche Bewerbungssituation erzeugt. Die Schüler bearbeiteten das Fragebogenpaket unter neutralen Bedingungen.

In beiden Stichproben wurden die Ausbildungs- bzw. Schulleistungen der Probanden als abhängige Variablen verwendet. Die betrachteten Variablen entsprachen somit ebenfalls in vollem Umfang den in der Praxis gängigen Erfolgsmaßen. Auch der Untersuchungszeitraum entsprach dem normalen Ausbildungszeitraum.

Einschränkend muss angemerkt werden, dass zur Erfassung der relevanten Konstrukte Fleiß, Ordnung und Selbstwirksamkeit jeweils nur ein Fragebogenmaß eingesetzt wurde. Alle erhobenen Variablen in dieser Untersuchung basieren zudem ausschließlich auf Selbstbeschreibungen, was zu einer Common-Method-Bias geführt haben könnte. Aufgrund der zeitlichen Begrenzungen in allen Untersuchungen war der Einsatz weiterer Fragebogenmaße oder die Befragung relevanter Drittpersonen jedoch nicht möglich. Außerdem entsprach dies dem praktischen Vorgehen in Auswahl-situationen von externen Bewerbern, was wiederum für die Nützlichkeit der Ergebnisse für die Praxis spricht. Bei den meisten verwendeten Verfahren handelte es sich zudem um ausreichend validierte Tests, die ihre Nützlichkeit im Kontext der Personalauswahl bereits mehrfach unter Beweis gestellt haben.

9.1.3 Statistische Validität der Untersuchung

Die interne und externe Validität wurde von Cook und Campbell (1979, zitiert nach Bortz & Döring, 1995) um die *statistische Validität* ergänzt. Diese setzt sich kritisch mit den in der Untersuchung angewendeten statistischen Verfahren sowie weiterer methodischer Aspekte der Datenberechnungen auseinander. Einschränkungen der Interpretierbarkeit der Ergebnisse der Untersuchung können sich auch durch methodische Probleme ergeben. Entsprechend wird auch das methodische Vorgehen einer kritischen Betrachtung unterzogen.

Das methodische Vorgehen in dieser Studie kann in vielerlei Hinsicht als gründlich betrachtet werden. So wurden nach einer Aufbereitung der Daten und Ersetzung der fehlenden Werte zuerst die arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS) hinsichtlich ihrer Skalengqualität gründlich überprüft, um die Nützlichkeit dieser zur Erfassung der Gewissenhaftigkeitsaspekte Fleiß und Ordnung zu erfassen. Zu diesem Zweck wurden u.a. Reliabilitätsanalysen, MTMM-Analysen und konfirmatorische Faktorenanalysen berechnet. In der Untersuchung wurde somit neben bereits gut etablierten und validierten Testverfahren mit den AGS ein ebenfalls messgenaues Instrument eingesetzt.

Anschließend wurde die Vorhersagekraft der arbeitsbezogen formulierten Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie Selbstwirksamkeit zur Prädiktion der verschiedenen Leistungskriterien mittels Korrelationsanalysen überprüft. Mögliche Einflüsse auf die Daten aufgrund von Varianzeinschränkungen oder Unreliabilitäten der Maße zur Erfassung der unabhängigen bzw. abhängigen Variablen wurden kontrolliert.

Wie beispielsweise Bühner (2011) jedoch anmerkt, kann man aus Korrelationsanalysen keine kausalen Schlüsse ziehen. Zur Aussagekraft der Korrelationsanalysen ist in dieser Studie allerdings anzumerken, dass es sich bei der vorliegenden Studie um eine Längsschnittuntersuchung handelt und die Kriteriumsdaten mit großem zeitlichem Abstand nach der Erhebung

der unabhängigen Variablen erhoben wurden. Eine kausale Interpretation der beobachteten Konstruktbeziehungen ist daher möglich.

Darüber hinaus wurden verschiedene Moderations- und Mediationsmodelle berechnet, um die genaue Wirkweise der Variablen bestimmen und verschiedene Modellannahmen zur Vorhersage der Leistungskriterien überprüfen zu können. Zuletzt wurde die inkrementelle Validität der Gewissenhaftigkeitsaspekte Fleiß und Ordnung gegenüber einem allgemeinen Gewissenhaftigkeitsmaß bestimmt und die Skaleneigenschaften der verschiedenen Gewissenhaftigkeitsmaße miteinander verglichen.

Neben diesen positiven Aspekten der statistischen Auswertung gab es allerdings auch einige methodische Probleme, die im Folgenden näher ausgeführt werden sollen.

Die Ergebnisse der Korrelationsanalysen, die Modellüberprüfungen sowie die Berechnungen zur inkrementellen Validität sind nur eingeschränkt interpretierbar, da die untersuchten Stichproben größtenteils zu klein waren. Dies ergab die Überprüfung der Power bzw. Teststärke der Untersuchung (s. Abschnitt 8.3.2). Dies könnte eine Erklärung für die wenigen signifikanten Ergebnisse in den verschiedenen Analysen sein.

Weiterhin konnte aufgrund der uneinheitlichen Antwortformate der Skala zur Erfassung der wahrgenommenen eigenen Fähigkeiten die Skalengüte von dieser nicht berechnet werden. Zwar hatte sich die Skala in vorherigen Untersuchungen als nützlich erwiesen (vgl. Dreier, 2008; Nottle & Robins, 2007), dennoch sollten im Vorfeld oder zumindest im Rahmen der Skalenüberprüfungen Erkenntnisse bezüglich der Messgüte aller verwendeten Maße bestehen. Entsprechend können die Zweifel an der Messgenauigkeit der Skala zur Erfassung der wahrgenommenen eigenen Fähigkeiten nicht ausgeräumt werden, was wiederum die inhaltliche Interpretation von Hypothesen erschwert, die dieses Konstrukt betreffen.

Die Modellüberprüfungen ergaben, dass ein großer Anteil der Varianz in den Daten nicht durch die unabhängigen Variablen erklärt werden konnte. Einige Ergebnisse deuteten darauf hin, dass weitere Einflussvariablen, wie z.B. sozial erwünschtes Antwortverhalten, Reifungsprozesse oder relevante Veränderungen in den Leben der Probanden den Zusammenhang zwischen Prädiktoren und Kriterien sowie auch zwischen den Prädiktoren selbst beeinflusst haben könnten. Zwar wurde ansatzweise der Einfluss dieser Variablen erhoben. Eine tiefergehende Analyse des Einflusses dieser potenziellen Störvariablen wurde jedoch nicht vorgenommen. Es ist möglich, dass hierdurch mögliche Einflüsse zwischen den untersuchten Variablen nicht entdeckt wurden.

Eine alternative Erklärung für die wenigen signifikanten Ergebnisse der Korrelationsanalysen und Modellüberprüfungen könnten auch methodische Artefakte sein, die auf die Schiefe der Kriteriumsvariablen zurückgehen (Le et al., 2011). Da die untersuchten Auszubildenden im Vergleich zu dem durchschnittlichen Auszubildenden der relevanten IHK vergleichsweise bessere Prüfungsergebnisse und Berufsschulnoten zeigten, ist es möglich, dass es auch aufgrund dieses Deckeneffektes zu weniger signifikanten Ergebnissen gekommen ist.

Zudem wurde im Rahmen der wiederholten Modellüberprüfungen keine Adjustierung des Alpha-Fehler-Niveaus vorgenommen. Bei der mehrfachen Analyse derselben Daten ist es statistisch wahrscheinlich, dass die globale Hypothese H_0 aufgrund eines zufällig signifikanten Ergebnisses fälschlicherweise verworfen wird. Aufgrund der mehrfachen Analyse derselben Stichproben wäre in dieser Studie daher die Berechnung einer „test wise error rate“ beispielsweise mittels Bonferroni-Korrektur angemessen gewesen (Bortz, 1993, S. 248 f). Allerdings führt diese eher konservative Korrektur zu einer Verringerung des Alpha-Niveaus, was wiederum das Erreichen signifikanter Ergebnisse zusätzlich erschwert. Aufgrund der geringen Teststärke der vorliegenden Untersuchung ist es allerdings unerheblich, ob das Alpha-Niveau angepasst wurde oder nicht, da ohnehin kaum signifikante Ergebnisse erreicht

wurden. In ähnlichen, zukünftigen Untersuchungen sollte jedoch eine Alpha-Adjustierung vorgenommen werden.

9.1.4 Zusammenfassende Bewertung der Validität der Untersuchung

Die interne und externe Validität einer Untersuchung stehen in einem steten Widerspruch zueinander. Je höher die interne Validität einer Untersuchung sein soll, desto stärker müssen die Einflüsse, die auf unabhängige und abhängige Variablen wirken (können), kontrolliert werden. Dies gelingt allerdings nur in vergleichsweise stark kontrollierten Laborstudien. Die Ergebnisse aus diesen stark reglementierten Laborsettings sind allerdings nur wenig auf die Realität übertragbar und weisen somit eine sehr geringe externe Validität auf. Extern valide Verfahren bergen demgegenüber die Gefahr, dass unkontrollierbare und/ oder nicht beachtete Störeinflüsse die Ergebnisse verzerren (Bortz, 1993; Bortz & Döring, 1995).

Die vorliegende Studie weist im Spannungsfeld zwischen interner und externer Validität als Felduntersuchung eine vergleichsweise geringe interne Validität, demgegenüber jedoch eine hohe externe Validität auf. Im Hinblick darauf, dass bereits viele Laborstudien die Vorhersagekraft von Persönlichkeitseigenschaften in Bezug auf berufliche und schulische Leistungskriterien aufzeigen konnten (Barrick & Mount, 2004; Poropat, 2009; 2011) und vor allem mehr Feldforschung zur Überprüfung der Laborergebnisse gefordert wurde (Sarges, 2010), wurde eine geringere interne Validität in Kauf genommen. Die nachfolgend diskutierten Ergebnisse der vorliegenden Studie sind vor diesem Hintergrund zu bewerten.

9.2 Diskussion der Befunde

Der erste Teil dieser Studie beinhaltete die psychometrische Überprüfung der AGS bzw. deren Items hinsichtlich ihrer Itemeigenschaften sowie Reliabilitäten sowie inhaltlichen Validität. Die Ergebnisse dieses Teils werden in Abschnitt 9.2.1 noch einmal aufgegriffen und diskutiert. Abschnitt 9.2.2 beschäftigt sich mit einer kritischen Reflektion der Überprüfung der Modellannahmen zur Vorhersagekraft der Variablen Fleiß, Ordnung und Selbstwirksamkeit sowie wahrgenommener eigener Fähigkeiten. Der Vergleich der Vorhersagekraft der AGS im Vergleich mit einem allgemeinen Gewissenhaftigkeitsmaß wird anschließend in Abschnitt 9.2.3 diskutiert.

9.2.1 Skalenüberprüfung der AGS

Der erste Teil dieser Studie beschäftigte sich mit der Skalenüberprüfung der Arbeitsbezogenen Gewissenhaftigkeitsskalen. Die Ergebnisse hierzu werden in den folgenden Abschnitten diskutiert. Die ersten Fragestellungen betreffen die Itemkennwerte und internen Konsistenzen der AGS (Abschnitte 9.2.1.1 und 9.2.1.2), bevor die den AGS zugrunde liegenden theoretischen Modellannahmen (9.2.1.3) und die Konstruktvalidität (9.2.1.4) der Skala diskutiert werden.

9.2.1.1 Interne Konsistenz der Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS)

Hypothese 1 betraf die Überprüfung der AGS auf Itemebene. Die Ergebnisse der Skalenüberprüfung ergaben, dass die AGS nach den Kriterien der Itemtrennschärfe und Itemschwierigkeit zwei Faktoren abbilden, die den latenten Konstrukten Fleiß und Ordnung entsprechen. Es war eine leichte Tendenz hin zu einer Ja-Sage-Tendenz der Probanden beobachtbar, da die Mehrheit der Antworten sich oberhalb des Skalenmittelwertes befand. Bei einigen Items wurde das Antwortformat nicht gänzlich ausgenutzt, was andeutet, dass das Antwortformat eventuell zu differenziert war oder die Items statistisch zu leicht waren (Bühner, 2011).

Die beiden Faktoren Fleiß und Ordnung erwiesen sich als reliabel im Sinne einer hohen internen Konsistenz (Cronbach's Alpha). Mit guten internen Konsistenzen ($\alpha = .80$ bis $.86$) für den Fleißaspekt sowie befriedigenden Ergebnissen ($\alpha = .75$ bis $.78$) für den Ordnungsaspekt erwiesen sich die AGS als zuverlässiges Maß zur Erfassung der beiden Gewissenhaftigkeitsaspekte. Hypothese 1 wurde somit bestätigt. Mit der Gewährleistung einer hohen Messgenauigkeit ist eine wesentliche Voraussetzung für die praktische Anwendbarkeit der AGS erfüllt.

9.2.1.2 Zeitliche Stabilität der Arbeitsbezogenen Gewissenhaftigkeitsskalen

Hypothese 2 betraf die Frage, inwieweit die Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie das Selbstwirksamkeitskonstrukt zeitlich stabile Merkmale sind. Diese Annahme wurde mittels einer Analyse der Retest-Reliabilität der beiden Aspekte überprüft. Die Ergebnisse ergaben sowohl für die Gewissenhaftigkeitsaspekte Fleiß ($r = .51$) und Ordnung ($r = .46$) als auch das Selbstwirksamkeitskonstrukt ($r = .50$) vergleichbare, mittlere Zusammenhänge. Eine Überprüfung der Mittelwertsunterschiede ergab für die Gewissenhaftigkeitsaspekte keine signifikanten Abweichungen der Werte zwischen T1 und T2. Dies ließ sich dadurch erklären, dass die untersuchten Bewerber zu T2 sowohl deutlich höhere als auch deutlich niedrigere Selbsteinschätzungen im Vergleich zu T1 abgaben. Ein eindeutiger Trend hin zu einer signifikanten Zu- bzw. Abnahme der Selbsteinschätzungen konnte nicht beobachtet werden. Interessanterweise wiesen die kaufmännischen Bewerber zu T1 im Vergleich zu den technischen Bewerbern höhere Ausprägungen auf den relevanten Variablen Fleiß, Ordnung und Selbstwirksamkeit auf, während sich dieser Effekt zu T2 umkehrte. Die Selbsteinschätzungen hinsichtlich der Selbstwirksamkeit ergaben bei einer Überprüfung der Mittelwertsunterschiede eine signifikante Abweichung dahingehend, dass sowohl die kaufmännischen als auch die technischen Auszubildenden die eigenen Fähigkeiten zu T2 signifikant niedriger einschätzten als zu T1. Letzteres verwundert etwas, da man annehmen könnte, dass die berufliche Selbstwirksamkeit mit der Menge der beruflichen Erfahrungen eher ansteigt. Die Abnahme der beruflichen Selbstwirksamkeit zu Beginn der Ausbildung könnte damit erklärt werden, dass die Auszubildenden von der Komplexität der Ausbildung und des Unternehmens beeindruckt sind und daher die eigenen Fähigkeiten im Vergleich zur Bewerbungssituation deutlich geringer einschätzten. Es ist anzunehmen, dass die berufliche Selbstwirksamkeit nach einer Weile der Arbeitstätigkeit parallel zu der zunehmenden Aneignung von Fachwissen und Fertigkeiten wieder ansteigt. Um dies zu überprüfen, sind jedoch längere Follow-up-Studien mit mehreren Messzeitpunkten notwendig.

Eine weitere Erklärung für die mäßige Retest-Reliabilität bzw. geringere Selbsteinschätzung der Auszubildenden auf dieser Skala zu T2 könnte auch in der Spezifität der Skala selbst begründet sein. Es ist möglich, dass die Auszubildenden zum Testzeitpunkt der Bewerbung einige spezifisch formulierte Items nicht gut einschätzen konnten.

Eine Beeinflussung der Testergebnisse durch die vorab gemachten Erfahrungen der Probanden mit den Selbsteinschätzungsverfahren im Rahmen der Auswahltestung ist aufgrund des langen Zeitintervalls zwischen Prä- und Post-Testung demgegenüber eher unwahrscheinlich. Vielmehr handelte es sich in der vorliegenden Untersuchung um ein vergleichsweise langes Zeitintervall zwischen T1 und T2, unabhängig davon, dass sich die Wahl des Zeitintervalls ohnehin an praktischen Überlegungen orientieren sollte (Bühner, 2011). Die geringen Retest-Reliabilitäten sind somit nicht aufgrund der vergangenen Zeit per se erklärbar.

Bezieht man die Retest-Reliabilität auf die intraindividuelle Werteverteilung einer Person, liegt eine perfekte Reliabilität ($\alpha = 1,00$) dann vor, wenn in aufeinander folgenden Messungen immer derselbe Wert erzielt wird. Da dies jedoch unwahrscheinlich ist, geht man in der Regel bei der Retest-Reliabilität im Vergleich zur internen Konsistenz von etwas geringeren Werten aus (Bühner, 2011).

Laut Schimmack (2010) unterschätzen Retest-Reliabilitäten, die nur auf Selbstauskünften basieren, zudem die Stabilität einer Messung. Dem Autor zufolge werde die beobachtete Stabilität für Selbstauskünfte derzeit mit .46 beziffert. Dieser Wert entspricht den Befunden dieser Studie. Außerdem könnten auch Korrekturen, beispielsweise hinsichtlich der Unreliabilität, zu deutlichen Steigerungen führen. Die Berechnung des Reliable Change Index¹ (RCI) deutete an, dass Messungenauigkeiten für die mäßigen Retest-Reliabilitäten zuständig waren. Eine Korrektur der Unreliabilität führte zu Retest-Reliabilitäten von .60, die als knapp ausreichend bezeichnet werden können.

Die mäßigen beobachteten Retest-Reliabilitäten können darüber hinaus auch auf verschiedene andere Ursachen zurückgeführt werden, die man grob in methodische Artefakte und inhaltliche Überlegungen unterteilen kann.

Eine Voraussetzung für eine möglichst genaue Retest-Reliabilität ist grundsätzlich, die Messungen unter möglichst vergleichbaren Bedingungen durchzuführen (Bühner, 2011). Dies war in der vorliegenden Untersuchung nur bedingt der Fall. Einerseits handelte es sich wiederum um eine Gruppentestung und die Testinstruktion wurde durch die Ausbildungsleiter des Unternehmens vorgenommen. Andererseits war die Testung zu T1 durch die Auswahl-situation charakterisiert, in der die Kandidaten motiviert waren, sich so gut wie möglich darzustellen. Demgegenüber wurde die Nacherhebung im Rahmen einer Einführungsveranstaltung der Auszubildenden durchgeführt, während dieser die Probanden vermutlich weniger motiviert waren, sich besonders vorteilhaft darzustellen. Die signifikanten Zusammenhänge mit dem Maß zur Erfassung der sozialen Erwünschtheit können hierfür als Hinweis dienen, dass sozial erwünschte Antworttendenzen die Testwerte beeinflusst haben könnten. Allerdings müssten dann sämtliche Auszubildende zu T2 signifikant geringere Selbsteinschätzungen aufweisen, was jedoch nicht der Fall ist. Es müssen somit weitere Einflüsse auf die Auszubildenden bzw. Testwerte gewirkt haben.

Grundsätzlich ist es möglich, dass Bewerber, die nach der Höhe ihres Testwertes ausgewählt wurden und gleichzeitig ehrlich geantwortet haben, in einer nachfolgenden Testung niedrigere Werte aufweisen. Es handelt sich hierbei um das statistische Artefakt der Regression zur Mitte (Nye, Do, Drasgow & Fine, 2008), der mit einer Tendenz zu sozial erwünschtem Antwortverhalten jedoch nichts zu tun hat.

In sehr homogenen Stichproben kann es auch aufgrund der geringen Fehlervarianz zu geringeren Retest-Reliabilitäten kommen. In diesem Falle ist die Retest-Methode zur Bestimmung der Reliabilität ungeeignet (Schimmack, 2010).

Weitere Messfehler könnten die geringe Retest-Reliabilität verursacht haben. So könnte es sein, dass spezifische Messfehler aufgetreten sind, wenn die Probanden unterschiedliche Auffassungen bezüglich gleicher Begriffe hatten. Der Kontextbezug der Skalen sollte diese Effekte eigentlich minimieren. Dennoch ist es denkbar, dass die Probanden während des Auswahlverfahrens ein anderes Verständnis der Items hatten, als nach Aufnahme der Arbeitstätigkeit. Dies könnte die unterschiedlichen Skalenausprägungen erklären. Auch könnten zufällige Test- oder Itembearbeitungen, generelle Verständnisschwierigkeiten oder ein Motivationsverlust der Probanden die mäßige Retest-Reliabilität verursacht haben.

Wenn man davon ausgeht, dass die Retest-Korrelation der einzige Indikator für die Stabilität einer Messung ist, läuft man allerdings Gefahr, einen falschen Eindruck von der Stabilität von Persönlichkeitseigenschaften zu vermitteln und die Wahrscheinlichkeit für die Veränderbarkeit der untersuchten Merkmale zu maskieren.

So könnten auch variable äußere oder innere Einflüsse (= States) die Beantwortung der Fragebögen beeinflusst haben, wie z.B. Testangst in Auswahl-situationen. Es ist möglich, dass unterschiedliche Stimmungen, Emotionen und andere mentale Zustände zum Testzeitpunkt

die Messung der Konstrukte bzw. die Selbstauskünfte beeinflusst haben, die zu systematischen Fehlern und somit zur scheinbaren Instabilität der Merkmalsausprägungen geführt haben (Barrick et al., 2001). Bühner (2011) nennt diesbezüglich das Latent-State-Trait-Modell als integrierenden Erklärungsansatz für den (mäßigen) Zusammenhang von Prä-Post-Messungen. Das Modell nimmt an, dass psychologische Messungen nie frei von situationalen Einflüssen sind und neben der systematischen Traitvarianz und der unsystematischen Fehlervarianz auch eine systematische Statevarianz sowie eine systematische Varianz der Wechselwirkung zwischen Trait und Situation den Anteil der Gesamtvarianz beeinflussen. In klassischen Reliabilitätsschätzungen sind jedoch nur die systematischen Varianzanteile enthalten. Die zusätzliche Statevarianz könnte zu einer erhöhten Korrelation zwischen den Items und folglich zu einer höheren Reliabilität zu T1 geführt haben, da für die Bewerber zu T1 die gleichen Mechanismen (z.B. Auswahl-situation, Testangst) gewirkt haben. Der State der Probanden ist allerdings zu jedem Messzeitpunkt ein anderer, was zu geringeren Retest-Reliabilitäten führt.

Wie bereits dargestellt (z.B. Abschnitt 2.5) deuten neuere Forschungsergebnisse zudem an, dass sich Persönlichkeit über das Leben hinweg verändert und hierfür sowohl intrinsische Reifungsprozesse als auch grundlegende Veränderungen der Lebensumgebung bzw. -umstände verantwortlich sind (Specht et al., 2011). Aktuelle Definitionen von Persönlichkeit beziehen sich auf die relative Stabilität der interindividuellen Charakteristika in Bezug auf Gedanken, Gefühle und Verhaltensdispositionen. Dennoch wurden in unterschiedlichen Studien über einen längeren Zeitraum hinweg Unterschiede bzw. Veränderungen der Eigenschaftsausprägungen gefunden (Jackson et al., 2009; Roberts, Walton & Viechtbauer, 2006). Neuere Forschungsarbeiten deuten darauf hin, dass Persönlichkeitseigenschaften sich über das Leben hinweg verändern und nicht so stabil sind, wie immer geglaubt. Roberts und Mroczek (2008) fassten die bestehende Literatur zur Stabilität von Persönlichkeitsmerkmalen zusammen und kamen zu dem Schluss, dass Veränderungen v.a. im jungen Erwachsenenalter auftreten, diese aber auch im mittleren oder späteren Alter noch auftreten können. Diese Veränderungen scheinen durch bestimmte Lebensereignisse bedingt zu werden, die auch an bestimmte Lebensphasen eines Menschen gebunden sind. In Bezug auf die Gewissenhaftigkeitsdimension deuten die Ergebnisse u.a. auf einen vergleichsweise starken Anstieg der Eigenschaftsausprägungen im jungen Erwachsenenalter hin. Allerdings stammen diese Ergebnisse weitgehend aus querschnittlichen Untersuchungen, sodass auch Sozialisationseffekte die Eigenschaftsausprägungen mitbeeinflusst haben könnten. Zudem postulierten bereits Costa und McCrae (1988), dass die intrinsische Reifung um das 30. Lebensjahr herum abgeschlossen sei. Erst danach sei von stabilen Eigenschaftsausprägungen auszugehen. Die Debatte darum, welche Umwelteinflüsse die Entwicklung und Veränderung der Persönlichkeit beeinflussen, dauert noch an.

Das Transaktionale Modell der Persönlichkeit nimmt an, dass Veränderungen sowohl auf intrinsische Reifung als auch auf den Einfluss sozialer Rollen, normativer Veränderungen und grundlegender Lebensereignisse zurückzuführen sind (Roberts et al., 2005). Die Bezeichnung „grundlegende Lebensereignisse“ („major life event“) beinhaltet normative Veränderungen im Leben (z.B. Antritt der ersten Arbeitsstelle), bedeutsame Änderungen (z.B. Geburt eines Kindes) oder allgemein große Erfahrungen (z.B. Arbeitslosigkeit). Da Gewissenhaftigkeit diejenige Eigenschaft ist, die einen bedeutsamen Einfluss auf den eigenen Arbeitsstatus und -erfolg hat, ist es allgemein vorteilhaft, wenn man bei einem Eintritt in den Arbeitsmarkt eine erhöhte Gewissenhaftigkeit zeigt. Den Studienergebnissen von Specht et al. (2011) zufolge stieg die Gewissenhaftigkeit bei Personen, die ihre ersten Berufserfahrungen machten, signifikant mehr an. Die Forscher nahmen an, dass sich Persönlichkeitsausprägungen dann verändern, wenn man am Übergang zu einem neuen Lebensabschnitt steht (z.B. Eintritt in die Berufstätigkeit), einen starken Druck verspürt, sich auf eine bestimmte Art und Weise zu ver-

halten (z.B. in einer Auswahl-situation), bestimmte Antworten oder Reaktionen unerwünscht sind und klare Informationen vorliegen, wie man sich verhalten sollte. Diese Einflüsse haben vermutlich auch auf die Auszubildenden bzw. Bewerber gewirkt und zu den mäßigen Retest-Reliabilitäten geführt.

Da man jedoch auch nicht nur passiv die neuen Lebensumstände auf sich einwirken lässt, sondern diese auch durch das eigene Verhalten bzw. die eigene Persönlichkeit beeinflusst, handelt es sich um ein reziprokes Verhältnis, das die Schwankungen der Selbstbeschreibungen noch besser erklären kann. So tragen die individuellen Persönlichkeitseigenschaften vor allem während des frühen Erwachsenenalters dazu bei, die Umstände der Arbeit zu beeinflussen, was sich wiederum auf die Persönlichkeit einer Person auswirkt (Sutin & Costa, 2010). Diese sogenannten „Sozialisierungseffekte“ sind vor allem in kleineren Firmen oder Abteilungen tätigkeitsübergreifend beobachtbar. Unter „beruflicher Sozialisierung“ versteht man die sozialen Prozesse, mit denen sich neue Mitarbeiter an die Anforderungen einer neuen Tätigkeit anpassen und sich in ihrer organisationalen Rolle einfinden. Der Fokus hierbei liegt auf tätigkeitsrelevanten Verhaltensweisen und Anforderungen, an die sich der neue Mitarbeiter anpassen muss (Satterwhite, Fleenor, Braddy, Feldman & Hoopes, 2009). Mitarbeiter, die schon länger im Unternehmen tätig sind und die relevanten Ziele, Überzeugungen und Werte bereits verinnerlicht haben, dienen hierfür oftmals als Rollenmodelle. Dieser Prozess der persönlichen Assimilation führt zu einer vergleichsweise hohen Homogenität innerhalb der Belegschaft eines Unternehmens. Diese „Tendenz zur Mitte“ kann als eine weitere Ursache für die unterschiedlichen Veränderungen der Selbstbeschreibungen und somit der Retest-Reliabilität angesehen werden. Auch kann dieser Effekt die gegenläufigen Tendenzen der kaufmännischen und technischen Auszubildenden erklären.

Konkret bedeutet dies, dass man bei Eintritt in ein Unternehmen (neue) Verhaltensweisen annimmt, die in diesem wichtig sind (z.B. Pünktlichkeit), um im Beruf bzw. der Firma erfolgreich zu sein. Insbesondere Auszubildende, die ihre ersten Arbeitserfahrungen machen, sind durch die expliziten und impliziten Normen eines Unternehmens bzw. dessen Unternehmenskultur sehr beeinflussbar. Das Verhalten der Auszubildenden beeinflusst darüber hinaus auch wiederum das Verhalten des Unternehmens ihnen gegenüber. Auf diese Weise verfestigen sich bestimmte Verhaltensweisen. Da jeder Auszubildende mit einer individuellen Persönlichkeit in das Unternehmen eintrat, ist es daher möglich, dass einige Probanden ihre Ausprägungen des Fleiß- und Ordnungsaspekts nach oben bzw. unten anpassen mussten bzw. sich zu T2 in dieser Anpassungsphase befanden. Eine Erhebung einer „allgemeinen Unternehmensgewissenhaftigkeit“ und ein mehrfacher Abgleich dieser mit den sich verändernden Merkmalsausprägungen der neu eingestellten Auszubildenden könnte aufklären, ob diese Annahme zutreffend ist. Auch Heine und Kollegen (2008) fanden, dass sich die Kultur und das Selbst einer Person gegenseitig beeinflussen. Es ist also möglich, dass diese Effekte auch im organisationalen Kontext wirken. Eine Überprüfung der beruflichen Sozialisation war jedoch nicht Inhalt oder Ziel dieser Untersuchung. Zukünftige Forschungsarbeiten sollten sich mit der Stabilität bzw. Veränderbarkeit arbeitsbezogener Persönlichkeitsmerkmale bei Berufsanfängern beschäftigen. Diesbezüglich könnten auch Beschäftigte untersucht werden, die den Arbeitgeber wechseln. Auch hier ist es möglich, dass es zu Veränderungen von Eigenschaftsausprägungen als Reaktion auf die neue, veränderte Arbeitsumgebung kommt. Interessant wäre dabei auch zu klären, bis zu welchem Alter solche Veränderungen auftreten, wie stark diese sind und wie langanhaltend.

Es ist auch denkbar, dass die Selbstauskünfte zu T2 durch die ersten Erfahrungen in der Ausbildung beeinflusst wurden, da die Auszubildenden im Sinne einer „realistic job preview“ nun besser einschätzen konnten, inwiefern die Items auf sie zutreffen oder nicht. Dadurch, dass die Auszubildenden zu T2 ein Verständnis der Arbeitsanforderungen hatten, konnten sie die

eigenen Eigenschaften besser einschätzen und veränderten ihre Aussagen von T1 zu T2 entsprechend.

Allerdings haben sich die meisten Studien bisher auf die Untersuchung der allgemeinen Gewissenhaftigkeitsdimension beschränkt. Veränderungen einzelner Subskalen oder Facetten wurden nur selten untersucht. So ist es möglich, dass die gefundenen Veränderungen der allgemeinen Gewissenhaftigkeitsdimension auf unterschiedliche Veränderungen der darunter liegenden Facetten bzw. Aspekte zurückzuführen sind, sich eventuell sogar gegenseitig maskieren. Jackson und Kollegen (2009) untersuchten in einer längsschnittlichen Untersuchung die Veränderbarkeit verschiedener Gewissenhaftigkeitsaspekte. Die Ergebnisse deuten an, dass die Ausprägungen auf der Fleiß-Facette sich vor allem zwischen jungen und mittelalten Erwachsenen unterschieden. Die Forschergruppe berichtete von einem kurvilinearen Effekt, der nach einem frühen Anstieg im jungen Erwachsenenalter zu einer Abnahme der Merkmalsausprägung im hohen Alter führte. Im Gegensatz dazu fanden sie für die Ordnungsfacette keine signifikanten Unterschiede zwischen den verschiedenen Altersgruppen, was sie als Hinweis dafür werteten, dass Ordnung nur wenig zur Veränderung der Gewissenhaftigkeit allgemein beiträgt. Terracciano et al. (2005) fanden ähnliche Ergebnisse in einer längsschnittlichen Untersuchung der Facetten des NEO-PI-R.

Festzuhalten bleibt, dass sich die Facetten von Gewissenhaftigkeit zu verschiedenen Zeitpunkten im Leben zu verändern scheinen. Zukünftige Forschung sollte sich tiefergehend mit der Veränderlichkeit der Gewissenhaftigkeitsfacetten beschäftigen. Vor allem sollte es darum gehen, die Ursachen für die Veränderungen aufzudecken, beispielsweise den Einfluss kritischer Lebensereignisse wie z.B. Eintritt in das Berufsleben oder Berufswechsel. Auch genetische, gesellschaftliche sowie kulturelle Einflüsse sollten hierbei berücksichtigt werden. So zeigen neuere metaanalytische Daten beispielsweise, dass Menschen, die in den späten 1960er Jahren junge Erwachsene waren, einen weniger starken Anstieg der Gewissenhaftigkeit zeigen als diejenigen, die früher oder später im 20. Jahrhundert geboren wurden (Roberts et al., 2006). Es ist möglich, dass das Erleben des sozialen Klimas der 68er Bewegung in der prägenden Jugendphase zu kleineren Anstiegen der Gewissenhaftigkeit geführt hat oder sogar zu Abnahmen der Gewissenhaftigkeitsausprägung.

Letztlich führen diese inhaltlichen Überlegungen zu vier verschiedenen Arten, wie sich die beobachteten Eigenschaftsausprägungen über die Zeit hinweg verändert haben könnten. Zum einen könnte es sich einfach um eine Veränderung des wahren Wertes handeln. Auch eine Veränderung der Eigenschaftsausprägungen aufgrund der veränderten Lebenssituation der Auszubildenden ist möglich. Zum anderen könnte ebenfalls eine Wechselwirkung zwischen der jeweiligen Persönlichkeit und veränderten Lebenssituation zu einer Veränderung der Werte zu T2 geführt haben. Die vierte Möglichkeit ist die Annahme unsystematischer Messfehler, die die mäßigen Retest-Reliabilitäten verursacht haben könnten.

9.2.1.3 Überprüfung des latenten Konstrukts der AGS

Hypothese 3 betraf die Überprüfung des theoretischen Modells, das den AGS zugrunde liegt. Die Überprüfung der Daten konnte mittels konfirmatorischer Faktorenanalysen tendenziell bestätigen, dass es sich bei den Gewissenhaftigkeitsaspekten Fleiß und Ordnung um homogene Faktoren handelte. Zudem konnte trotz einiger methodischer Einschränkungen davon ausgegangen werden, dass den AGS ein zweifaktorielles Modell zugrunde liegt. Die Modellannahmen von DeYoung (2010) für die Gewissenhaftigkeit konnten bestätigt werden. Alle Itemladungen luden eindeutig auf die zugehörige latente Variable. Hypothese 3 wird insgesamt als bestätigt angesehen, womit wiederum eine Voraussetzung für die praktische Anwendung der AGS gegeben ist.

9.2.1.4 Überprüfung der Konstruktvalidität der AGS

Zur Überprüfung der Konstruktvalidität der AGS wurden zwei verschiedene Analyseverfahren angewendet. Zum einen wurde mittels einer MTMM-Analyse die konstrukt-konvergenten und -divergenten Validitäten der AGS überprüft (Hypothese 4). Zum anderen wurde die faktorielle Validität der beiden Gewissenhaftigkeitsaspekte im Rahmen einer konfirmatorischen Faktorenanalyse (CFA) überprüft.

Die Ergebnisse der MTMM-Analyse wiesen hohe Interkorrelationen der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung mit dem allgemeinen Gewissenhaftigkeitsmaß auf, was die Annahmen zur konstrukt-konvergenten Validität stützte. Die Skaleninterkorrelationen zwischen den Aspekten und der allgemeinen Gewissenhaftigkeitsdimension waren vergleichbar mit den Forschungsergebnissen anderer Wissenschaftler zu den Zusammenhängen zwischen der Gewissenhaftigkeitsdimension und den zugehörigen Facetten (Lievens et al., 2008; Lord, 2011). Sowohl der Fleiß- als auch der Ordnungsaspekt erwiesen sich als unabhängige, aber dennoch konstrukt-konvergente Skalen. Die Zusammenhänge mit der allgemeinen Gewissenhaftigkeitsdimension waren hoch. Aufgrund der Asymmetrie zwischen den spezifischeren Aspekten und der allgemeinen Dimension bzw. der unterschiedlich breiten Messung des Gewissenhaftigkeitskonstrukts konnte ohnehin keine maximale Korrelation auftreten (Bühner, 2011). Die Ergebnisse sprechen somit für eine gute Konstruktvalidität der Gewissenhaftigkeitsaspekte. Die Zusammenhänge der beiden Gewissenhaftigkeitsaspekte mit den konstrukt-divergenten Skalen Neurotizismus sowie Offenheit lagen deutlich darunter und deuteten an, dass es sich hierbei um konstrukt-divergente Faktoren handelte. Mit dem Summenwert der LPS-Skalen 3 und 4 ergaben sich keine signifikanten Zusammenhänge, was für die Unabhängigkeit der Konstrukte sprach. Hypothese 4 konnte somit bestätigt werden. Die Ergebnisse der MTMM-Analyse bestätigen die Annahmen zur Konstruktvalidität der AGS.

Hypothese 5 widmete sich der Überprüfung der Konstruktvalidität der AGS mittels CFA. Trotz einiger Einschränkungen, die beispielsweise auf eine verringerte Konstruktvalidität aufgrund des Arbeitsbezugs der Skalen zur Erfassung der beiden Gewissenhaftigkeitsaspekte sowie der Selbstwirksamkeit oder die geringen Zellbesetzungen zurückgehen könnten (Salgado et al., 2003), sind die Ergebnisse der vorliegenden Studie eingeschränkt interpretierbar. Diese bestätigten einerseits unter Berücksichtigung der Fit-Indizes sowie der Faktorladungen tendenziell die theoretischen Modellannahmen, sodass von einer grundsätzlichen Konstruktvalidität des Fleiß- und Ordnungsaspektes ausgegangen werden kann. Andererseits unterschieden sich die Gewissenhaftigkeitsaspekte Fleiß und Ordnung jeweils inhaltlich und strukturell von den Konstrukten Neurotizismus sowie Offenheit des NEO-FFI, sodass es sich um eigenständige Faktoren handelte. Insgesamt konnten die Annahmen von Hypothese 5 somit eher bestätigt werden. Die AGS erwiesen sich gegenüber der Gewissenhaftigkeitsdimension sowie Neurotizismus und Offenheit gegenüber als faktoriell valide.

Zur Neurotizismus-Skala ergab sich ein negativer Zusammenhang, was mit der negativen Valenz des Neurotizismus-Konstruktes im alltäglichen Leben und insbesondere dem Kontext der Berufsauswahl zu erklären ist.

Insgesamt handelt es sich bei den AGS somit um eine faktoriell und inhaltlich valide Skala zur Erfassung der Gewissenhaftigkeitsaspekte Fleiß und Ordnung.

Allerdings wurden die AGS lediglich mit einigen Skalen des NEO-FFI sowie des LPS verglichen. Zukünftige Forschungsarbeiten sollten die Übereinstimmung der Gewissenhaftigkeitsaspekte mit weiteren Persönlichkeitsinventaren überprüfen. Hierzu könnten auch Testverfahren herangezogen werden, die nicht auf dem FFM basieren.

Zudem kann es sein, dass die Faktorenstruktur der AGS in verschiedenen Stichproben unterschiedlich ausfällt. Da bislang auch noch kein Konsens dahingehend herrscht, wie viele

Facetten bzw. Aspekte jeder Dimension des FFM zugrundeliegen (Sutherland et al., 2007), sollte diese Fragestellung weiter untersucht werden. Eine Überprüfung der Konstruktvalidität in weiteren Stichproben ist daher angebracht. Da das Ziel dieser Untersuchung jedoch die Vorhersage von Ausbildungsleistung war, war die hier gewählte Stichprobe zur Konstruktvalidierung der AGS angemessen. Vor einem Einsatz der Skala in anderen Stichproben sollte jedoch eine weitere Konstruktvalidierung der Skala erfolgen. Hierzu sollte auch darüber nachgedacht werden, weitere Verfahren zu verwenden als lediglich MTMM-Analysen und CFA-Verfahren. Jackson et al. (1996) sehen diese Methoden kritisch an, da diese dazu tendierten, die Annahmen nur zu bestätigen und keine Informationen über möglicherweise besser passende Alternativmodelle lieferten.

9.2.2 Überprüfung der Modellpassung

Der zweite Teil dieser Arbeit widmete sich der Überprüfung des Vorhersagemodells bzw. der Überprüfung verschiedener Modellannahmen zur Vorhersage von IHK-Noten mittels der Variablen Fleiß, Ordnung und Selbstwirksamkeit sowie wahrgenommener eigener Fähigkeiten. Die erste Fragestellung betraf die einfachen Zusammenhänge der Prädiktorvariablen mit den verschiedenen Kriteriumsvariablen (Abschnitt 9.2.2.1). Im zweiten Abschnitt werden die Ergebnisse der Moderations- und Mediationsanalysen (9.2.2.2) zur Überprüfung der Modellannahmen diskutiert.

9.2.2.1 Vorhersagekraft von Fleiß, Ordnung, Selbstwirksamkeit und wahrgenommenen eigenen Fähigkeiten

Die Hypothesen 6a bis 6d bezogen sich auf die Vorhersagekraft der Variablen Fleiß, Ordnung und Selbstwirksamkeit sowie der wahrgenommenen eigenen Fähigkeiten. Für die verschiedenen Stichproben lagen unterschiedliche Kriteriumsmaße vor. Sowohl in der kaufmännischen als auch in der technischen Substichprobe dienten die allgemeine Berufsschulleistung der Auszubildenden und eine Vorgesetztenbeurteilung als Kriterien. Für die Substichprobe der technischen Auszubildenden lagen darüber hinaus auch Informationen über die IHK-Zwischenprüfungen vor, die einerseits als Gesamtnote bzw. Gesamtprozente und andererseits auch in schriftliche und praktische Prüfungsergebnisse unterteilt waren. Für die Kontrollgruppe der Schüler lagen lediglich die Abschlussnoten als Kriteriumsmaß vor.

Die Ergebnisse unterschieden sich teilweise deutlich zwischen den verschiedenen Stichproben. In den Auszubildendenstichproben ergaben sich fast keine signifikanten Ergebnisse, was den Powerberechnungen zufolge auf zu geringe Stichprobengrößen zurückzuführen war. Die Ergebnisse müssen daher unter Berücksichtigung dieser Einschränkung interpretiert werden. Weitere Untersuchungen sollten sich der Überprüfung der Vorhersagekraft bzw. Replikation der Ergebnisse in größeren Stichproben widmen.

Die Kontrollgruppe wies demgegenüber eine annehmbare Teststärke auf, sodass die Ergebnisse dieser Stichprobe gegenüber der Feldstichprobe der Auszubildenden interpretiert werden können.

Für den Gewissenhaftigkeitsaspekt Fleiß ergab sich in der Kontrollgruppe ein geringer negativer Zusammenhang mit den Abschlussnoten der Schüler, was bedeutet, dass Schüler, die sich tendenziell als zielstrebig und ehrgeiziger beschrieben, bessere Noten erhielten als weniger fleißige Schüler. Dieses Ergebnis stimmt mit den Annahmen überein.

In der kaufmännischen Stichprobe ergaben sich keine signifikanten Korrelationen.

Auch in der technischen Subgruppe wurden keine signifikanten Zusammenhänge aufgefunden. Allerdings zeigten sich hier mittlere, nicht-signifikante Zusammenhänge mit der Vorgesetztenbeurteilung. Technische Auszubildende, die sich höhere Ausprägungen im Fleißaspekt

zugeschrieben hatten, wurden von ihren Ausbildungsleitern tendenziell besser bewertet. Dieser Effekt zeigte sich konstant sowohl für die Eigenschaftsausprägungen zu T1 als auch bei T2. Ein ähnlicher Effekt zeigte sich für die Ergebnisse des praktischen Prüfungsteils der IHK-Zwischenprüfung. Auszubildende, die sich selbst als fleißiger eingeschätzt hatten, erreichten in dieser tendenziell bessere Ergebnisse als – eigenen Angaben zufolge – „weniger fleißige“ Auszubildende. Auch hinsichtlich der allgemeinen Berufsschulnote konnte dieser Effekt für den Fleißaspekt zu T2 aufgefunden werden. Die Leistungen im schriftlichen Teil der IHK-Zwischenprüfung zeigten weniger große Zusammenhänge mit dem Fleißaspekt.

Bei einer zukünftigen Untersuchung der Vorhersagekraft des Fleißaspektes sollten somit größere Stichproben und unterschiedliche Kriterien überprüft werden. Die vorliegenden Ergebnisse deuten an, dass Fleiß mit verschiedenen Aufgabenanforderungen bzw. Kriterien unterschiedlich stark zusammenhängt. So ist es beispielsweise möglich, dass in interaktiven Settings, wie z.B. praktischen oder mündlichen Prüfungen, sich die Eigenschaften des Fleißaspektes besser auswirken können als beispielsweise in schriftlichen Prüfungen oder Lernkontrollen. Da die Abschlussnoten der Kontrollgruppe sowohl aus schriftlichen als auch mündlichen Leistungen bestehen, war eine nachträgliche Differenzierung des Kriteriums bezüglich dieser Fragestellung nicht mehr möglich.

Der Ordnungsaspekt zeigte in der Kontrollgruppe einen mittleren negativen Zusammenhang mit dem Kriterium der Abschlussnote der Schüler. Dies deutet – vergleichbar zu den Ergebnissen des Fleißaspektes – an, dass Schüler, die sich selbst als strukturierter und planvoller vorgehend beschrieben, auch bessere Abschlussnoten erhielten. Diese Ergebnisse widersprechen einer Vielzahl von Studien, die einen negativen Zusammenhang zwischen Ordnungs-Facetten und (Schul-) Leistungsvariablen berichten (MacCann et al., 2009; Schmit et al., 1995; Ziegler et al., 2010).

In der kaufmännischen Stichprobe ergaben sich keine signifikanten Effekte. Allerdings zeigte der Ordnungsaspekt zu T1 tendenziell einen negativen mittleren Zusammenhang mit der allgemeinen Berufsschulnote der Auszubildenden auf. Das heißt, dass Auszubildende mit höheren Ausprägungen auf der Ordnungsskala tendenziell bessere Berufsschulnoten erhielten als weniger strukturiert arbeitende Auszubildende. Es ergab sich weiterhin eine niedrige positive Korrelation zu T2 mit der Vorgesetztenbeurteilung. Das heißt, dass die Ausbildungsleiter diejenigen Auszubildenden besser beurteilten, die sich selbst als weniger ordentlich beschrieben. Dieser Vorzeichenunterschied ist auffällig. In Hinblick auf die Tatsache, dass sich vor allem die kaufmännischen Auszubildenden sehr hohe Ausprägungen auf dieser Skala zuschrieben, könnte dies andeuten, dass sich die sich als sehr ordentlich arbeitend beschreibenden Auszubildenden zu stark an Regeln und Vorschriften orientieren, längere und aufwendigere Einarbeitungen benötigen und ebenfalls weniger gut in der Lage sind, übergreifende Strukturen zu erkennen. Zukünftige Forschungsarbeiten sollten sich mit dieser Überlegung tiefer auseinandersetzen. Es ist möglich, dass zu hohe Ausprägungen im Ordnungsaspekt in kaufmännischen Berufen erfolgshinderlich sind. Eine Untersuchung dieser Fragestellung könnte beispielsweise durch den Vergleich von Extremgruppen geschehen. Es ist anzunehmen, dass sehr ordentliche Auszubildende zwar bessere Leistungen in fest umschriebenen Settings, wie z.B. der Berufsschule zeigen. In variableren Settings, wie z.B. der betrieblichen Ausbildung, könnte diese Eigenschaft jedoch auch hinderlich sein.

In der technischen Subgruppe ergaben sich diesbezüglich gegenläufige Ergebnisse. Hier zeigte der Ordnungsaspekt zu T2 eine niedrige negative, nicht-signifikante Korrelation mit dem Ergebnis des schriftlichen Teils der IHK-Zwischenprüfung. Das heißt, dass weniger ordentliche Prüflinge höhere Punktzahlen bzw. bessere Leistungen in der Prüfung erreichten. Im praktischen Prüfungsteil ergab sich ein geringer positiver, ebenfalls nicht-signifikanter Zusammenhang mit dem Ordnungsaspekt zu T2. Dies deutet an, dass sehr planvoll vorgehende Auszubildende bessere Noten in der praktischen Prüfung erhielten. Dies könnte man so inter-

pretieren, dass ordentlichere Auszubildenden bzw. Prüflinge ihre Arbeitsschritte nicht nur besser planen, sondern auch erklären konnten und daher bessere Noten erhielten. Weniger ordentlichen Schülern könnte dies weniger gut gelungen sein. Es könnte sein, dass diese Schüler sich in der schriftlichen Zwischenprüfung zu sehr an Details aufgehalten haben und daher beispielsweise weniger Aufgaben bearbeitet haben als weniger ordentliche Auszubildende. Allerdings sind dies nur Vermutungen. Weitere Untersuchungen an größeren Stichproben müssten überprüfen, ob sich sehr ordentliche Auszubildende in den verschiedenen Prüfungsleistungen von weniger ordentlichen Auszubildenden signifikant unterscheiden. Hierzu könnten Subgruppen analysiert werden, die hinsichtlich ihrer Ausprägungen in zwei Gruppen unterteilt werden.

Die Ergebnisse zur prognostischen Validität der AGS deuten an, dass die beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung je nach untersuchter Stichprobe und abhängig vom Kriterium unterschiedlich hohe Vorhersagetendenzen haben. Die Ergebnisse dieser Studie bestätigen tendenziell die Annahmen früherer Forschungsarbeiten, dass die Erforschung der Substruktur von Gewissenhaftigkeit zu einem besseren Verständnis der Vorhersagekraft der Persönlichkeitsdimension führen kann (De Fruyt & Salgado, 2000; Van der Zee et al., 2003) und die Subfaktoren von Gewissenhaftigkeit differenzierte Vorhersagekennwerte aufweisen (Jackson et al., 1996). Zwar ergaben sich – vergleichbar mit den Ergebnissen der Metaanalysen zur Vorhersagekraft der allgemeinen Gewissenhaftigkeitsdimension (Barrick et al., 2001; O'Connor & Paunonen, 2007; Poropat, 2009; Schmidt & Hunter, 1998) – nur moderate Zusammenhänge zwischen Fleiß und Ordnung mit den verschiedenen Leistungskriterien, doch auch diese geringen Korrelationskoeffizienten sind bedeutungsvoll (Viswesvaran et al., 2007). Vor allem in Hinblick auf die starken Varianzeinschränkungen und Spezifika der untersuchten Stichproben ist davon auszugehen, dass die Zusammenhänge in heterogeneren Stichproben noch deutlicher zutage treten. Zukünftige Forschungsarbeiten in größeren Populationen könnten diesbezüglich weiter zur Aufklärung der Zusammenhänge beitragen.

Auffällig waren auch die differenzierten Effekte der beiden Gewissenhaftigkeitsaspekte. Diese Studie ist nicht die erste, die differenzierte Effekte der Fleiß- und Ordnungsaspekte findet. Bereits MacCann und Kollegen (2009) sowie Schmit et al. (1995) fanden, dass Ordnung weniger stark mit Leistungskriterien korrelierte als Fleiß. In dieser Untersuchung zeigte der Ordnungsaspekt in der Subpopulation der kaufmännischen Auszubildenden einen höheren Zusammenhang mit den Kriteriumsvariablen als der Fleißaspekt (T1). In der Subpopulation der technischen Auszubildenden wies demgegenüber Fleiß (T1) höhere Korrelationen mit den verschiedenen Kriteriumsvariablen auf. Diese Ergebnisse können entweder auf Spezifika der Stichproben selbst oder Besonderheiten in der Umgebung der Auszubildenden zurückzuführen sein. So ist es möglich, dass bestimmte Situationsparameter das Zeigen fleißiger bzw. ordentlicher Verhaltensweisen begünstigt bzw. gehemmt haben könnten. Bereits Sutherland und Kollegen (2007) schlussfolgerten aus den nicht-signifikanten Zusammenhängen zwischen Gewissenhaftigkeit und Arbeitsleistung in einem von ihnen durchgeführten Feldexperiment, dass eine hohe Strukturierung der Arbeitsbedingungen dazu führen könnte, dass Unterschiede in den Gewissenhaftigkeitseigenschaften weniger zum Tragen kommen. Sutin und Costa (2010) nahmen an, dass sich gewissenhaftere Menschen eher (Arbeits-) Umgebungen suchen, die ihnen mehr Freiheit zur selbstständigen Gestaltung ihrer Arbeit lassen. Der Erfolg, den die strukturiert und eigeninitiiert arbeitenden Menschen dabei haben, führt dazu, dass diese in den entsprechenden Umgebungen verbleiben. In Bezug auf diese Studie könnte eine Arbeitsanalyse Aufschluss darüber geben, wie stark strukturiert die jeweiligen Arbeitsumgebungen der Auszubildenden sind. Es ist denkbar, dass die Strukturiertheit eines Arbeitsplatzes als Mediator des Zusammenhangs zwischen Gewissenhaftigkeitsaspekt und Arbeitsleistung fungiert.

Eine weitere Erklärungsmöglichkeit der differenziellen Vorhersageeffekte könnte auf eine höhere inhaltliche Übereinstimmung der untersuchten Gewissenhaftigkeitsaspekte mit den verschiedenen Kriterien zurückzuführen sein. Gemäß den Schlussfolgerungen von Dudley et al. (2006) weisen Facetten bzw. Aspekte im Vergleich mit einer allgemeinen Persönlichkeitsdimension mehr Gemeinsamkeiten mit den Kriterien auf, was zu höheren Korrelationen führt. Auch dies könnte mit einer Arbeitsanalyse überprüft werden.

In Bezug auf das Selbstwirksamkeitskonstrukt ergab sich ein mittlerer, negativer Effekt in der Kontrollgruppe. Das heißt, Schüler mit höheren Abschlussnoten schätzten ihre eigenen Fähigkeiten tendenziell besser ein als Schüler mit schlechteren Abschlussnoten. Diese Ergebnisse sind kongruent zu bestehenden Forschungsarbeiten, die die Vorhersagekraft des Selbstwirksamkeitskonstruktes im schulischen Bereich zeigen konnten (Bell & Kozlowski, 2002).

In der Gruppe der kaufmännischen Auszubildenden ergaben sich keine signifikanten Zusammenhänge mit dem Selbstwirksamkeitskonstrukt. Für die Berufsschulleistung und die Beurteilung durch die Ausbildungsleiter war es demzufolge unerheblich, ob ein Auszubildender sich viel oder wenig zutraute. Möglich ist aber auch, dass Bewerber, die sich eine sehr hohe Selbstwirksamkeit zuschrieben, ihre eigenen Fähigkeiten tendenziell überschätzten. Dies ist Bandura und Wood (1989) zufolge zwar eine natürliche Reaktion „normaler“ Menschen, könnte hier aber zu einer Beeinträchtigung der Kriteriumsvalidität geführt haben, da die später gezeigte Leistung nicht den Annahmen entsprach.

In der Gruppe der technischen Auszubildenden ergab sich ein mittlerer, negativer Effekt der Selbstwirksamkeit zu T2 in Bezug auf die Vorgesetztenbeurteilung. Dieser Effekt war vergleichbar mit den Ergebnissen, die andere Forscher bei der Untersuchung von Berufsanfängern gefunden haben (Abele et al., 2000). Technische Auszubildende, die ihre eigenen Fähigkeiten allgemein optimistischer einschätzten, wurden auch durch ihre Ausbildungsleiter besser beurteilt. Ein etwas geringerer, allerdings nicht-signifikanter negativer Effekt, ergab sich auch für das Selbstwirksamkeitskonstrukt zu T1. Zudem zeigte die allgemeine Berufsschulleistung der technischen Auszubildenden einen mittleren, negativen, aber nicht-signifikanten Zusammenhang mit dem Selbstwirksamkeitskonstrukt, welches zu T2 erhoben wurde. Insgesamt scheint Selbstwirksamkeit also ein Prädiktor des Zusammenhanges mit den verschiedenen Leistungskriterien zu sein.

Zukünftige Forschungsarbeiten sollten den Zusammenhang des Konstruktes in größeren Stichproben replizieren, um aussagekräftige Ergebnisse zu erhalten. Dabei sollte vor allem auch hinsichtlich der verschiedenen Ausbildungsarten unterschieden werden. Die Befunde dieser Studie deuten an, dass Selbstwirksamkeit im technischen Bereich bedeutsamer ist als in kaufmännischen Berufen. Diese Ergebnisse sind konsistent zu Befunden von Schyns und Moldzio (2009), die in einer vergleichbaren Stichprobe von Ausbildungsbewerbern um eine kaufmännische Lehre nur kleine Zusammenhänge des Selbstwirksamkeitskonstruktes mit verschiedenen schulischen Leistungskriterien auffinden konnten. Die Autoren untersuchten allerdings lediglich Zusammenhänge mit zuvor abgefragten Schulnoten. Die von ihnen gefundenen Zusammenhänge waren zwar signifikant, aber sehr gering ($r = -.06$ bis $-.13$). Auch der Einfluss weiterer, bislang nicht beachteter Variablen ist möglich. Zudem handelte es sich sowohl in der Studie von Schyns und Moldzio (2009) als auch in dieser Untersuchung um eine bereits stark vorselektierte Stichprobe. Die positiven Erfahrungen, die diese Probanden bereits in der Schule und im Verlaufe des Auswahlprozesses gemacht hatten, könnten die Selbstwirksamkeit dieser Probanden nachhaltig positiv beeinflusst haben (Bandura, 1977). Zudem waren die kaufmännischen Bewerber vergleichsweise älter als die technischen Bewerber und die Probanden der Kontrollgruppe. So könnte das Alter der untersuchten Probanden die Zusammenhänge zwischen Selbstwirksamkeit und den verschiedenen Leistungsvariablen beeinflusst haben. Darüber hinaus ist auffällig, dass die Streuung der Mittelwerte in der kaufmännischen Stichprobe deutlich geringer ist als in der technischen Stichprobe oder der Kontrollgruppe. Es

ist auch denkbar, dass die geringe Varianz der Daten die geringen Zusammenhänge der Variablen verursacht hat und es sich somit um ein methodisches Artefakt handelt. Diesbezüglich sollten weitere Studien die Zusammenhänge der Variablen in heterogeneren Stichproben untersuchen.

Des Weiteren ist es möglich, dass die Kontextualisierung der Items sowohl der AGS als auch der Selbstwirksamkeitsskala sich in dieser Untersuchung negativ auf die Vorhersagekraft der Skalen in Bezug auf die Kriterien ausgewirkt hat. So zeigten verschiedene Forschungsarbeiten, dass sich ein inkongruenter Bezugsrahmen nachteilig auf die Kriteriumsvalidität des Maßes auswirken kann (Lievens et al., 2008; Reddock et al., 2010). In der vorliegenden Studie bearbeiteten die Ausbildungsbewerber sowie die Schüler der Kontrollgruppe arbeitsbezogen formulierte Skalen zur Erfassung der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie Selbstwirksamkeit. Allerdings sollten in dieser Untersuchung vor allem schulbezogene Kriterien vorhergesagt werden. Es lag somit eine inhaltliche Diskrepanz zwischen Prädiktor und Kriterium vor. Lediglich die Vorgesetztenbeurteilung der Auszubildenden war ein eher arbeitsbezogenes Kriterium. Mit diesem Kriterium ergaben sich tendenziell auch die meisten signifikanten Ergebnisse in den beiden Auszubildendenstichproben. Dieser Punkt sollte bei der Interpretation der vorliegenden Ergebnisse zur Vorhersagekraft und dem Zusammenhang der Gewissenhaftigkeitsaspekte und des Selbstwirksamkeitskonstrukts in Bezug auf die verschiedenen Leistungskriterien und Vorhersagemodelle bedacht werden.

Die letzte Teilhypothese zu dieser Fragestellung (H6d) postulierte einen Zusammenhang zwischen den eigenen wahrgenommenen Fähigkeiten und den verschiedenen Leistungskriterien.

In der Kontrollgruppe ergab sich diesbezüglich ein geringer negativer, aber nicht-signifikanter Zusammenhang mit der Schulabschlussnote. Die Teststärke war groß genug, um von einem aussagekräftigen Ergebnis sprechen zu können.

In der kaufmännischen Substichprobe ergaben sich ebenfalls keine signifikanten Ergebnisse mit dem Konstrukt. Allerdings war hier die Teststärke nicht groß genug, um von aussagekräftigen Ergebnissen sprechen zu können. Es ergab sich ein positiver, mittlerer, aber nicht-signifikanter Zusammenhang mit der allgemeinen Berufsschulleistung sowie der Vorgesetztenbeurteilung. Dieser positive Zusammenhang könnte andeuten, dass Auszubildende, die ihre eigenen Fähigkeiten allgemein optimistischer bzw. besser einschätzen als die ihrer Mitschüler bzw. Kollegen, auch bessere Leistungen in der Berufsschule bzw. Ausbildung erreichen, was den Befunden von Nofle und Robins (2007) sowie Dreier (2008) entspricht.

In der Subgruppe der technischen Auszubildenden ergaben sich differenzierte Ergebnisse, die teilweise signifikant wurden. Hinsichtlich der Vorgesetztenbeurteilung ergab sich ein mittlerer, positiver Effekt, der andeutete, dass Techniker, die ihre eigenen Fähigkeiten im Vergleich zu ihren Kollegen als besser einschätzten, auch von ihren Vorgesetzten besser beurteilt wurden. Diese Auszubildenden erreichten auch bessere Noten in der Berufsschule. Mit den Ergebnissen des praktischen Teils der IHK-Zwischenprüfung ergaben sich ebenfalls signifikante, negative Zusammenhänge. Da dieses Kriterium als maximal erreichte Punktzahl operationalisiert wurde, ist dieses Ergebnis inhaltlich kongruent zu den zuvor berichteten positiven Zusammenhängen. In Bezug auf die schriftlichen Prüfungsleistungen konnte kein Effekt der wahrgenommenen eigenen Fähigkeiten gefunden werden. Diese schien unabhängig von der Einschätzung der eigenen Fähigkeiten zu sein.

Insgesamt sind die Ergebnisse der eigenen wahrgenommenen Fähigkeiten betreffend uneindeutig. Aufgrund der geringen Stichprobengrößen können keine eindeutigen Aussagen getroffen werden. Die signifikanten Zusammenhänge in der technischen Substichprobe deuten an, dass die Wahrnehmung der eigenen Fähigkeiten als gut bzw. besser als andere prädiktiv für das Erreichen guter Ausbildungsergebnisse ist, was die Befunde von Nofle und Robins (2007) stützt.

Auffällig war auch die unterschiedliche Höhe der Ausprägungen zwischen Schülerinnen und Schülern. Den Ergebnissen zufolge nahmen die Schülerinnen die eigenen Fähigkeiten signifikant besser wahr als die Schüler. Dieser Effekt sollte weiter untersucht werden. Da Schülerinnen tendenziell auch bessere Leistungen in der Schule zeigen als ihre männlichen Mitschüler, sollte überprüft werden, ob dies die Ursache für die diskrepanten Selbsteinschätzungen war. Auch sollte längsschnittlich überprüft werden, inwiefern diese Beobachtung zeitlich stabil ist oder sich im Laufe der Berufsausbildung bzw. dem Erwachsenwerden eventuell die Einschätzungen der eigenen Fähigkeiten zwischen weiblichen und männlichen Probanden angleichen.

Zukünftige Studien sollten das Konstrukt weiter erforschen, da es im Vergleich zu den anderen Variablen vergleichsweise hohe Zusammenhänge mit den Leistungskriterien aufwies. Allerdings sind die wahrgenommenen eigenen Fähigkeiten ein noch relativ wenig beforschtes Konstrukt, das vor allem im Kontext der beruflichen Eignungsdiagnostik bisher weder in der Forschung noch in der Praxis beachtet wurde. Diese Arbeit kann als erster Schritt in diese Richtung angesehen werden. Weitere Forschungsarbeiten sollten sich sowohl der Überprüfung des Konstruktes selbst sowie seiner Nützlichkeit in der Eignungsdiagnostik widmen.

Zusammenfassend zur prognostischen Validität der vier Konstrukte ist anzumerken, dass auch Methodenfaktoren zu einer Verringerung der Kriteriumskorrelationen geführt haben könnten. So besagt Bühner (2011), dass Korrelationen niedriger ausfallen können, wenn verschiedene Methoden verwendet werden, um Prädiktor und Kriterium zu erfassen, was in der vorliegenden Untersuchung der Fall war. Es ist auch möglich, dass per se Messfehler die nicht-signifikanten Ergebnisse verursacht haben könnten. Messfehler kann man nicht verhindern, vor allem bei der Evaluation so komplexer Variablen wie Persönlichkeitseigenschaften (Schimmack, 2010).

Umstritten ist nach wie vor, ab welchem Zeitraum man von prädiktiver Validität sprechen kann. Je nach Studie bzw. Lehrbuch variieren diese Angaben (Bortz & Döring, 1995; Bühner, 2011; Lienert & Raatz, 1994). Es ist bei der vorliegenden Studie jedoch davon auszugehen, dass zwischen Prädiktor und Kriterium ein ausreichend großes Zeitintervall vorlag, um von prognostischer Validität sprechen zu können. Allerdings merken Judge et al. (2007) an, dass aufgrund der geringeren Kontrollierbarkeit von Störvariablen es in Feldstudien tendenziell zu geringeren Validitäten kommen kann. Andere Studien sprechen sich sogar für sehr lange Zeiträume aus, da sie annehmen, dass auch kleinere Effekte über die Zeit hinweg größere Auswirkungen haben können. So können beispielsweise über das Leben hinweg kleinere schulische Leistungsunterschiede oder leichte Unterschiede in den individuellen Ausprägungen der Gewissenhaftigkeit sich kumulieren und in der späten Zukunft zu großen Leistungsunterschieden führen. Verhalten wird verstärkt und aufrechterhalten ähnlich eines „Schneeballprinzips“, was auch „kumulative Kontinuität“ genannt wird (Noftle & Robins, 2007).

Die signifikanten Vorhersagewerte in der Kontrollgruppe sowie die Tendenzen in den beiden Auszubildendenstichproben entsprachen in ihrer Höhe allerdings den Vorhersagewerten, die in den einschlägigen Metaanalysen berichtet werden. In Anbetracht der starken Varianzeinschränkungen in der vorliegenden Studie sind diese Ergebnisse hervorzuheben. Eine Replikation der Studie mit größeren und ggf. heterogeneren Stichproben könnte zu deutlich höheren Vorhersagewerten führen. In diesem Sinne sind die hier gefundenen Ergebnisse für die Vorhersage von Ausbildungsleistung vielversprechend.

9.2.2.2 Überprüfung des Vorhersagemodells

Die Hypothesen 7 und 8 bezogen sich auf die Zusammenhänge der Gewissenhaftigkeitsaspekte mit dem Selbstwirksamkeitskonstrukt und den wahrgenommenen eigenen Fähigkeiten in Bezug auf eine Verbesserung der Vorhersagekraft. Die Überprüfung der Annahmen er-

folgte mittels Moderations- und Mediationsanalysen. Ziel der Analysen war es, die Beziehungen zwischen den Persönlichkeitskonstrukten und Leistungsvariablen genauer zu untersuchen, wie es von verschiedenen Forschern gefordert wird (Dudley et al., 2006; Funder, 2006; Lord, 2011).

Hypothese 7 postulierte, dass Selbstwirksamkeit sowie die wahrgenommenen eigenen Fähigkeiten sich aufgrund der Erfahrungen, die die Auszubildenden bzw. Schüler machen, verändern und aus diesem Grund die Richtung und/ oder Stärke zwischen den Gewissenhaftigkeitsaspekten Fleiß und Ordnung in Bezug auf die verschiedenen Leistungskriterien beeinflussen.

Die angenommenen moderierenden Effekte konnten jedoch weder in der kaufmännischen noch in der technischen Subgruppe gefunden werden.

Auch in der Kontrollgruppe ergaben sich weitestgehend keine signifikanten Moderationseffekte. Lediglich Selbstwirksamkeit schien den Zusammenhang zwischen dem Ordnungsaspekt und der Schulleistung zu moderieren. Die Ergebnisse deuteten an, dass je mehr ein Schüler von den eigenen Fähigkeiten überzeugt ist und optimistisch in die schulische Zukunft blickte, desto eher konnte er die eigenen Bewältigungsmechanismen nutzen, um bessere Schulleistungen zu erreichen. Konkret deutet dies an, dass erfolgreiche Schüler strukturierter und planvoller vorgehen, um gute Noten zu erreichen.

Hypothese 8 postulierte, dass Selbstwirksamkeit bzw. die wahrgenommenen eigenen Fähigkeiten als Mediatoren den Einfluss zwischen den Gewissenhaftigkeitsaspekten und den verschiedenen Leistungskriterien beeinflussen.

Die erste Teilhypothese nahm einen vermittelnden Einfluss des Selbstwirksamkeitskonstrukts zwischen dem Fleißaspekt und den verschiedenen Ausbildungserfolgs- bzw. Schulleistungskriterien an. Ein solcher Effekt konnte in den verschiedenen Stichproben lediglich für das Kriterium der praktischen Zwischenprüfung in der technischen Auszubildendengruppe gefunden werden. In der Kontrollgruppe der Schüler ergab sich eine partielle Mediation des Selbstwirksamkeitskonstrukts für den Fleißaspekt und die Abschlussnote.

Für den Ordnungsaspekt ergaben sich in der technischen Subgruppe wiederum Tendenzen, die andeuteten, dass Selbstwirksamkeit in dieser Substichprobe als Mediator zwischen Ordnung und der Vorgesetztenbeurteilung bzw. der allgemeinen Berufsschulleistung agiert. In der Kontrollgruppe der Schüler ergab sich eine signifikante, partielle Mediation des Selbstwirksamkeitskonstrukts zwischen dem Ordnungsaspekt und der Schulabschlussnote.

Als weiterer Mediator wurden auch die wahrgenommenen eigenen Fähigkeiten untersucht. Hier ergab sich in der Subgruppe der kaufmännischen Auszubildenden zwar kein signifikantes Ergebnis, es gab jedoch eine Tendenz hin zu einer partiellen Mediation der wahrgenommenen eigenen Fähigkeiten zwischen dem Fleißaspekt und den Kriterien. In der technischen Subgruppe ergab sich eine signifikante Mediation des Konstrukts für das Kriterium der allgemeinen Berufsschulleistung und eine vermittelnde Tendenz in Bezug auf das Kriterium der Vorgesetztenbeurteilung. In der Kontrollgruppe ergab sich demgegenüber keine signifikante Mediation.

Zwischen dem Ordnungsaspekt und den Kriterien ergaben sich für die wahrgenommenen eigenen Fähigkeiten in der Subgruppe der kaufmännischen Auszubildenden wiederum Tendenzen in Bezug auf das Kriterium der Vorgesetztenbeurteilung. In der Subgruppe der technischen Auszubildenden lagen signifikante Mediationseffekte für alle Kriterien vor, mit Ausnahme der Ergebnisse der schriftlichen IHK-Zwischenprüfungen. In der Kontrollgruppe konnten wiederum keine mediierenden Effekte nachgewiesen werden.

Die Ergebnisse der Modellüberprüfungen sprechen somit eher dafür, dass Selbstwirksamkeit und wahrgenommene eigene Fähigkeiten nicht als Moderatoren, sondern Mediatoren wirken.

Allerdings waren die Ergebnisse je nach untersuchter Stichprobe und abhängig vom Kriterium sehr uneinheitlich, weshalb keine klare Aussage für oder gegen ein Modell getroffen werden kann. Es gibt bisher auch nur wenige Versuche, Selbstwirksamkeit und die wahrgenommenen eigenen Fähigkeiten hinsichtlich ihres Einflusses auf die Vorhersagekraft zwischen Persönlichkeitseigenschaften und Berufsschulerfolg zu überprüfen. Laut Jackson et al. (2007) müssen hierzu auch viele Randbedingungen konstant gehalten oder kontrolliert werden können, um signifikante Ergebnisse zu erhalten. Dies war jedoch in der vorliegenden Felduntersuchung aufgrund der o.g. Gründe (s. Abschnitt 9.1) schwierig.

Als problematisch erwiesen sich auch hier wieder die kleinen Stichproben, die oftmals dazu führten, dass bereits die Vorbedingungen der Moderations- und Mediationsanalysen nicht erfüllt waren. Die geringen Teststärken sind vermutlich die Ursache für die nicht-signifikanten Ergebnisse in den Moderations- und Mediationsanalysen. So wiesen auch die Alternativmodelle in den meisten Fällen keine signifikante Modellpassung auf, was ein weiteres Indiz für die mangelnde Teststärke als Ursache für die fehlenden Effekte gewertet werden kann. Allerdings waren die untersuchten Moderations- bzw. Mediationsmodelle dem Alternativmodell gegenüber oftmals überlegen im Sinne einer besseren Varianzaufklärung. Es ist möglich, dass eine Replikation der Studien mit größeren Stichproben die angenommenen Effekte bestätigen kann oder aber auch andere Effekte auffindet.

Positiv zu bewerten ist der Ansatz der Studie, der verschiedene Ansätze zur Erhöhung der prädiktiven Validität von Persönlichkeitseigenschaften in der beruflichen Eignungsdiagnostik miteinander verbindet. So wurde auf der Seite der Prädiktoren nicht nur Gewissenhaftigkeit allgemein evaluiert, sondern auf einer spezifischeren Ebene die Gewissenhaftigkeitsaspekte Fleiß und Ordnung hinsichtlich ihrer Vorhersagekraft überprüft. Zudem wurden diese mit Selbstwirksamkeit kombiniert, um im Sinne des interaktionistischen Ansatzes sowohl stabile als auch variable Prädiktoren miteinander zu kombinieren. Zudem wurde ein eher unbekanntes Konstrukt – die wahrgenommenen eigenen Fähigkeiten – in die Studie aufgenommen, das zuvor vielversprechende Zusammenhänge mit (schulischen) Leistungskriterien gezeigt hat (Noftle & Robins, 2007; Dreier, 2008). Wie von Funder (2006) gefordert, wurden Mediations- und Moderationsanalysen berechnet. Unabhängig von der mangelnden Teststärke der Untersuchung klärten die vier soeben aufgezählten Faktoren jedoch in allen berechneten Modellen vergleichsweise wenig Varianz auf. Es ist möglich, dass das postulierte Vorhersagemodell weiterhin eine zu starke Vereinfachung des Zusammenhangs darstellt und die Beziehung zwischen Persönlichkeit und Ausbildungsleistung noch komplexer ist als angenommen. Verschiedene Autoren postulieren, dass der Zusammenhang zwischen Arbeit und Persönlichkeit nicht einfach ist (Jackson et al., 2010; Lord, 2011). Ihnen zufolge ist Verhalten von verschiedenen Determinanten abhängig, zu denen neben Persönlichkeitseigenschaften im Sinne von Traits und erlernten Fähigkeiten auch Motive, Ziele, Interessen und situationale Faktoren zählen. Insofern sollte die vorliegende Studie als erster Schritt hin zu einem komplexeren Vorhersagemodell gewertet werden. Das Vorhersagemodell sollte nicht nur noch einmal in einer größeren Stichprobe evaluiert werden, sondern auch durch weitere Einflussvariablen ergänzt werden. Welche dies genau sein könnten, wird in Abschnitt 9.4.1 (Implikationen für die Forschung) näher erläutert.

Allgemein einschränkend ist zu den vorliegenden Überprüfungen der prognostischen Validität und des Vorhersagemodells anzumerken, dass lediglich lineare Zusammenhänge überprüft wurden (Sarges, 2010). Es ist denkbar, dass die unterschiedlichen unabhängigen und abhängigen Variablen auch auf eine andere Art und Weise zusammenhängen. Beispielsweise berichteten O'Connor und Paunonen (2007) in ihrer Metaanalyse von einem umgekehrt u-förmigen Zusammenhang zwischen Gewissenhaftigkeit und Schulleistung in dem Sinne, dass sowohl extrem niedrige als auch extrem hohe Ausprägungen von Gewissenhaftigkeit mit schlechteren schulischen Leistungen in der Schule assoziiert sind. Auch Minbashian und

Kollegen (2010) zeigten in ihrer quasiexperimentellen Studie, dass sehr rigide agierende Manager weniger in der Lage waren, sich an veränderte Umgebungsbedingungen anzupassen und daher weniger erfolgreich waren. Ab einem bestimmten Punkt ist eine hohe Gewissenhaftigkeit oder eine hohe Ausprägung im Ordnungsaspekt nicht mehr nur hilfreich, sondern kann auch hinderlich sein, wenn diese Verhaltenstendenz in Rigidität, Inflexibilität, Perfektionismus und Pedanterie umschwenkt (Le et al., 2011). Entsprechend sind diese Verhaltensäußerungen in der Klinischen Psychologie teilweise auch als Zwangsstörungen codiert. Personen, die Details zu viel Aufmerksamkeit schenken und sich ausschließlich mit diesen beschäftigen, verlieren den Überblick und somit die übergeordneten (Arbeits-) Ziele aus den Augen. Die meisten Arbeiten mit geringem Komplexitätsgrad beinhalten Aufgaben, die beides – Schnelligkeit und Genauigkeit – erfordern. Die (über-) vorsichtige und pflichterfüllende Natur (über-) hoch gewissenhafter Menschen führt jedoch dazu, dass diese Zeit verschwenden und somit eine geringere Arbeitsleistung aufweisen als andere. Das bedeutet, dass eher moderate Ausprägungen von Gewissenhaftigkeit erstrebenswert sind. Die Zerteilung von Gewissenhaftigkeit in die Aspekte Fleiß und Ordnung könnte dieses Verhältnis differenziert aufklären. Es ist möglich, dass auch in der vorliegenden Studie andere als lineare Zusammenhänge zwischen den Variablen bestanden haben, die jedoch nicht ermittelt wurden. Eine Überprüfung weiterer Möglichkeiten der Zusammenhänge der verschiedenen unabhängigen und abhängigen Variablen über lineare Modelle hinaus sollte im Rahmen zukünftiger Studien durchgeführt werden.

Darüber hinaus könnte eine Fehlerquelle in den Modellüberprüfungen aufgetreten sein, wenn der untersuchte Mediator zeitlich zu weit entfernt von der Prädiktorvariable liegt (Müller, 2007). Dies könnte für die wahrgenommenen eigenen Fähigkeiten der Fall gewesen sein, die erst zu T2 erhoben wurden, aber mit den Gewissenhaftigkeitsaspekten Fleiß und Ordnung sowie dem Selbstwirksamkeitskonstrukt zu T1 in Verbindung gesetzt wurden.

9.2.3 Vergleich der AGS mit einem allgemeinen Gewissenhaftigkeitsmaß

Der dritte Teil der Untersuchung widmete sich dem Vergleich der AGS mit einem allgemeinen Gewissenhaftigkeitsmaß. Um diese Fragestellung zu untersuchen, wurden zuerst die prädiktiven Validitäten der AGS mit der Gewissenhaftigkeitsdimension des NEO-FFI (1993, 2008) verglichen und anschließend die inkrementelle Validität der Gewissenhaftigkeitsaspekte Fleiß und Ordnung über die allgemeine Gewissenhaftigkeitsskala hinaus berechnet (9.2.3.1). Zudem wurden die intra- und interpersonellen Antwortvariabilitäten der unterschiedlichen Gewissenhaftigkeitsskalen als Indikatoren für die Nützlichkeit der Kontextualisierung ermittelt und miteinander verglichen (9.2.3.2).

9.2.3.1 Vergleiche zur Vorhersagekraft der verschiedenen Gewissenhaftigkeitsmaße

Hypothese 9 widmete sich dem Vergleich der Vorhersagekraft des Fleiß- und Ordnungsaspektes mit dem allgemeinen Gewissenhaftigkeitsmaß des NEO-FFI. Die Kriteriumskorrelationen beider Gewissenhaftigkeitsaspekte waren sowohl zu T1 als auch zu T2 deutlich geringer und meist nicht signifikant, sodass Hypothese 9 verworfen werden musste. Dies ist kongruent mit den Ergebnissen von Schmit et al. (1995), in deren Untersuchung die Gewissenhaftigkeitsdimension ebenfalls etwas höher mit dem Kriterium (GPA) korrelierte als die sechs Gewissenhaftigkeitsfacetten. Allerdings sind die Ergebnisse der vorliegenden Studie nur bedingt aussagekräftig, da die untersuchten Substichproben zu klein waren, wie bereits zuvor beschrieben (s. Abschnitte 8.3.2 und 9.2.2.2). So zeigten die Gewissenhaftigkeitsaspekte teilweise zwar hohe, aber nicht signifikante Zusammenhänge mit den verschiedenen Kriterien. Auch das allgemeine Gewissenhaftigkeitsmaß korrelierte nur in einem Fall (Vorgesetztenbeurteilung in der technischen Substichprobe) signifikant mit dem Kriterium. Ein Vergleich der Vorhersagekraft der Maße war unter diesen Umständen nur bedingt möglich. Eine ge-

meinsame Analyse kaufmännischer und technischer Auszubildender ergab teilweise deutlich höhere Zusammenhänge der Gewissenhaftigkeitsaspekte Fleiß und Ordnung mit den verschiedenen Kriterien (s. Appendix, Tabelle A-16). Die Ableitung von Schlussfolgerungen und Implikationen ist in Bezug auf die Fragestellung von Hypothese 9 daher eher nicht angezeigt.

Hypothese 10 nahm an, dass die AGS inkrementelle Validität gegenüber dem allgemeinen Gewissenhaftigkeitsmaß des NEO-FFI besitzen. Die Ergebnisse deuteten an, dass der Fleißaspekt in keiner der beiden untersuchten Substichproben der Auszubildenden zu einer signifikanten Erhöhung der Varianzaufklärung beitrug und somit keine inkrementelle Validität gegenüber dem allgemeinen Gewissenhaftigkeitsmaß besaß. Durch die Hinzunahme des Ordnungsaspekts ergab sich in beiden Auszubildendengruppen eine signifikante Erhöhung der Varianzaufklärung und inkrementelle Validität für das Kriterium der Vorgesetztenbeurteilung. Bezüglich der weiteren Kriterien ergaben sich allerdings keine signifikanten Ergebnisse. Die inkrementelle Validität der beiden Gewissenhaftigkeitsaspekte über die Gewissenhaftigkeitskala des NEO-FFI hinaus ergab sich entsprechend auch nur für das Kriterium der Vorgesetztenbeurteilung. Dieser Effekt ist scheinbar vor allem auf die Varianzerhöhung durch die Hinzunahme des Ordnungsaspektes verantwortlich. Auch diese wenig aussagekräftigen Ergebnisse sind vor allem auf die nicht ausreichende Teststärke in den beiden untersuchten Stichproben zurückzuführen. Da bereits die Kriteriumskorrelationen der verschiedenen Gewissenhaftigkeitsmaße kaum einmal signifikant wurden, war zu erwarten, dass auch die Berechnungen der inkrementellen Validitäten nicht signifikant würden. Auch hier wäre die Analyse einer größeren Stichprobe von Vorteil gewesen, um signifikante und somit aussagekräftige Ergebnisse zu erhalten.

Darüber hinaus muss angefügt werden, dass die inkrementelle Validität von Aspekten über Faktoren hinaus bisher noch nicht erforscht wurde. Lediglich einige Untersuchungen zur inkrementellen Validität von Facetten über Dimensionen hinaus liegen bis dato vor (O'Connor & Paunonen, 2007; O'Neill & Allen, 2007; Paunonen, 1998; Sutin & Costa, 2010). In diesen war die inkrementelle Validität mit durchschnittlich 5-7 % zusätzlich aufgeklärter Varianz allerdings nicht besonders hoch (O'Connor & Paunonen, 2007), sodass die in der vorliegenden Studie gefundenen Effekte des Ordnungsaspekts mit 17 % bzw. 24 % zusätzlicher Varianzaufklärung hervorzuheben sind. Weitere Forschungsarbeiten zur inkrementellen Validität der Gewissenhaftigkeitsaspekte Fleiß und Ordnung – auch über andere Gewissenhaftigkeitsmaße hinaus – sollten in größeren Stichproben und mit weiteren Kriteriumsmaßen durchgeführt werden.

Eine Alternativerklärung für die inkrementelle Validität des Ordnungsaspektes bzw. der AGS über das allgemeine Gewissenhaftigkeitsmaß hinaus kann allerdings auch die Kontextualisierung der Items der AGS sein, denn auch kontextualisierte Items weisen inkrementelle Validität gegenüber generischen Items auf (Pace & Brannick, 2010; Schmit et al., 1995). Um Klarheit darüber zu erlangen, welcher bzw. wie viel Anteil der Varianzaufklärung auf die Kontextualisierung der AGS zurückzuführen ist, sollte in einer zukünftigen Studie auch ein generisches Maß zur Messung von Fleiß und Ordnung eingesetzt werden (z.B. die Gewissenhaftigkeitsskalen der BFAS; DeYoung et al., 2007). Ein Vergleich der prädiktiven und inkrementellen Validitäten könnte Aufschluss über den Anteil der Varianzaufklärung der Kontextualisierung geben.

Wie bereits in Abschnitt 9.2.2.1 diskutiert wurde, ist es auch möglich, dass sich die inkongruente (arbeitsbezogene) Kontextualisierung der AGS-Items nachteilig auf die Kriteriumskorrelation der damit erfassten Gewissenhaftigkeitsaspekte Fleiß und Ordnung ausgewirkt hat. Falls dies der Fall ist, können die Ergebnisse zur inkrementellen Validität der AGS an dieser Stelle nicht interpretiert werden und es müsste davon ausgegangen werden, dass die Kriteriumskorrelationen sowie die inkrementelle Validität der Gewissenhaftigkeitsaspekte Fleiß und Ordnung deutlich über den gefundenen Ergebnissen dieser Studie liegen.

9.2.3.2 Vergleich der intra- und interpersonellen Antwortvariabilitäten

Die letzte Hypothese widmete sich dem Vergleich der intra- und interpersonellen Antwortvariabilitäten der AGS mit dem allgemeinen Gewissenhaftigkeitsmaß, um die Nützlichkeit der Kontextualisierung der Items erfassen zu können. Hierzu wurden die internen Konsistenzen und Streuungen der Skalen miteinander verglichen. Den Ergebnissen zufolge erwies sich der Fleißaspekt dem allgemeinen Gewissenhaftigkeitsmaß gegenüber als homogener im Sinne höherer Reliabilitäten und geringerer Streuungen. Die Ergebnisse den Ordnungsaspekt betreffend sind inkonsistent. Hier zeigten sich teilweise höhere, aber auch niedrigere interne Konsistenzen bzw. niedrigere und größere Streuungen im Vergleich mit dem allgemeinen Gewissenhaftigkeitsmaß.

Die Ergebnisse verfügen aufgrund der Analyse der Grundgesamtheit über eine ausreichende Power, um interpretiert werden zu können.

Die bestätigenden Ergebnisse der Fleißskala betreffend deuten darauf hin, dass die Kontextualisierung der Items zu einem intrapersonell homogeneren Antwortverhalten der getesteten Personen geführt hat. Das heißt, die einzelnen Personen haben in ihrem Antwortverhalten weniger Varianz gezeigt, was zu einer Verringerung der intrapersonellen Antwortvariabilität geführt hat. Unterschiede zwischen den Personen (interpersonelle Antwortvariabilität) sind aufgrund der Kontextualisierung ebenfalls geringer geworden, da alle untersuchten Probanden auf das gleiche Konzept zurückgriffen, um die Items zu beantworten.

Die inkonsistenten Ergebnisse in Bezug auf die Ordnungsskala können die Annahmen zur intra- und interpersonellen Antwortvariabilität weder bestätigt noch verworfen werden. Hier zeigten sich sowohl Zu- als auch Abnahmen der inter- und intrapersonellen Antwortvariabilität.

Allgemein spiegeln die hier gefundenen Ergebnisse die Befunde von Lievens et al. (2008) sowie Reddock et al. (2010) nur teilweise wider, das eine Kontextualisierung von Items zu einer Verringerung der intrapersonellen Antwortvariabilität führt. In zukünftigen Studien sollte überprüft werden, inwiefern die intrapersonelle Antwortvariabilität als eigenständige Eigenschaft gewertet werden kann und ggf. sogar prädiktive Validität hinsichtlich der Kriterien besitzt.

Allerdings sollte in einer solchen Untersuchung der Referenzrahmen des Prädiktors konzeptuell mit dem Kriterium übereinstimmen. Andererseits kann man trotz hoher Reliabilität eine nur niedrige Validität erreichen. Dies war in der vorliegenden Studie der Fall, auch wenn „Schule“ und „Arbeit“ den Autoren zufolge konzeptuell als überlappend zu betrachten sind. Zukünftige Forschungsarbeiten zur Überprüfung der AGS sollten sich daher weiter dem Vergleich der Skaleneigenschaften unter kongruenten Bedingungen widmen, um diesbezüglich eindeutige Aussagen treffen zu können.

Die inkonsistenten Ergebnisse bezüglich des Ordnungsaspektes könnten auf die teilweise noch fehlende Arbeitserfahrung der Ausbildungsbewerber zu T1 zurückzuführen sein. Es ist möglich, dass vor allem die Inhalte der Ordnungs-Items für Menschen ohne Berufserfahrung schwerer zu verstehen waren und es so zu inkonsistenterem Antwortverhalten gekommen ist.

Hierzu ist auch anzumerken, dass während der Testdurchführung in der Kontrollgruppe v.a. in den Hauptschulklassen einige Nachfragen zum inhaltlichen Verständnis einiger Items gestellt wurden. Es ist demzufolge anzunehmen, dass auch nicht alle Ausbildungsbewerber die Inhalte der Items vollständig erfasst haben, sich aber nicht getraut haben, dies in der Auswahl-situation nachzufragen. Dies könnte zu inkonsistenten Antworten geführt haben. Da in der Substichprobe der technischen Auszubildenden im Vergleich zu den kaufmännischen Auszubildenden deutlich mehr Bewerber auch mit geringerer schulischer Vorbildung enthalten waren, ist es möglich, dass dies vor allem die Ergebnisse in der technischen Subpopulation betroffen

haben könnte. Es ist denkbar, dass die Items der AGS trotz gründlicher Itemformulierung für Bewerber bzw. Schüler mit schlechten semantischen Kenntnissen der deutschen Sprache schwerer verständlich sind. Dies könnte noch einmal evaluiert werden. Das Problem der semantischen Verständlichkeit der Items könnte auch mit der Muttersprache der Probanden konfundiert sein, da in den Hauptschulklassen deutlich mehr Schüler mit Migrationshintergrund waren als in der Gymnasialklasse. Diesbezüglich muss kritisch angemerkt werden, dass im Rahmen der Überprüfung der Kontrollvariablen in der vorliegenden Studie nicht die Muttersprache der Probanden erhoben wurde. Es ist möglich, dass Versuchsteilnehmer, deren Muttersprache nicht Deutsch ist, Verständnisschwierigkeiten mit den Items der Fragebogen hatten.

Die technischen Bewerber verfügten zudem auch über weniger berufliche Vorerfahrungen als die kaufmännischen, was ebenfalls die inkonsistenten Ergebnisse erklären könnte.

Zukünftige Forschungsarbeiten sollten überprüfen, inwieweit es signifikante Unterschiede in der Beantwortung von allgemeinen und kontextualisierten Items in verschiedenen Subpopulationen gibt. So ist es möglich, dass berufserfahrene Bewerber arbeitsbezogene Items anders beantworten als Bewerber ohne Berufserfahrung. Auch könnte überprüft werden, ob Mitarbeiter einer Firma berufsbezogene Items konsistenter Beantworten als Bewerber, die noch nicht in dem Unternehmen tätig sind. Zudem ist es möglich, dass die Beantwortung der Items von der schulischen Vorbildung oder auch dem Alter abhängig ist. Auch hier sollten Untersuchungen überprüfen, inwiefern es diesbezüglich Gruppenunterschiede gibt.

Bing und Kollegen (2004) merkten kritisch an, dass auch das Beantworten eines neutral gehaltenen Fragebogens, wie hier der NEO-FFI (Borkenau & Ostendorf, 2008), unter Bewerber-Bedingungen Kontextualisierung genug sei. Dies könnte erklären, warum die Bewerberstichprobe – über die starke Vorselektion hinaus – auch auf der allgemeinen Gewissenhaftigkeitsdimension vergleichsweise geringe Streuungen zeigte.

9.2.4 Überprüfung der Kontrollvariablen

Bezüglich der Kontrollvariablen kann ein Einfluss der untersuchten Störvariablen größtenteils ausgeschlossen werden. Weder das Geschlecht oder das Alter der Probanden noch die Schulbildung oder das Fehlen bzw. Vorhandensein erster beruflicher Erfahrungen hatten den Überprüfungen zufolge einen Einfluss auf die Ergebnisse der vorliegenden Untersuchung.

Ein Einfluss sozial erwünschten Antwortverhaltens kann demgegenüber allerdings nicht ausgeschlossen werden. So zeigte die Skala zur Erfassung des sozial erwünschten Antwortverhaltens in allen untersuchten Stichproben signifikante Zusammenhänge mit den unabhängigen und abhängigen Variablen. Darüber hinaus wiesen die Auszubildenden in allen relevanten Eigenschaften deutlich höhere Selbstbeschreibungen auf als die Schüler in der Kontrollgruppe. Diese Effekte könnten einerseits durch die hohe Vorselektion der Bewerber, andererseits aber auch durch sozial erwünschtes Antwortverhalten erklärt werden. Die Tatsache, dass jedoch nicht alle Auszubildenden zu T2 ihre Selbsteinschätzungen verringerten, sondern auch erhöhten oder – in wenigen Fällen – stabil blieben, deutet an, dass weitere Einflüsse die Selbsteinschätzungen beeinflusst haben, die im Rahmen dieser Untersuchung allerdings nicht erfasst wurden. Insofern ist die vorliegende Untersuchung repräsentativ für den uneindeutigen Forschungsstand in Bezug auf sozial erwünschtes Antwortverhalten (Kanning, 2011; Lee et al., 2010).

Einschränkend muss hinzugefügt werden, dass das Maß zur Erfassung sozial erwünschter Antworttendenzen (SDS-17; Stöber, 2001) erst zu T2 erhoben wurde. Es ist wahrscheinlich, dass die Bewerber in dieser Situation keinen oder einen deutlich geringeren Druck verspürt haben, sich auf eine bestimmte Art und Weise zu präsentieren und weniger sozial erwünscht

antworteten als in der Bewerber-Bedingung (T1). Die vorliegenden Ergebnisse sollten daher unter Berücksichtigung dieser Einschränkung betrachtet und interpretiert werden.

Im Vorwege der Untersuchung wurde durch verschiedene Maßnahmen versucht, den Einfluss sozial erwünschten Antwortverhaltens soweit wie möglich zu minimieren. Ziegeler (2011) kam in einer vergleichbaren Studie zu dem Ergebnis, dass Faking im Rahmen der Testung von Ausbildungsbewerbern nicht auszuschließen ist. Der Zusammenhang von sozial erwünschtem Antwortverhalten und Facetten bzw. Aspekten ist aber noch wenig erforscht, so dass die Einordnung der vorliegenden Ergebnisse noch schwerer möglich ist. Zukünftige Forschungsarbeiten sollten sich der Identifikation der bedingenden Faktoren sowie eindeutigen Identifizierbarkeit sozial erwünschten Antwortverhaltens widmen. Der Einsatz von Fragebögen zur Identifikation von Fakern (z.B. SDS-17; Stoeber, 2001) und eine Berechnung von Korrelationen mit den anderen eingesetzten Testverfahren ist keine ausreichende Lösung dieses Problems (Pauls & Crost, 2004). Eine Analyse sozial erwünschten Antwortverhaltens stand jedoch nicht im Fokus der vorliegenden Arbeit.

Bis die genauen Bedingungen für das Auftreten von Faking nicht identifiziert sind und keine effektiven Maßnahmen zur eindeutigen Identifikation bzw. Elimination von Fakern existieren, ist sozial erwünschtes Antwortverhalten eine Störvariable, die in allen Situationen beruflicher Eignungsdiagnostik mehr oder weniger stark auftritt. Zwar kann sozial erwünschtes Antwortverhalten die Ergebnisse der vorliegenden Studie beeinflusst haben, aber dies entspricht den natürlichen Gegebenheiten in einer Feldstichprobe, die Personaler stets bedenken müssen. Kriteriumsvalidität liegt auch trotz Einfluss von Faking oftmals vor (Blickle et al., 2011). Vor diesem Hintergrund ist das Auftreten sozial erwünschten Antwortverhaltens eher ein Vorteil der vorliegenden Untersuchung, da diese dadurch an Generalisierbarkeit für die Praxis gewinnt.

Da in der vorliegenden Feldstichprobe eine vergleichsweise große Anzahl von Bewerbern eingestellt wurde, ist davon auszugehen, dass sozial erwünschtes Antwortverhalten im Sinne einer Rangfolgenvertauschung der Bewerber keinen Effekt auf die Auswahlentscheidung des Unternehmens hatte.

Darüber hinaus ist zudem ungeklärt, ob sozial erwünschtes Antwortverhalten nicht auch ein Indikator für eine höhere Intelligenz (Bing et al., 2004), eine bessere Anpassungsfähigkeit der Bewerber (Marcus, 2003; Viswesvaran et al., 2007) oder schlicht eine persönliche Neigung im Sinne eines Traits ist (Stöber, 2001). In diesem Falle wäre Faking nicht negativ zu bewerten, sondern könnte im Gegenteil sogar als weiteres Einstellungskriterium gelten. Aus diesem Grund ist es derzeit nicht angezeigt, „Faker“ pauschal auszusortieren.

Es ist somit allgemein wichtig, sozial erwünschtes Antwortverhalten bei der Planung und Durchführung einer eignungsdiagnostischen Untersuchung zu bedenken (s. hierzu Abschnitte 2.4.3 und 3.2.3) und bei der Auswertung als weitere Alternativerklärung der Testergebnisse zu berücksichtigen. Da es weiterhin keine gesicherten Erkenntnisse gibt, ob und in welchem Maße sozial erwünschtes Antwortverhalten im Einzelfall auftritt oder nicht, sollten in der eignungsdiagnostischen Praxis auch weiterhin Persönlichkeitstests zur Bewerberauswahl eingesetzt werden. Eine Möglichkeit, diese weitestgehend zu neutralisieren, bietet eine Normierung relevanter Testverfahren in Praxisstichproben. Forschung und Praxis sollten hierzu sowie bei der Identifikation der bedingenden Faktoren von Faking zukünftig enger kooperieren. Vor allem auch der Zusammenhang von Faking und Facetten bzw. Aspekten relevanter Persönlichkeitseigenschaften wie Gewissenhaftigkeit bzw. Fleiß und Ordnung sollte weiter erforscht werden. Auch die Aufnahme der Variable in Vorhersagemodelle ist angezeigt.

9.3 Zusammenfassende Bewertung der Ergebnisse

Im Folgenden soll ein kurzes Resümee zu den Ergebnissen dieser Studie gezogen werden.

Hinsichtlich der Skalvalidierung können die Ziele der Untersuchung als erreicht angesehen werden. Es kann davon ausgegangen werden, dass die AGS das zweifaktorielle latente Modell zur Erfassung der Gewissenhaftigkeitsaspekte Fleiß und Ordnung zuverlässig abbilden. Hinsichtlich der Stabilität wiesen die Skalen der AGS allerdings nur mäßige Retest-Reliabilitäten auf. Dies könnte auf das lange Zeitintervall zwischen den Erhebungen der Variablen zurückgeführt werden, währenddessen viele unkontrollierte Einflüsse und Ereignisse auf die Studienteilnehmer eingewirkt haben könnten.

Hinsichtlich ihrer Vorhersagekraft zeigten die arbeitsbezogen formulierten Gewissenhaftigkeitsaspekte Fleiß und Ordnung in den Feldstichproben keine signifikanten Zusammenhänge mit den Kriteriumsvariablen. Allerdings deuteten die Zusammenhänge an, dass die AGS ein Maß zur differenzierten Vorhersage verschiedener Leistungskriterien zur Erfassung des Ausbildungserfolges sein könnten. Studien mit größeren Stichproben könnten Aufschluss hierüber geben. In der knapp ausreichend großen Kontrollgruppe zeigten sich entsprechend die angenommenen signifikanten Zusammenhänge. Sowohl der Fleiß- als auch der Ordnungsaspekt sagten signifikant die Abschlussnoten der Schüler voraus. Ähnliche Ergebnisse ergaben sich für das Selbstwirksamkeitskonstrukt, das in den Feldstichproben nicht in der Lage war, die verschiedenen Leistungskriterien signifikant vorauszusagen, wohl aber in der Kontrollgruppe signifikant mit den Abschlussnoten der Schüler korrelierte.

Bezüglich der Vorhersagemodelle ergaben sich ebenfalls nur in der Kontrollgruppe einige signifikante Ergebnisse. Allgemein deuteten die Überprüfungen der Modelle an, dass sowohl Selbstwirksamkeit als auch die wahrgenommenen eigenen Fähigkeiten eher als Mediatoren denn als Moderatoren den Zusammenhang der Gewissenhaftigkeitsaspekte mit den abhängigen Variablen beeinflussten. Allerdings deuteten die geringen Varianzaufklärungen an, dass noch weitere Variablen die Zusammenhänge zwischen Prädiktoren und Kriterien beeinflussen. Diese zu identifizieren sollte das Ziel zukünftiger Forschungsarbeiten sein. Eine weitere Analyse der verschiedenen Zusammenhänge in Bezug auf nicht-lineare Zusammenhänge zwischen unabhängigen und abhängigen Variablen könnte darüber hinaus ebenfalls angedacht werden.

Letztlich deutete die Überprüfung der inkrementellen Validität der Gewissenhaftigkeitsaspekte Fleiß und Ordnung im Vergleich mit einer allgemeinen Gewissenhaftigkeitsskala an, dass vor allem der Ordnungsfaktor zusätzliche Varianz hinsichtlich des Ausbildungserfolges erklären kann. Zwar ist es möglich, dass der Berufsbezug der Skala und die damit einhergehende inhaltliche Übereinstimmung zwischen Prädiktor und Kriterium zu diesem Ergebnis geführt hat. Da sich dieser Effekt allerdings nicht für den Fleißaspekt zeigte, ist eher davon auszugehen, dass dies nicht der Fall ist. Die Ergebnisse zur intrapersonellen Variabilität deuten zudem an, dass die Kontextualisierung der Items zu einer Homogenisierung des Antwortverhaltens zwischen und innerhalb der Probanden geführt hat. Die Annahme, dass die Probanden durch die Kontextualisierung auf ein einheitliches Schema zur Beantwortung der Items zurückgriffen, kann somit als bestätigt angesehen werden.

Die Überprüfung der Kontrollvariablen deutete auf einen Einfluss sozial erwünschten Antwortverhaltens hin. Alter, Geschlecht, Schulbildung sowie das Vorhandensein von ersten beruflichen Erfahrungen hatten keinen Einfluss auf die Ergebnisse der vorliegenden Studie.

Der größte Nachteil der vorliegenden Studie ergibt sich aus der geringen Teststärke der Feldstichprobe. Aufgrund der geringen Zellbesetzungen in den kaufmännischen und technischen Substichproben wurden viele Effekte nicht signifikant. Zudem konnten viele Störfaktoren im

Feldexperiment nicht kontrolliert oder identifiziert werden. Die geringe interne Validität wurde jedoch für eine hohe externe Validität in Kauf genommen.

So zeichnet sich diese Studie gegenüber vielen anderen, die ebenfalls den Zusammenhang zwischen Persönlichkeit bzw. Gewissenhaftigkeit und Leistungsvariablen untersucht haben, qualitativ hinsichtlich der Generalisierbarkeit ihrer Ergebnisse aus. Die bestehende Forschungsliteratur wird weiterhin vor allem durch Laborstudien dominiert, deren Ergebnisse nur bedingt auf die eignungsdiagnostische Praxis übertragbar sind. Die vorliegende Untersuchung überprüfte die theoretischen Annahmen und Erkenntnisse dieser Laborstudien in einer Feldstudie und trug damit zu einer Annäherung zwischen Forschung und Praxis bei. Der wesentliche Vorteil der Studie liegt nicht nur aus methodischer Sicht darin, dass reale Kandidaten in einer echten Bewerbungssituation mit verschiedenen Testverfahren und deren Zusammenhang mit tatsächlichen Leistungsvariablen überprüft wurde. Infolgedessen sind die Ergebnisse in hohem Maße praxisrelevant.

In Bezug auf die Stichprobe und die untersuchten Kriteriumsvariablen ist allerdings anzumerken, dass die Ergebnisse nur auf Deutschland übertragbar sind. Dies ist mit der Einzigartigkeit des deutschen Berufsausbildungssystems erklärbar, das nach einer dreigliedrigen Schullaufbahn eine kombinierte Berufs- und Schulausbildung aufweist. Da mit den AGS allerdings auch ein deutschsprachiger Fragebogen überprüft wurde, wiegt dieses Argument nicht so stark. Im Gegenteil – es ist von Vorteil, dass für dieses spezielle Ausbildungssystem praxisrelevante Ergebnisse gefunden wurden, die in einer Vielzahl von Ausbildungsberufen bzw. der Auswahl von Auszubildenden für diese angewendet werden können.

9.4 Ausblick und Implikationen

Die folgenden Abschnitte enthalten die Implikationen aus der vorliegenden Studie unterteilt für die Forschung (Abschnitt 9.4.1) und Praxis (Abschnitt 9.4.2). Einige Ableitungen, vor allem für die Forschung, wurden bereits in den vorherigen Abschnitten, besonders in der Darstellung der Ergebnisse, gegeben. Die folgenden beiden Abschnitte beschäftigen sich vor allem mit den Fragestellungen der Faktorenstruktur der Gewissenhaftigkeitsdimension, arbeitsbezogenem Testen sowie der Modellbildung zur Vorhersage von Ausbildungserfolg. Abschnitt 9.4.3 enthält ein Fazit zur vorliegenden Studie.

9.4.1 Implikationen für die Forschung

Ausgehend von der Interpretation der Ergebnisse lassen sich weitere Fragestellungen ableiten, die an die vorliegende Untersuchung anknüpfen.

Die Konstruktvalidität der AGS kann den vorliegenden Ergebnissen zufolge als gesichert angesehen werden. Das Instrument erfasst zuverlässig die arbeitsbezogenen Gewissenhaftigkeitsaspekte Fleiß und Ordnung. Es sollte darüber nachgedacht werden, die AGS weiter zu entwickeln, beispielsweise mittels der Erstellung von Fremdeinschätzungsversionen sowie Übersetzungen der Skala in andere Sprachen. Nach der Übersetzung der Skala sollte folglich auch die kulturelle Übertragbarkeit der AGS überprüft werden. Erste Ansätze sind hierzu bereits vorhanden (Dreier, 2008).

Entgegen der Annahmen erwies sich die Retest-Reliabilität der Skala allerdings nicht als besonders hoch. Diese sollte noch einmal in einer anderen Studie überprüft werden. Als Alternativerklärung für dieses Resultat könnte auch in Folgeuntersuchungen die Möglichkeit sich verändernder instabiler Konstrukte überprüft werden. Dies würde allerdings den Annahmen der klassischen Traittheorie widersprechen (Schuler & Höft, 2006).

Die mäßigen Korrelationen der AGS zwischen T1 und T2 waren in der vorliegenden Studie vergleichbar zur Retest-Reliabilität des Selbstwirksamkeitskonstruktes. Es ist denkbar, dass die Selbsteinschätzungen der Aspekte Fleiß und Ordnung sich ebenfalls verändern, vergleichbar mit dem Selbstwirksamkeitskonstrukt.

Selbstwirksamkeit ist ein variables Konstrukt, das sich laut sozial-kognitiver Theorie basierend auf Leistungs- und Stellvertretererfahrungen, mentalen Vorstellungen, verbalen Überzeugungen sowie physiologischen und emotionalen Zuständen beständig weiter entwickelt. Auch in Anbetracht der Tatsache, dass die Modellannahmen der vorliegenden Untersuchung nicht bestätigt werden konnten, sollten zukünftige Forschungsarbeiten somit auch Einflüsse zu identifizieren und kontrollieren bzw. eliminieren versuchen, die diese Veränderungen der Einschätzungen von T1 zu T2 hervorgerufen haben könnten. Die Einflussvariablen der sozial-kognitiven Lerntheorie könnten hierfür eine Ausgangsbasis bilden. Denkbar ist aber auch der Einfluss weiterer Eigenschaften, wie z.B. situationale Einflüsse (States), wie motivationale Einflüsse während einer Testung. Längsschnittliche Untersuchungen von Arbeitnehmern könnten Aufschluss darüber geben, welche Faktoren bei der Arbeit die Selbstwirksamkeit einer Person positiv bzw. negativ beeinflussen. Ein erster Ansatz hierzu wurde bereits von Minbashian und Kollegen (2010) getätigt, die eine Anpassungsform der Gewissenhaftigkeit („momentary conscientiousness“) hinsichtlich ihrer prädiktiven Validität untersucht haben.

Neben den zuvor geschilderten unidirektionalen Ansätzen zum Zusammenhang von kontextbezogener Persönlichkeit und schulischer bzw. beruflicher Leistung gibt es aber auch Überlegungen, dass Persönlichkeitseigenschaften und das Innehaben einer Tätigkeit sich gegenseitig beeinflussen könnten. Viswesvaran und Kollegen (2007) bezeichneten dies als *Kriteriumsdynamik*. Es gibt entsprechend auch erste Hinweise, denen zufolge Gewissenhaftigkeit sich auch dynamisch in dem Maße anpasst, in dem Aufgaben mehr oder weniger gewissenhaftes Verhalten von einem Mitarbeiter erfordert (Minbashian et al., 2010).

Auch ein reziprokes Verhältnis zwischen Leistung und den Gewissenhaftigkeitsaspekten ist denkbar (Bipp 2010; DeRaad & Schouwenburg, 1995), in dem Sinne, dass die Belohnungen, die Personen für gute Leistungen erhalten, das Verhalten, das den Erfolg hervorruft (= Fleiß und Ordnung), indirekt verstärkt und auf diese Weise vermehrt auftritt. Diese Überlegung könnte auch den Anstieg der Gewissenhaftigkeit über das Leben hinweg erklären, vor allem die gefundenen Zunahmen auf den Fleiß-Facetten und die eher gleichbleibenden Niveaus der Ordnungs-Facetten. Auch die leichte Abnahme der Gewissenhaftigkeit im hohen Alter könnte durch das (dann ausbleibende) reziproke Verhältnis zwischen Leistung und Gewissenhaftigkeit erklärt werden.

Darüber hinaus ist es möglich, dass die Effekte der Gewissenhaftigkeitsaspekte sich erst nach einiger Zeit im Beruf manifestieren, da die starke Strukturierung des Arbeitsalltags der neuen Auszubildenden noch verhindert, dass sich die Gewissenhaftigkeitsaspekte im Verhalten der Auszubildenden bereits zu Beginn zeigen. Zu ähnlichen Ergebnissen kamen Sutherland und Kollegen (2007) in einer Feldstudie an südafrikanischen Kundendiensttechnikern in einer sehr stark strukturierten Arbeitsumgebung. Auch Lord (2011) schrieb, dass Situationen bei der Arbeit Druck auf das Verhalten einer Person ausüben und folglich das Zeigen relevanter Eigenschaften fördern oder hemmen können. Die Retestung der vorliegenden Studie wurde zu Beginn der Lehrtätigkeit durchgeführt. Es ist möglich, dass eine erneute Testung der mittlerweile fest im Beruf stehenden Auszubildenden zu höheren Retest-Reliabilitäten führt. Zukünftige Untersuchungen sollten überprüfen, inwiefern der Strukturierungsgrad von Arbeitsumgebungen Auswirkungen auf das Verhalten bzw. das Ausdrücken der individuellen Persönlichkeit einer Person hat. Besonderes Augenmerk sollte auf die Analyse des Verhaltens neu eingestellter Mitarbeiter und Auszubildender im Vergleich mit bereits länger im Unternehmen tätigen Personen gelegt werden. Dies ist auch praktisch sehr relevant, beispielsweise für die Bewertung von Mitarbeitern während und kurz nach der Probezeit.

Hinsichtlich der Vorhersagekraft der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie der Selbstwirksamkeit ergaben sich lediglich in der Kontrollgruppe signifikante Effekte, die in ihrer Höhe den Zusammenhängen entsprechen, die in den relevanten Metaanalysen berichtet werden (z.B. Barrick et al., 2001; Poropat, 2009; O'Connor & Paunonen, 2007; Schmidt & Hunter, 1998). In den Stichproben der kaufmännischen und technischen Bewerber ergaben sich Hinweise auf Zusammenhänge, die allerdings nicht signifikant wurden. Es deutete sich an, dass die AGS differenzielle Vorhersagekraft besitzen – abhängig von der untersuchten Stichprobe und dem Kriterium. Zukünftige Forschungsarbeiten sollten die prädiktive Validität der AGS in größeren Stichproben untersuchen und hierbei verschiedene Stichproben und Kriterien miteinander vergleichen. Ein genaueres Verständnis der Einflüsse der beiden Aspekte kann zu einem besseren Verständnis der Zusammenhänge zwischen Persönlichkeit und Erfolg beitragen (De Fruyt & Salgado, 2003) und bisher maskierte Effekte offenlegen (Van der Zee et al., 2003).

Es sollten auch ältere Bewerber sowie Mitarbeiter in Personalentwicklungsmaßnahmen untersucht werden, um die Vorhersagekraft der Gewissenhaftigkeitsaspekte auf weitere eignungsdiagnostische Fragestellungen generalisieren zu können.

Insbesondere sollte aber das Vorhersagemodell weiter überprüft und entwickelt werden. Zum einen sollten die Moderations- und Mediationsmodelle in größeren Stichproben noch einmal hinsichtlich ihrer Gültigkeit überprüft werden. Zum anderen sollte das postulierte Modell um weitere Einflussvariablen ergänzt werden, da die Gewissenhaftigkeitsaspekte gemeinsam mit dem Selbstwirksamkeitskonstrukt bzw. den wahrgenommenen eigenen Fähigkeiten nicht genügend Varianz aufklären konnten. Hierzu könnten wie bereits oben beschrieben die Einflussvariablen der sozial-kognitiven Theorie herangezogen werden, z.B. individuelle Leistungserfahrungen, Stellvertretererfahrungen etc. abgefragt werden. Auch die Hinzunahme weiterer Persönlichkeitskonstrukte aus dem FFM, wie z.B. Ausgeglichenheit oder Offenheit für Erfahrungen könnten zur Varianzaufklärung beitragen und als weitere Variablen in einem komplexeren Vorhersagemodell fungieren.

Das Konstrukt der wahrgenommenen eigenen Fähigkeiten (Noftle & Robins, 2007) erwies sich für einige der untersuchten Kriterien als prädiktiv und wies sich tendenziell als guter Moderator bzw. Mediator der Zusammenhänge auf. Die wahrgenommenen eigenen Fähigkeiten sollten in Zukunft weiter beforscht werden. Sie scheinen vor allem für die Vorhersage von mündlichen bzw. praktischen Leistungen gut geeignet zu sein und zeigten mittlere Zusammenhänge mit den Persönlichkeitsvariablen Fleiß, Ordnung und Selbstwirksamkeit. Im Vergleich zu letzterer zeigten die wahrgenommenen eigenen Fähigkeiten sogar meist höhere Vorhersagekennwerte. Hierzu muss einschränkend angefügt werden, dass die wahrgenommenen eigenen Fähigkeiten jedoch erst zu T2 erhoben wurden und nicht bereits am Testtag mit erhoben wurden. Es sollte überprüft werden, inwiefern das Konstrukt bei gleichzeitiger Erhebung mit den weiteren prädiktiven Variablen eventuell besser geeignet ist, Leistung vorherzusagen und das Selbstwirksamkeitskonstrukt in der Praxis der beruflichen Eignungsdiagnostik zu ergänzen oder gar zu ersetzen. In einer Studie von Cury, Sarrazin und Famose (1997) deutete sich an, dass nur sehr selbstfokussierte Personen sich mit anderen verglichen. Aufgabenorientierte Personen interessierten sich nicht dafür, welche Leistungen andere erbracht hatten. Es ist daher möglich, dass sich selbstzentrierte und aufgabenfokussierte Personen in ihrer Ausprägung der wahrgenommenen eigenen Fähigkeiten unterscheiden. Bei der weiteren Überprüfung des Konstrukts sollte dies berücksichtigt werden.

Auch weitere Persönlichkeitseigenschaften bzw. deren Interaktion mit den Gewissenhaftigkeitsdimensionen könnten untersucht werden. So beschrieb bereits Lord (2011), dass Gewissenhaftigkeit in Kombination mit Offenheit das Lernverhalten einer Person beschreibe. Gerade bei der Vorhersage von Ausbildungserfolg könnte diese Kombination interessant sein, zumal auch Schmidt und Hunter (1998) in ihrer Metaanalyse die Offenheitsdimension für die

Vorhersage von Ausbildungserfolg als bedeutsam identifizierten. Auch Bell und Kozlowski (2002) berichteten, dass hohe Ausprägungen von Lernorientierung als „Puffer“ dient, der Schüler angesichts schlechter Leistungen vor der Selbstabwertung bewahrt. Konkret bedeutet dies, dass das Selbstwirksamkeitsempfinden aufgrund der hohen Lernmotivation auch angesichts schlechter Leistungen nicht abgewertet wird.

Darüber hinaus sollten auch andere Persönlichkeitsfaktoren als weitere Einflüsse der Vorhersagekraft berücksichtigt werden (Rode et al., 2008). Beispielsweise könnte die Kombination von Gewissenhaftigkeit und Neurotizismus bzw. Ausgeglichenheit den Einfluss negativer Emotionen auf die Zielerreichung erklären (Lord, 2011). Ein Zusammenhang zwischen dem Fleißaspekt und Neurotizismus könnte beispielsweise Auskunft über die Wirkweise störender Emotionen oder physiologischer Zustände („Stress“) auf die (berufliche) Zielorientierung einer Person geben (Rode et al., 2008). Eine Verbindung mit dem Motivationskonstrukt ist in diesem Zusammenhang ebenfalls denkbar. Auch die Anpassungsfähigkeit eines Individuums könnte eine bedeutende Einflussvariable bei der Vorhersage von beruflicher bzw. schulischer Leistung sein (Minbashian et al., 2010; Moutafi et al., 2004). Allerdings sollten grundsätzlich nur nachweislich für die berufliche Eignungsdiagnostik nützliche und relevante Kriterien erforscht werden (Hülshager & Maier, 2008) und auch Störvariablen der Zusammenhänge weiter identifiziert werden. Bezüglich des letztgenannten Punktes sollte vor allem der Einfluss von sozial erwünschtem Antwortverhalten weiter erforscht werden. Aber auch die Veränderbarkeit der Gewissenhaftigkeitsdimensionen über die Zeit hinweg sollte als weitere Konfundierende in Betracht gezogen und längsschnittlich untersucht werden (Jackson et al., 2009).

Neben der Untersuchung des Zusammenhanges der Fleiß- und Ordnungsaspekte mit weiteren Persönlichkeitskonstrukten sollte auch der Beziehung zu allgemeinen kognitiven Fähigkeiten wissenschaftliche Beachtung geschenkt werden. Schmidt und Hunter (1998) waren die ersten, die die inkrementelle Validität von Gewissenhaftigkeit über allgemeine kognitive Fähigkeiten hinaus metaanalytisch belegten. Eine Überprüfung der inkrementellen Validität der Fleiß- und Ordnungsaspekte sollte diesbezüglich vorgenommen werden. Zudem identifizierten Moutafi et al. (2004) eine hohe Gewissenhaftigkeit bei weniger kognitiv begabten Kindern als Copingmechanismus, um gute Schulleistungen zu erreichen. Durch erhöhtes Lernen und das regelmäßige Erledigen der Hausaufgaben erzielten ihren Ergebnissen zufolge auch weniger intelligente Kinder gute Noten. Die Autoren bezeichneten Gewissenhaftigkeit entsprechend als adaptive Eigenschaft, in dem Sinne, dass man sich bzw. seinen Arbeitsaufwand der Umgebung bzw. den Anforderungen anpasst. Dies spiegelt auch die Ergebnisse von Minbashian et al. (2010) wider, die Gewissenhaftigkeit im Arbeitskontext ebenfalls als adaptiv erfassten. In beiden Studien wird allerdings zwischen aufgabenbezogener bzw. momentaner Gewissenhaftigkeit und Gewissenhaftigkeit als Trait unterschieden. Es ist fraglich, ob sich diese Ergebnisse auch auf die Berufsausbildung übertragen lassen. Ansatzweise könnte man dies versuchen, mit den vorliegenden Daten zu überprüfen. Sollten sich die Ergebnisse bestätigen lassen, dass Fleiß und/ oder Ordnung den Zusammenhang zwischen kognitiven Fähigkeiten und schulischer Anstrengung bzw. Leistung beeinflussen, würde dies den diagnostischen Nutzen der Gewissenhaftigkeitsaspekte für die Auswahl von Auszubildenden noch weiter erhöhen. Ein solcher Befund würde auch die theoretischen Annahmen von Andresen (eingereicht) bestätigen, dass der Wille, etwas zu leisten, einen Einfluss auf den Erfolg eines Individuums hat.

Der Arbeitsbezug der verwendeten Skalen zur Erfassung der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie der Selbstwirksamkeit sollte in dieser Untersuchung zur Erhöhung der Kriteriumsvalidität und Reduktion der Messfehler beitragen (Bing et al., 2004; Hunthausen et al., 2003; Lievens et al., 2008; Reddock et al., 2010, 2011; Schmit et al., 1995). Allerdings wurde lediglich in der Bewerberstichprobe auch ein generisches Gewissenhaftigkeitsmaß eingesetzt. Frühere Studien konnten die Überlegenheit der AGS gegenüber einem

neutralen Maß zur Erfassung der beiden Gewissenhaftigkeitsaspekte Fleiß und Ordnung (BFAS; DeYoung et al., 2007) zwar belegen (Dreier, 2008; Struckmeier, 2009), allerdings handelte es sich bei den dort verwendeten AGS um eine Vorversion der in dieser Studie verwendeten Skala. Außerdem fehlt ein Vergleich der arbeitsbezogenen Selbstwirksamkeitsskala mit einem neutralen Selbstwirksamkeitsmaß. Auf Basis der vorliegenden Ergebnisse kann nicht eindeutig belegt werden, inwiefern der Arbeitsbezug der verwendeten Skalen zu einer Verbesserung der Kriteriumsvalidität geführt hat. Dies sollte in einer gesonderten Untersuchung nachgeholt werden.

Die Überprüfung der inkrementellen Validität der AGS deutete eine Überlegenheit des Ordnungsaspekts gegenüber der allgemeinen Gewissenhaftigkeitsdimension an. Allerdings ist nicht nachvollziehbar, welcher Anteil der zusätzlich aufgeklärten Varianz des Ordnungsaspektes auf den Arbeitsbezug der Skala oder die Spezifität des Aspektes zurückzuführen ist. Auch die AGS wiesen gegenüber dem allgemeinen Gewissenhaftigkeitsmaß inkrementelle Validität auf und spiegeln damit Ergebnisse anderer Untersuchungen wider, die ebenfalls die Überlegenheit der Facetten gegenüber den allgemeinen Dimensionen aufzeigten (O'Connor & Paunonen, 2007; O'Neill & Allen, 2007; Sutin & Costa, 2010). Zukünftige Untersuchungen sollten die inkrementelle Validität und deren bedingende Faktoren systematisch untersuchen, um Aussagen darüber treffen zu können, inwiefern der Arbeitsbezug einer Skala zu einer (zusätzlichen) Erhöhung der Varianzaufklärung beiträgt (Pace & Brannick, 2010).

Allgemein deuteten die Ergebnisse zur intrapersonellen Variabilität an, dass der Berufsbezug der Skala zu einer homogenen Antwortweise der Probanden führte. Dies kann als Beleg für die Nützlichkeit der Kontextualisierung der Items angesehen werden. Allerdings erfolgte zu keinem Zeitpunkt der vorliegenden Untersuchung eine explizite Überprüfung der Wirksamkeit des Referenzrahmens. Es bleibt unklar, ob die Schüler bzw. Berufsschüler die AGS und Selbstwirksamkeitsskala mit einem Arbeits- oder Schulbezug beantworteten. Es ist zudem möglich, dass die hier verwendeten kontextualisierten Items immer noch zu allgemein waren und zur Verbesserung der Kriteriumsvalidität noch spezifischere, ggf. tätigkeitsbezogen formulierte Items notwendig sind (Schmit et al., 1995). Zukünftige Untersuchungen sollten sich der Überprüfung der verwendeten Referenzrahmen bei den Probanden widmen, um Aussagen darüber treffen zu können, inwiefern die Kontextualisierung eines Fragebogens hilfreich ist und wirklich zu homogenerem Antwortverhalten führt.

Zudem war der Arbeitskontext der AGS und des Selbstwirksamkeitsmaßes nicht konsistent zum Referenzrahmen der meisten Kriterien, die eher schulbezogen waren. Der falsche Kontext könnte sich in der vorliegenden Studie auch nachteilig ausgewirkt und sogar zu verringerten Kriteriumsvaliditäten geführt haben (Lievens et al., 2008). Lediglich bezüglich der Vorgesetztenbeurteilung stimmte der Kontext der arbeitsbezogenen Skalen inhaltlich mit dem Kriterium überein. Hier ergaben sich tendenziell auch die höchsten Zusammenhänge. In zukünftigen Forschungsarbeiten sollte darauf geachtet werden, dass Fragebogen und Kriterium inhaltlich kongruent sind. Auch sollte der Kontext der Items systematisch variiert werden, um den Einfluss einer falschen Kontextualisierung besser abschätzen und vor diesem Hintergrund die Ergebnisse dieser Studie neu bewerten zu können.

Darüber hinaus könnte der Situationsdruck variiert werden, um den Einfluss der Testsituation erfassen zu können. Laut Lord (2011) kann man zwischen „starken“ und „schwachen“ Situationen unterscheiden. In dieser Untersuchung stellte das Bewerbungsverfahren eine starke Situation dar, die die persönliche Varianz im Antwortverhalten eingeschränkt haben könnte. Die Testsituation in der Kontrollgruppe ist demgegenüber als schwach zu bezeichnen, da die Antworten der Schüler keinen Einfluss auf ihren Schulabschluss o.ä. hatten. Dies könnte zu mehr Varianz im Antwortverhalten der Schüler geführt haben. Der Situationsdruck ist demzufolge eine mögliche Erklärung für die größere Datenvarianz in der Kontrollgruppe. Zukünftige Studien sollten überprüfen, inwiefern der Situationsdruck in Auswahl-situationen zu ho-

mogenerem Antwortverhalten zwischen den Probanden führt. Dies ist ein weiterer Faktor, der zu einer erhöhten Varianzeinschränkung führt. Auch in dieser Studie könnte dies die Aussagekraft der Ergebnisse der Bewerberstichprobe eingeschränkt haben.

Der Nutzen der Kontextualisierung sollte zusätzlich bei Personalern bzw. Bewerbern untersucht werden. Hierzu ist denkbar, Bewerber generische sowie kontextualisierte Testverfahren bearbeiten zu lassen und hinterher bezüglich ihrer subjektiven Nützlichkeit für die Vorhersage von Ausbildungs- und Arbeitsleistung zu befragen. Auch die Personalern könnten hinsichtlich des empfundenen Mehrwertes durch den Einsatz der arbeitsbezogener Skalen befragt werden. Diese Untersuchung könnte auch die Wahrnehmung dieser Testverfahren in der eignungsdiagnostischen Praxis verändern und die Lücke zwischen Wissenschaft und Praxis zu schließen helfen.

Allgemein sollte man beim Vergleich generischer und kontextualisierter Verfahren allerdings darauf achten, dass keine „carry-over-Effekte“ auftreten, die zur Folge haben, dass Probanden den Kontext eines ersten Fragebogens erinnern und mit diesem auch einen zweiten bzw. folgende Testverfahren beantworten (Bing et al., 2004). Da in der vorliegenden Studie immer zuerst der neutrale Fragebogen administriert wurde, ist das Vorliegen eines solchen Effektes in dieser Untersuchung allerdings auszuschließen.

Insgesamt könnte auch über eine Überprüfung der Verständlichkeit der AGS nachgedacht werden, da – wie bereits in Abschnitt 9.2.3.2 beschrieben – vor allem Schüler mit geringerem Bildungsgrad Schwierigkeiten bei der Itembeantwortung der AGS äußerten. DuVernet et al. (2010) besagen, dass der Antwortprozess auf Fragebogenitems in fünf Stufen unterteilt werden kann. Im ersten Schritt erfolgt die semantische Interpretation der Wörter, was die Autoren als literarisches Verstehen der Items bezeichnen. Der zweite Schritt ist das Verstehen des Inhaltes und die Interpretation des Items. Im dritten Schritt sucht der Proband das Gedächtnis nach relevanten Informationen oder Verhaltensweisen ab, die zur Beantwortung des Items geeignet sind. Der vierte Schritt ist die Antwortformulierung. Hierzu müssen die aus dem Gedächtnis abgerufenen Informationen nach Relevanz gewichtet, integriert und verglichen werden. Der letzte Schritt ist dann die Anpassung der Antwort auf das Antwortformat. Dies kann dadurch beeinflusst werden, dass jeder Mensch danach strebt, möglichst konsistent zu erscheinen und seine Antworten tendenziell aneinander anpasst. Den Autoren zufolge ist jeder dieser Schritte im Antwortprozess von der allgemeinen kognitiven Leistungsfähigkeit des Beantwortenden abhängig. Wenn jemand also kognitiv nicht in der Lage ist, eine oder mehrere Stufen im Antwortprozess zu bewältigen, kann er den Fragebogen nicht korrekt beantworten und wird ggf. frustriert und demotiviert. Zwar wurde bei der Generierung der Items zwar darauf geachtet, möglichst kurze und leicht verständliche Aussagen zu formulieren, dennoch ist es möglich, dass es bei der Itembeantwortung individuell zu Problemen gekommen ist. In einer zukünftigen Arbeit könnten daher noch einmal die Items hinsichtlich ihrer Verständlichkeit vor allem in geringer gebildeten Probandengruppen überprüft werden. Hierzu könnten beispielsweise Vergleiche der Itemschwierigkeiten in Versuchsgruppen mit unterschiedlichem Bildungsgrad herangezogen werden.

Abschließend ist allgemein kritisch zur Forschung der differenziellen und diagnostischen Psychologie anzumerken, dass die meisten Untersuchungen bzw. deren Ergebnisse nicht die Entscheider in der Praxis der beruflichen Eignungsdiagnostik erreichen (Hülshager & Maier, 2008; König et al., 2010; Ortner et al., 2007) und nur wenige Praxisstudien durchgeführt werden (Schuler et al., 2007). Nach wie vor herrscht eine große Diskrepanz zwischen wissenschaftlichen Erkenntnissen zu und praktischer Durchführung von eignungsdiagnostischen Verfahren, was bereits Schmidt und Hunter (1998) in ihrer Metaanalyse kritisierten. Dies liegt vor allem daran, dass relevante Erkenntnisse nicht nur weitestgehend in Laborstudien durchgeführt werden, sondern meist auch lediglich in Journals veröffentlicht werden, die fast ausschließlich von Studenten und Forschern gelesen werden oder auf Fachkongressen „intern“

den ausschließlich forschenden Kollegen vorgestellt werden. Gerade die deutsche Forschung krankt daran, dass vor allem in englischer Sprache veröffentlicht wird. Der Einfluss der deutschen Forschung auf den internationalen Markt ist vergleichsweise gering (Hülshager & Maier, 2008). So werden auch seitens der Forschung oftmals lediglich US-Verfahren übersetzt und für den deutschen Sprachraum angepasst und verwendet. Prominentestes Beispiel hierfür ist der NEO-PI-R bzw. NEO-FFI. Hülshager und Maier (2008) forderten, dass die deutsche Forschung sichtbarer werden müsse. Die vorliegende Studie bietet einen ersten Schritt in diese Richtung, zumal die AGS in einer Vorversion auch bereits schon in englischer Sprache vorliegen (Dreier, 2008). Auch im Hinblick auf die zunehmende Internationalisierung vieler Unternehmen sollten Verfahren sprachlich sowie kulturell angepasst werden (Kersting, 2011).

Selbst wenn die Praktiker Interesse an neueren wissenschaftlichen Erkenntnissen haben, scheitert die Umsetzung dieser in die Praxis zudem oftmals schlicht an der fehlenden Zeit seitens der Personaler, sich intensiv mit den relevanten wissenschaftlichen Arbeiten auseinander zu setzen (Nachtwei & Schermuly, 2009). Auch ein mangelnder Zugang zu psychologischer Fachliteratur kann eine Ursache für die fehlende Umsetzung wissenschaftlicher Erkenntnisse in der Praxis sein. So sollten Personaler über die Ergebnisse der Nützlichkeitsstudien der arbeitsbezogenen Gewissenhaftigkeitsaspekte und Selbstwirksamkeit bei der Vorhersage von Ausbildungs- und Arbeitsleistung informiert werden. Diese Maßnahme sollte dazu beitragen, die Vorurteile gegenüber Persönlichkeitstests bzw. deren Nützlichkeit abzubauen und die Akzeptanz sowie die Verwendung dieser Instrumente in Deutschland zu erhöhen. Denn wie bereits zu Beginn dieser Arbeit beschrieben, ist die Persönlichkeit der Bewerber bzw. sind die sog. „Softskills“ das Kriterium in Deutschland, das die meisten Unternehmen (88 %) als am meisten relevant für ihre Einstellungsentscheidungen empfinden (Backhaus, 2004a; CRF Deutschland, 2009) und immer beliebter werden (Anderson & Ones, 2003; DuVernet et al., 2010; Rammstedt & John, 2005). Auch seitens der Forschung werden Persönlichkeitstests als nützliche Ergänzung des Auswahlprozesses empfohlen (Lord, 2011; Risavy & Hausdorf, 2011). Dennoch werden Persönlichkeitstests weiterhin nur von wenigen Unternehmen (32 %) auch in der Praxis verwendet (Hülshager & Maier, 2008; König et al., 2010; Schuler et al., 2007). Fast die Hälfte der befragten Personaler (48 %) gaben in der Untersuchung von Nachtwei und Schermuly (2009) an, Bedenken gegenüber Persönlichkeitstests zu haben. Die Mehrheit der Bedenkenträger setzte demzufolge auch keine Persönlichkeitstests ein und äußerte eine Vielzahl an (teilweise unberechtigten) Vorbehalten bzw. Vorurteilen gegenüber diesen Verfahren.

Die wissenschaftliche Gemeinde sollte sich darüber Gedanken machen, in welcher Form relevante Forschungsergebnisse zukünftig anschaulich und präzise für die Praxis aufbereitet sowie an relevante Zielpersonen versendet werden können. Vor allem die Praktikabilität und einfache Durchführbarkeit sowie Ökonomie der Verfahren sollten hierbei betont werden. Einige Verlage verschicken beispielsweise Rundmails mit Zusammenfassungen von Fachbüchern. Zwar ist hier der Marketinggedanke ausschlaggebend, aber auch die Forschung könnte sich dieser Methode bedienen, um eine höhere praktische Relevanz erreichen zu können. Denkbar wäre auch das Verlegen einer entsprechenden Zeitschrift oder das Durchführen von speziellen Kongressen von Wissenschaftlern für Praktiker. Hierdurch könnten wiederum Kooperationen zwischen Forschern und Praktikern entstehen, die eine Brücke zwischen Wissenschaft und Praxis schlagen. Zudem könnten die relevanten Anwender auf diese Weise lernen, seriöse, wissenschaftlich fundierte Verfahren von pseudo- oder parawissenschaftlichen Methodiken zu unterscheiden, für die in der Praxis nicht nur immer noch zu viel Geld bezahlt wird, sondern die sich nebenbei oftmals auch am Rande der Legalität bewegen (Bartoschek, 2011).

9.4.2 Implikationen für die Praxis

Gerade in der beruflichen Ausbildung und nachfolgenden Fortbildungen machen nicht nur Wissen und Wissensaneignung Berufserfolg aus, sondern v.a. der Umgang mit Kunden, Kollegen und Vorgesetzten sowie die Arbeitsweise im Allgemeinen und Speziellen sagen Berufserfolg vorher. Vor allem auch, da die Aneignung von (Fach-) Wissen mit der Zeit immer unwichtiger wird, sind es vor allem persönliche Eigenschaften, die später maßgeblich zum beruflichen Erfolg einer Person beitragen. Die Gewissenhaftigkeitsaspekte Fleiß und Ordnung in Kombination mit der Selbstwirksamkeit einer Person könnten zur Aufklärung beitragen, inwiefern eine Person geeignet ist, sich in der tendenziell immer weniger strukturierten (Arbeits-) Welt zurechtzufinden und erfolgreich zu sein. Die zahlreichen Fehler der letzten Jahre, die bis hin zu drastischen, weltweit spürbaren Fehlentscheidungen seitens der Manager und Führungskräfte großer Unternehmen reichen, sprechen eine deutliche Sprache (Sarges, 2010). Daher ist es wichtig, bereits bei der Auswahl von Auszubildenden, auf diese Persönlichkeitseigenschaften bei Bewerbern zu achten. Vor allem, wenn eine Vorauswahl nach kognitiven Fähigkeiten bereits stattgefunden hat, sollte man in weiteren Schritten vor allem die Persönlichkeit als Auswahlkriterium heranziehen.

Auch wenn sich in der vorliegenden Untersuchung nicht eindeutig die prädiktive Validität der untersuchten Eigenschaften Fleiß, Ordnung und Selbstwirksamkeit gezeigt hat und die Überprüfung der Vorhersagemodelle nicht in der Lage war, die Komplexität zwischen den Gewissenhaftigkeitsaspekten und dem Leistungskriterien aufzuklären, so liefert diese Studie dennoch wertvolle Erkenntnisse für die Praxis der beruflichen Eignungsdiagnostik sowie ein neues, reliables Testverfahren zur berufsbezogenen Erfassung der Gewissenhaftigkeitsaspekte Fleiß und Ordnung.

Die Überprüfung der Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS) ergab, dass das Instrument zuverlässig die arbeitsbezogenen Gewissenhaftigkeitsaspekte Fleiß und Ordnung erfasst und die zugrunde liegenden latenten Konstrukte abbildet. Bezüglich der prädiktiven Validität ergaben sich teilweise signifikante Zusammenhänge mit Schulleistung sowie Tendenzen, dass Fleiß bzw. Ordnung unterschiedlich stark mit verschiedenen Leistungs- und Erfolgskriterien im Bereich der Berufsausbildung zusammenhängen bzw. diese vorhersagen. Auch die berufsbezogene Selbstwirksamkeitsskala zeigte vergleichbare Ergebnisse.

In der Praxis der beruflichen Eignungsdiagnostik sollten die AGS erst einmal parallel zu anderen Auswahlinstrumenten eingesetzt und hinsichtlich ihrer praktischen Nützlichkeit sowie Vorhersagekraft überprüft werden. Vor allem der Vergleich der AGS mit anderen Gewissenhaftigkeitsmaßen sollte in der Praxis vorangetrieben werden.

Zwar ist mit der Einführung der AGS die Problematik der Vielzahl der angebotenen Testverfahren nicht verbessert worden. Dem Verfahren ist aber zugute zu halten, dass es sich um ein wissenschaftlich sorgfältig konstruiertes Verfahren handelt, das lediglich die Gewissenhaftigkeitsaspekte Fleiß und Ordnung erfasst. Es ist somit einfach und schnell mit anderen Auswahlverfahren kombinierbar. Auch eine Entwicklung weiterer arbeitsbezogener Persönlichkeitsmaße, beispielsweise zur Erfassung von Ausgeglichenheit, mit dem Ziel der Entwicklung eines „Baukasten-Modells“ zur Kombination verschiedener arbeitsbezogener Persönlichkeitsmaße bzw. Auswahlverfahren ist denkbar.

Mit Bezug auf die von Nachtwei und Schermuly (2009) aufgelisteten Bedenken von Personalern gegenüber Persönlichkeitstests ist anzuführen, dass der Einsatz der AGS sowie der Selbstwirksamkeitsskala nicht nur in großen Unternehmen nützlich ist, sondern grundsätzlich zu einer Erhöhung der Varianzaufklärung in Personalauswahlprozessen beitragen kann. Vor allem der Ordnungsaspekt wies in der vorliegenden Untersuchung bereits gegenüber einem allgemeinen Gewissenhaftigkeitsmaß inkrementelle Validität auf. Der Arbeitsbezug der Skalen sollte die Akzeptanz der Verfahren sowohl bei Personalern als auch Bewerbern bzw. Mit-

arbeitern erhöhen (Görlich & Schuler, 2007) und dazu beitragen, dass Bewerber von den Tests bzw. erfragten Inhalten nicht abgeschreckt werden. Auch durch die neuen Technologien, wie beispielsweise die sozialen Netzwerke wie Facebook o.ä., werden Anwendern immer mehr die Ängste gegenüber Testverfahren genommen. Vor allem junge Menschen reagieren weitaus weniger kritisch und neugieriger auf Persönlichkeitstests als ältere Probanden (Sarges, 2010). Falls die Praktiker die Testverfahren zu unpersönlich halten oder die getesteten Inhalte lieber im Interview erfragen, können die Skalen auch als Grundlage für gezielte Erkundigungen der Interviewer genutzt werden.

Sowohl die Gewissenhaftigkeitsaspekte Fleiß und Ordnung als auch das Selbstwirksamkeitskonstrukt erheben nicht den Anspruch, die gesamte Komplexität einer Person abbilden zu können. Vielmehr sollte es darum gehen, gezielt bestimmte, arbeits- und erfolgsrelevante persönliche Voraussetzungen abzufragen.

Länderübergreifend gibt es rechtliche Richtlinien, die u.a. für Auswahlprozesse festlegen, dass diese „*[should] be based on the important knowledge, skills, abilities, and other attributes required for successful performance in the job in question*“ (Satterwhite et al., 2009, S. 155). Das bedeutet konkret, dass auch Persönlichkeitsinventare zum Standard der eignungsdiagnostischen Verfahren gehören (sollten). So besagt auch Lord (2011), dass der Schlüssel dafür, ob es nützlich ist, Persönlichkeitseinschätzungen in den Methodenmix mit aufzunehmen, in der Frage liegt, ob die Persönlichkeitsschätzung dem etwas hinzufügt und über das hinausgeht, was anhand anderer Aspekte der Person vorhergesagt werden kann. Gewöhnlich ist dies der Fall. Generell sollte dem Einsatz der verschiedenen Verfahren immer eine ausgiebige Tätigkeitsanalyse der zu besetzenden Stellen vorausgehen, um abzuschätzen, welche Ausprägungen ein Bewerber mitbringen sollte. Denn eine anforderungsgeleitete Personalauswahl und -entwicklung sind eine Grundvoraussetzung für eine erfolgreiche Stellenbesetzung. Hierbei ist stets ein multimethodales Vorgehen am besten geeignet (Schuler & Höft, 2006), die Komplexität einer Person am besten zu erfassen. Vor allem Persönlichkeitstest sind nützliche, effektive und ökonomische Ergänzungen im Auswahlprozess (Backhaus, 2004b; Hagemeister et al., 2010; Lord, 2011; Risavy & Hausdorf, 2011; Rode et al., 2008). Bei sorgfältiger Auswahl und Zusammenstellung der Auswahlinstrumente sowie einer wertschätzenden Instruktion der Bewerber sollten Bewerbungstests – vor allem Persönlichkeitstests – die Bewerber nicht zusätzlich unter Druck setzen und somit auch keine Angst messen (Lee et al., 2010). Gerade in Verfahren, wie z.B. Assessment Centern oder anderen Gruppentestungen, in denen z.B. die Gewissenhaftigkeit der Probanden nicht direkt beobachtbar ist, bietet der Einsatz entsprechender Persönlichkeitstests eine nützliche Unterstützung des diagnostischen Portfolios (Hagemeister et al., 2010; Van der Zee et al., 2003).

Dem theoretisch bestehenden Problem der möglichen Durchschaubarkeit von Persönlichkeitstests ist entgegen zu setzen, dass Bewerber in der Regel nicht wissen, wie Persönlichkeitstests konstruiert sind oder ausgewertet werden und folglich auch nicht in der Lage sind, die verwendeten Instrumente zu durchschauen (König et al., 2011). Unabhängig davon ist nach wie vor umstritten, inwiefern sich das Verfälschen von Testergebnissen überhaupt nachteilig auf die Auswahlentscheidung auswirkt (Marcus, 2003). Die prognostische Validität der Verfahren bleibt dennoch bestehen (Blickle et al., 2011).

Derzeit gibt es für die Fleiß- und Ordnungsaspekte lediglich Informationen über die Nützlichkeit bei der Auswahl von Auszubildenden. Dennoch ist zu vermuten, dass die Gewissenhaftigkeitsaspekte auch in älteren Stichproben zum Erkenntnisgewinn in der beruflichen Eignungsdiagnostik beitragen können. Auch bei Managern und Führungskräften scheinen Fleiß und Ordnung differenziert Berufserfolg vorhersagen zu können (MacCann et al., 2009; Schmit et al., 1995). Denkbar ist aber auch ein negativer Zusammenhang des Ordnungsaspektes mit den Leistungskriterien dieser Stichproben (Ziegler et al., 2010). So ist es möglich, dass eine zu hohe Ordnungsliebe in Pedanterie umschlägt und effektives, zielorientiertes

Arbeiten behindert. Unternehmen sollten daher sowohl Fleiß als auch Ordnung in der Praxis der Personalauswahl mit erheben, um die Nützlichkeit der AGS für weitere Zielgruppen zu ermitteln.

Dem uneingeschränkten Einsatz der AGS in der Praxis der beruflichen Eignungsdiagnostik steht somit grundsätzlich – außer der noch ausstehenden Kriteriumsvalidierung – nichts entgegen. Die AGS erfüllen die Forderung von Hagemeister und Kollegen (2010) nach sorgfältig konstruierten Testverfahren. Neben einer theoretischen Verankerung im FFM bzw. dem hierarchischen Modell der Persönlichkeit von DeYoung (2010) sind die Skalen gemäß der Regeln der KTT konstruiert. Der Arbeitsbezug sollte die Akzeptanz des Tests erhöhen bzw. die Ablehnung der Ergebnisse verringern und die Situationsabhängigkeit der Antworten verringern. Die AGS können somit einige der in Abschnitt 2.4.2 aufgeführten Nachteile von Persönlichkeitstests ausgleichen. Das Verfahren ist eine nützliche und hilfreiche Erweiterung der Auswahlmethoden der beruflichen Eignungsdiagnostik.

Viele Personaler, die von der geringen Validität von Persönlichkeitstests in der Personalauswahl und -entwicklung enttäuscht sind, sei empfohlen, diese nicht zur Top-Down-Auswahl einzusetzen, sondern eher Cut-off-Werte festzulegen, anhand derer Bewerber (mit) ausgewählt werden (Le et al., 2010). Es ist angebracht, solche Verfahren bereits frühzeitig im Bewerbungsprozess als Screeningverfahren einzusetzen, was letztlich auch einen positiven Effekt auf die prädiktive Validität der Instrumente haben sollte. Auch das Problem des sozialerwünschten Antwortverhaltens könnte auf diese Weise minimiert werden, da Personen mit auffälligen Werten vorab gekennzeichnet oder ggf. aussortiert werden könnten. Allerdings ist es keine leichte Aufgabe, die Cut-off-Werte für eine optimale Vorhersage zu definieren. Dies sollte jeweils erst nach einer sorgfältigen Anforderungsanalyse und Sichtung relevanter Normdaten geschehen. Insgesamt können somit die Ergebnisse aus den Personalauswahlverfahren aber auch im Sinne einer effektiven Personalentwicklung bereits zu Beginn der Arbeitstätigkeit zur Formulierung von Strategien zur Zielerreichung genutzt werden.

Doch nicht nur zur Selektion der „Besten“ bzw. zur Erfolgsprognose können die untersuchten Persönlichkeitseigenschaften Fleiß und Ordnung sowie Selbstwirksamkeit herangezogen werden. Da die Gewissenhaftigkeitsdimension auch signifikante Zusammenhänge mit z.B. Durchfallraten zeigt, sind diese Eigenschaften auch für die Vorhersage bzw. Vermeidung von Misserfolg relevant (Poropat, 2009). Beispielsweise zur Antizipation potenzieller Lernprobleme und der frühzeitigen Einleitung von Gegenmaßnahmen. So könnten weniger ordentlichen Auszubildenden vermehrt Lernhilfen bereitgestellt werden oder Strukturen für Lernen geschaffen werden. Wenn die Studenten aufgrund fehlerhafter Lernstrategien und unzureichender Zielsetzungsfähigkeiten durchfallen, kann ein Training eine wirksame Maßnahme sein, um dies zu bekämpfen. Alternativ könnten auch die Lehrmethoden angepasst werden, sodass die Schüler bei ihrem Lernen mehr unterstützt werden. Die AGS könnten als Screeninginstrument hierzu einen nützlichen Beitrag bei der Identifizierung solcher Auszubildenden leisten.

Ein wesentlicher Vorteil dieser Studie ist zudem, dass in dieser explizit Auszubildende des Dualen Ausbildungssystems in Deutschland untersucht wurden. Die meisten vergleichbaren Studien mit Berufsanfängern bzw. Schülern in einer beruflichen Ausbildung werden in den USA mit College-Studenten als Probanden durchgeführt. Es ist fraglich, inwiefern die Ergebnisse aus Untersuchungen mit College-Studenten, die zudem größtenteils unter künstlichen Labor-Bedingungen durchgeführt werden, auf das deutsche Ausbildungssystem bzw. die Praxis der Personalauswahl hierzulande übertragbar sind. So berichteten Hülshager und Maier (2008) von deutlich geringeren prädiktiven Validitäten allgemeiner kognitiver Fähigkeiten bei der Vorhersage von Ausbildungserfolg in Deutschland ($r = .47$) im Vergleich zu den USA ($r = .57$) bzw. Europa ($r = .54$).

Zusammenfassend bleibt zu sagen, dass Fleiß, Ordnung und Selbstwirksamkeit tendenziell als nützliche Prädiktoren zur differenzierten Vorhersage von schulischem und beruflichem Erfolg im Bereich der beruflichen Ausbildung angesehen werden können. In der Praxis sollten diese Persönlichkeitsmerkmale zukünftig parallel zu verwandten Auswahlverfahren angewendet und hinsichtlich ihrer Nützlichkeit sowie Vorhersagekraft evaluiert werden. Gerade in Anbetracht der geringen Vergleichbarkeit von Schulleistungen bzw. der „Fragwürdigkeit der Notengebung“ (Görlich & Schuler, 2007) können Persönlichkeitstests eine sinnvolle und ökonomische Ergänzung zur Vorauswahl von Bewerbern bieten. Personaler laufen durch den Einsatz dieser Testverfahren weniger stark Gefahr, grundsätzlich gut geeignete Bewerber aufgrund mangelhafter schulischer Leistungen zu früh auszuschließen (Kersting, 2011). Gerade mit Blick auf den demografischen Wandel und die andauernden Reformen im deutschen Schulsystem können die hier untersuchten Konstrukte eine gute Orientierungsbasis für Auswahlentscheidungen bieten.

9.4.3 Fazit

Es ist wichtig, die praktische Relevanz dieser Studie weder zu hoch noch zu niedrig zu bewerten. Auch wenn die gefundenen Zusammenhänge andeuten, dass Persönlichkeit nur einen kleinen Anteil der Varianz aufklärt, ist die Höhe dieser Effekte dennoch vergleichbar mit Testergebnissen z.B. aus der Biomedizin beispielsweise bei der Vorhersage von Herzproblemen durch EKG-Stresstests, Schwangerschaftsergebnisse durch Ultraschalluntersuchungen oder Brustkrebsscreenings durch Mammografien (Noftle & Robins, 2007). Außerdem kann man bei multipel bedingten Ergebnissen, wie es akademische Leistungsergebnisse sind, nur niedrige prädiktive Validitäten erwarten. Die Ergebnisse der Studie deuten an, dass bedingtes Bemühen in Kombination mit einer Selbstwahrnehmung als schulisch kompetenter Person eine wichtige Rolle bei der Vorhersage von Schulleistung spielen. Andere Faktoren wie z.B. Werte, Motive, Attributionsstil, Selbstwirksamkeit oder sogar praktische Fähigkeiten wie z.B. der Umgang mit Geld (im Sinne von der geringeren Notwendigkeit, außerhalb der Schule nebenbei noch Geld verdienen zu müssen), sind weitere wichtige Faktoren bei der Vorhersage von akademischem Erfolg. D.h. es ist eine Vielzahl von Prädiktoren notwendig, um die individuellen Unterschiede in schulischen Leistungen zuverlässig vorhersagen und erklären zu können. Diese zu identifizieren ist die Aufgaben zukünftiger Forschungsarbeiten. Bevor man in Interventionen investiert, um Persönlichkeit (irgendwie) zu verändern, sollten weitere beeinflussende Faktoren, Moderatoren, Mediatoren und z.B. auch die wahrgenommenen eigenen Fähigkeiten weiter untersucht werden, um möglichst genaue Aussagen treffen und Interventionen ableiten zu können.

Aufgrund der tendenziell positiven Ergebnisse können die AGS für Forschungszwecke und einen vorsichtigen Einsatz in Unternehmen zu Validierungszwecken empfohlen werden. Die Verwendung der AGS zu Zwecken der Individualdiagnostik setzt eine Verfügbarkeit geeigneter Normdaten voraus. An eine Veröffentlichung des Inventars ist erst nach Untersuchung einer geeigneten Eichstichprobe gedacht.

10. Literaturverzeichnis

- Abele, A. E., Stief, M., & Andrä, M. S. (2000). Zur ökonomischen Erfassung beruflicher Selbstwirksamkeitserwartungen – Neukonstruktion einer BSW-Skala. *Zeitschrift für Arbeits- und Organisationspsychologie*, 44 (3), 145-151.
- Ackerman, P. L. (1996). A theory of adult intelligent development: Process, personality, interest, and knowledge. *Intelligence*, 22, 229-259.
- Ackerman, P. L., & Heggstad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, 121 (2), 219-245.
- Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Holt.
- Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, 47, (Whole No. 211).
- Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (Hrsg.) (1999). *Intelligenz-Struktur-Test 2000 (I-S-T 2000)*. Göttingen: Hogrefe.
- Anderson, N., & Ones, D. S. (2003). The construct validity of three entry level personality inventories used in the UK: Cautionary findings from a multiple-inventory investigation. *European Journal of Personality*, 17, 39-66.
- Anderson, G., & Viswesvaran, C. (1998) *An update of the validity of personality scales in personal selection: A meta analysis of studies published after 1992*. Vortrag auf der 13th Annual Conference of the Society of Industrial and Organizational Psychology, Dallas.
- Andresen, B. (1995). "Risikobereitschaft" (R) – der sechste Basisfaktor der Persönlichkeit: Konvergenz multivariater Studien und Konstruktexplikation. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 16, 210-236.
- Andresen, B. (2000). Six basic dimensions of personality and a seventh factor of generalized dysfunctional personality: A diathesis system covering all personality disorders. *Neuropsychobiology*, 41, 5-23.
- Andresen, B. (eingereicht). *A factor-analytical study of thirteen personality questionnaires: Evidence for eight domain factors of personality*.
- Andressen, P., & Konradt, U. (2007). Messung von Selbstführung: Psychometrische Überprüfung der deutschsprachigen Version des RSLQ [Measuring self-leadership: Psychometric test of the German version of the RSLQ]. *Zeitschrift für Personalpsychologie*, 6, 117-128.
- Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11, 150-166.
- Backhaus, K. (2004a). *Seelenstriptease für den Traumjob? Persönlichkeitstests in der Personalauswahl*. Online-Dokument. Verfügbar unter: https://www.psyreon.de/content/e480/Publikationen/persoentlichkeitstests_in_der_personalauswahl.pdf. [Datum des Zugriffs: 14.04.2010].
- Backhaus, K. (2004b). *Persönlichkeit als Forschungsgegenstand der Psychologie. Eine Einführung in das Big Five-Persönlichkeitsmodell*. Online-Dokument. Verfügbar unter: <https://www.psyreon.de/content/e480/Publikationen/persoentlichkeit.pdf> [Datum des Zugriffs: 14.04.2010].
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84 (2), 191-215.

- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Bandura, A., & Wood, R. (1989). Effect of perceived controllability and performance standards on self-regulation of complex decision making. *Journal of Personality and Social Psychology*, 56 (5), 805-814.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51 (6), 1173-1182.
- Barrick, M. R. & Mount, M. K. (1991). The Big Five Personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44, 1-26.
- Barrick, M. R., & Mount, M. K. (2004). Yes, personality matters: Moving on to more important matters. *Human Performance*, 18 (4), 359-372.
- Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *Personality and Performance*, 9 (1/2), 9-30.
- Barrick, M. R., Mount, M. K., & Strauss (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. *Journal of Applied Psychology*, 78, 715-722.
- Bartoschek, S. (2011). Vorsicht, Falle: Warum gestandene Personaler auf unseriöse Auswahlverfahren hereinfallen. *Wirtschaftspsychologie aktuell*, 4/2011, 58-60.
- Beauducel, A., & Kersting, M. (2010). *START-P. Testbatterie für Berufseinsteiger – Persönlichkeit*. Göttingen: Hogrefe.
- Becker, P. (1996). Wie big sind die Big Five? *Zeitschrift für Differentielle und Diagnostische Psychologie*, 17 (4), 209-221.
- Bell, B. S., & Kozlowski, S. W. J. (2002). Goal orientation and ability: Interactive effects on self-efficacy, performance, and knowledge. *Journal of Applied Psychology*, 87, 497-505.
- Betz, N. E., & Hackett, G. (1981). The relationship of career-related self-efficacy to perceived career options in college women and men. *Journal of Counseling Psychology*, 28, 399-410.
- Bierhoff, H.-W., Wegge, J., Bipp, T., Kleinbeck, U., Attig-Grabosch, C., & Schulz, S. (2005). Entwicklung eines Fragebogens zur Messung von Eigenverantwortung oder: „Es gibt nichts Gutes, außer man tut es“. [Development of a questionnaire to measure personal responsibility or: “Don’t talk about it, just do it”]. *Zeitschrift für Personalpsychologie*, 4, 4-18.
- Bidjerano, T., & Dai, D. (2007). The relationship between the big-five model of personality and self-regulated learning strategies. *Learning and Individual Differences*, 17, 69-81.
- Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the Frame-of-Reference Effect in personality scale scores: A replication and extension. *Journal of Applied Psychology*, 89 (1), 150-157.
- Bipp, T. (2010). What do people want from their jobs? The Big Five, core self-evaluations and work motivation. *International Journal of Selection and Assessment*, 18 (1), 28-39.
- Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, 14, 317-335.

- Blickle, G., Schnitzler, A., von Below, A., & Johannes, A. (2011). *Konstrukt- und kriterienbezogene Validität eines Selbsteinschätzungsinstrumentes in der Personalauswahl: Ergebnisse zweier experimenteller Studien*. Vortrag auf der 7. Tagung der Fachgruppe Arbeits- und Organisationspsychologie der Deutschen Gesellschaft für Psychologie, Rostock, 09. September 2011.
- Borkenau, P., & Ostendorf, F. (1991). Ein Fragebogen zur Erfassung fünf robuster Persönlichkeitsfaktoren. *Diagnostica*, 37 (1), 29-41.
- Borkenau, P., & Ostendorf, F. (Hrsg.) (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa & McCrae*. (1. Aufl.). Göttingen: Hogrefe.
- Borkenau, P., & Ostendorf, F. (Hrsg.) (2008). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa & McCrae*. (2. Aufl.). Göttingen: Hogrefe.
- Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance*, 10, 99-109.
- Bortz, J. (Hrsg.) (1993). *Statistik für Sozialwissenschaftler*. (4. Aufl.). Berlin: Springer-Verlag.
- Bortz, J. & Döring, N. (Hrsg.) (1995). *Forschungsmethoden und Evaluation* (2. Aufl.). Berlin: Springer.
- Bowles, S., & Gintis, H. (1999). Comments on "The long shadow of work". *Critical Sociology*, 25 (2/3), 3-7.
- Brown, S. D., Lent, R. W., & Gore, P. A., Jr. (2000). Self-rated abilities and self-efficacy beliefs: Are they empirically distinct? *Journal of Career Assessment*, 8 (3), 223-235.
- Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, 50, 255-272.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson Studium.
- Bulheller, S., & Häcker, H. O. (Hrsg.) (2001). *Rechtschreibtest (RST). Neue Rechtschreibregelung*. Frankfurt am Main: Swets & Zeitlinger B.V., Swets Test Services.
- Butcher, J. N. (Hrsg.) (2011). *A beginner's guide to the MMPI-2* (3. Aufl.). Washington, DC: American Psychological Association.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 56, 81-105.
- Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology*, 38, 476-507.
- Chamorro-Premuzic, T., & Furnham, A. (2003a). Personality traits and academic examination performance. *European Journal of Personality*, 17, 237-250.
- Chamorro-Premuzic, T., & Furnham, A. (2003b). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality*, 37, 319-338.
- Chamorro-Premuzic, T., & Furnham, A. (2004). A possible model for explaining the personality-intelligence interface. *British Journal of Psychology*, 95, 249-264.
- Chamorro-Premuzic, T., & Furnham, A. (2006). Intellectual competence and the intelligent personality: A third way in differential psychology. *Review of General Psychology*, 10, 251-267.

- Chamorro-Premuzic, T., & Furnham, A. (2008). Personality, intelligence and approaches to learning as predictors of academic performance. *Personality and Individual Differences*, 44, 1596-1603.
- Chen, G., Casper, W. J., & Cortina, J. M. (2001). The role of self-efficacy and task complexity in the relationships among cognitive ability, conscientiousness, self-efficacy, and work-related performance: A meta-analytic examination. *Human Performance*, 14, 209-230.
- Connor-Smith, J. K., & Flachsbart, C. (2007). Relations between Personality and Coping: A Meta-Analysis. *Journal of Personality and Social Psychology*, 93, 1080-1107.
- Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987-2003. *Review of Educational Research*, 76 (1), 1-62.
- Corker, K. S., Oswald, F. L., & Donnellan, M. B. (2012). Conscientiousness in the classroom: A process explanation. *Journal of Personality*, 80 (4), 995-1028.
- Costa, P. T., Jr., & McCrae, R. R. (1982). An approach to the attribution of aging, period, and cohort effects. *Psychological Bulletin*, 92, 238-250.
- Costa, P. T. (Jr.), & McCrae, R. R. (1988). Personality in adulthood: A six-year longitudinal study of selfreports and spouse ratings on the NEO personality inventory. *Journal of Personality and Social Psychology*, 54, 853-863.
- Costa, P. T. (Jr.), & McCrae, R. R. (Hrsg.) (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI)*. Professional manual. Odessa, FL: Psychological Assessment Resources.
- CRF Deutschland (Hrsg.) (2009). *TOP Arbeitgeber Deutschland 2009*. Bielefeld: W. Bertelsmann Verlag GmbH & Co. KG.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Cronbach, L. J. (1990). *Essentials of psychological testing*. Row: Harper.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24, 349-354.
- Cury, F., Sarrazin, P., & Famose, J. P. (1997). Achievement goals, perceived ability and active search for information. *European Yearbook of Sport Psychology*, 1, 166-183.
- De Fruyt, F., & Mervielde, I. (1996). Personality and interests as predictors of educational streaming and achievement. *European Journal of Personality*, 10, 405-425.
- De Fruyt, F., & Salgado, J. F. (2003). Applied personality psychology: Lessons learned from the IWO field. *European Journal of Personality*, 17, 123-131.
- Denis, P. L., Morin, D., & Guindon, C. (2010). Exploring the capacity of NEO PI-R facets to predict job performance in two French-Canadian samples. *International Journal of Selection and Assessment*, 18 (2), 201-207.
- De Raad, B., & Perugini, M. (Hrsg.) (2002). *Big Five Assessment*. Göttingen: Hogrefe & Huber.
- De Raad, B., & Schouwenburg, H. C. (1996). Personality in learning and education: A review. *European Journal of Personality*, 10, 303-336.

- DeYoung, C. (2010). Personality neuroscience and the biology of traits. *Social and Personality Psychology Compass*, 4 (12), 1165-1180.
- DeYoung, C. (2012). *A hierarchical model of personality based on the Big 5*. Online-Dokument. Verfügbar unter: <http://www.tc.umn.edu/~cdeyoung/Research2.html> [Datum des Zugriffs: 07.06.2012].
- DeYoung, C. G., Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, J. R. (2010). Testing predictions from personality neuroscience: Brain structure and the Big Five. *Psychological Science*, 21 (6), 820-828.
- DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *Journal of Personality and Social Psychology*, 93 (5), 880-896.
- Digman, J. M. (1989). Five robust trait dimensions: Development, stability, and utility. *Journal of Personality*, 57 (2), 195-214.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41, 417-440.
- DIN Deutsches Institut für Normung e.V. (2002). *Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen – DIN 33430*. Berlin.
- Dollinger, S. J., & Orf, L. A. (1991). Personality and performance in “personality”: Conscientiousness and openness. *Journal of Research in Personality*, 25, 276-284.
- Dollinger, S. J., Matyja, A. M., & Huber, J. L. (2008). Which factors best account for academic success: Those which college students can control or those they cannot. *Journal of Research in Personality*, 42, 872-885.
- Donnay, D. A. C., & Borgen, F. H. (1999). The incremental validity of vocational self-efficacy: An examination of interest, self-efficacy, and occupation. *Journal of Counseling Psychology*, 46, 432-447.
- Dreher, G. F., & Bretz, R. D. (1991). Cognitive ability and career attainment: The moderating effects of early career success. *Journal of Applied Psychology*, 76, 392-397.
- Dreier, K. (2008). *The impact of industriousness and orderliness on students' performance and career interests: A mediational analysis*. Unveröffentlichte Diplomarbeit, Christian-Albrechts-Universität zu Kiel, Kiel.
- Dreier, K., & Moldzio, T. (2010). Karriere dank Ordnung und Fleiß: Ein neuer Test misst die Gewissenhaftigkeit. *Wirtschaftspsychologie aktuell*, 1, 14-16.
- Dreier, K., Kasper, S., & Moldzio, T. (2010). *Die Arbeitsbezogenen Gewissenhaftigkeits-skalen (AGS)*. Vortrag auf dem 8. Kongress für Wirtschaftspsychologie: „Psychologie in der Wirtschaft: Chancen und Herausforderungen“, Potsdam, 15. Mai 2010.
- Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, 16, 939-944.
- Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology*, 91 (1), 40-57.
- DuVernet, A., Meade, A. W., Coughlin, C., & Kantrowitz, T. (2010). *DIF in personality assessment: Does cognitive ability influence item interpretation?* Vortrag präsentiert auf dem “25th Annual Meeting of the Society for Industrial and Organizational Psychology”, Atlanta, GA, April 2010.

- Eby, L. T., Butts, M., & Lockwood, A. (2003). Predictors of success in the era of the boundaryless career. *Journal of Organizational Behavior*, 24, 689–708.
- Eden, D., & Kinnar, J. (1991). Modeling Galatea: Boosting self-efficacy to increase volunteering. *Journal of Applied Psychology*, 76 (6), 770–780.
- Edwards, A. L., Abbott, R. D., & Klockars, A. J. (1972). A factor analysis of the EPPS and PRF personality inventories. *Educational and Psychological Measurement*, 32, 23–29.
- Egberink, I. J. L., Meijer, R. R., & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality*, 44, 232–244.
- Fahrenberg, J., Hampel, R., & Selg, H. (Hrsg.) (2001). *Das Freiburger Persönlichkeitsinventar (FPI)*. Handanweisung (7. überarbeitete und neunormierte Aufl.). Göttingen: Hogrefe.
- Fahrenberg, J., Hampel, R., & Selg, H. (Hrsg.) (2010). *Das Freiburger Persönlichkeitsinventar (FPI-R)*. Handanweisung (8. überarbeitete Aufl.). Göttingen: Hogrefe.
- Farsides, T., & Woodfield, R. (2003). Individual differences and undergraduate academic success: The roles of personality, intelligence, and application. *Personality and Individual Differences*, 34, 1225–1243.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Feltham, R., & Woods, J. (Hrsg.) (1995). *Business Personality Indicator Questionnaire (BPI)*. London, UK: NFER-Nelson.
- Fiske, D. W. (1949). Consistency of factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, 44, 329–344.
- Fraboni, M., & Cooper, D. (1989). Further validation of three short forms of the Marlowe-Crowne Scale of Social Desirability. *Psychological Reports*, 65, 595–600.
- Funder, D.C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, 40 (1), 21–34.
- Funke, U., Krauss, J., Schuler, H., & Stapf, K. H. (1987). Zur Prognostizierbarkeit wissenschaftlich-technischer Leistungen mittels Personvariablen: Eine Metaanalyse der Validität diagnostischer Verfahren im Bereich Forschung und Entwicklung. *Gruppendynamik*, 18, 407–428.
- Furnham, A., & Chamorro-Premuzic, T. (2004). Personality and intelligence as predictors of statistics examination grades. *Personality and Individual Differences*, 37, 943–955.
- Furnham, A., Chamorro-Premuzic, T., & McDougall, F. (2003). Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and Individual Differences*, 14, 49–66.
- Glug, I. (2004). *Die Bedeutung der Gewissenhaftigkeit in der Berufseignungsdiagnostik: Entwicklung und Validierung eines Fragebogens zur differenzierten Erfassung des Gewissenhaftigkeits-Faktors*. Unveröffentlichte Diplomarbeit, Christian-Albrechts-Universität zu Kiel, Kiel.
- Goldberg, L. R. (1971). A historical survey of personality scales and inventories. In: Reynolds, P. (Hrsg.), *Advances in psychological assessment* (2. Aufl., S. 293–336). Palo Alto, CA: Science and Behavior Books.

- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26-34.
- Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In: Mervielde, I., Deary, I., De Fruyt, F., & Ostendorf, F. (Hrsg.), *Personality psychology in Europe* (7. Aufl., S. 7-28). Tilburg, NL: Tilburg University Press.
- Gore, P. A., Jr. (2006). Academic self-efficacy as a predictor of college outcomes: Two incremental validity studies. *Journal of Career Assessment*, 14 (1), 92-115.
- Görlich, Y., & Schuler, H. (2007). *Arbeitsprobe zur berufsbezogenen Intelligenz – technische und handwerkliche Tätigkeiten (AZUBI-TH)*. Göttingen: Hogrefe Verlag.
- Göthlich, S. E. (2009). Zum Umgang mit fehlenden Daten in großzahligen empirischen Erhebungen. In: Albers, S., Klapper, D., Konradt, U., Walter, A. & Wolf, J. (Hrsg.), *Methodik der empirischen Forschung* (3. Aufl.) (S. 119-135). Wiesbaden: Gabler Verlag.
- Gough, H. G. (1957). *Manual for the California Psychological Inventory*. Palo Alto, CA: Consulting Psychologists Press.
- Gough, H. G. (1957). *Manual for the California Psychological Inventory*. Palo Alto, CA: Consulting Psychologists Press.
- Gough, H. G. (1964). Academic achievement in high school as predicted from the California Psychological Inventory. *Journal of Educational Psychology*, 55, 174-180.
- Gough, H. G., & Hall, W. B. (1964). Prediction of performance in medical school from the California Psychological Inventory. *Journal of Applied Psychology*, 48, 218-226.
- Gray, E. K., & Watson, D. (2002). General and specific traits of personality and their relation to sleep and academic performance. *Journal of Personality*, 70, 177-206.
- Graziano, W. G., & Ward, D. (1992). Probing the Big Five in adolescence: Personality and adjustment during a developmental transition. *Journal of Personality*, 60, 425-439.
- Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology*, 18, 135-164.
- Hackett, G., & Betz, N. E. (1995). Career choice and development. In: Maddux, J. E. (Hrsg.), *Self-efficacy, adaptation, and adjustment: Theory, research, and application* (S. 249-280). New York: Plenum.
- Hagemeister, C., Lang, F., & Kersting, M. (2010). Einstellungen von Psychologinnen und Psychologen in Deutschland zu Tests. *Report Psychologie*, 35, 428-439.
- Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, 57, 639-683.
- Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of conscientiousness. *Psychological Science*, 19 (4), 309-313.
- Höft, S., & Schuler, H. (2001). *Grundlagen einer persönlichkeitsorientierten Berufseignungsdiagnostik*. Positionsreferat im Rahmen der 6. Fachgruppe für Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik, Leipzig, 13.-14. September 2001.
- Hogan, R.T. (1982). A socioanalytic theory of personality. In: Page, M. (Hrsg.), *Handbook of industrial and organizational psychology* (2. Aufl., S. 873-919). Palo Alto, CA: Consulting Psychologists Press.

- Hogan, R. T. (1991). Personality and personality measurement. In: Dunette, M. D., & Hough, L. M. (Hrsg.), *Handbook of industrial and organizational psychology* (2. Aufl., S. 873-919). Palo Alto, CA: Consulting Psychologists Press.
- Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, 92 (5), 1270-1285.
- Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American Psychologist*, 51, 469-477.
- Holtz, B. C., Ployhart, R. E., & Dominguez, A. (2005). Testing the rules of justice: The effects of frame-of-reference and pre-test validity information on personality test responses and test perceptions. *International Journal of Selection and Assessment*, 13, 75-86.
- Horn, W. (Hrsg.) (1983). *Leistungsprüfsystem (L-P-S)*. Handanweisung. (2. Auflage). Hogrefe: Göttingen.
- Hossiep, R., & Paschen, M. (Hrsg.) (2003). *Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung - BIP* (2. vollst. überarbeitete Aufl.). Göttingen: Hogrefe.
- Hossiep, R., & Collatz, A. (2010). *BIP – Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung*. Online-Dokument. Verfügbar unter: www.testentwicklung/bip/bip_aufbau.htm [Datum des Zugriffs: 27.06.2012].
- Hough, L. M. (1992). The “Big Five” personality variables-construct confusion: Description versus prediction. *Human Performance*, 5 (1/2), 139-155.
- Hough, L. M., & Schneider, R. J. (1996). Personality traits, taxonomies, and applications in organizations. In: Murphy, K. R. (Hrsg.), *Individual differences and behavior in organizations* (S. 31-88). San Francisco, CA: Jossey Bass.
- Howard, P. J., & Howard, M. J. (2001). *Professional manual for the Workplace Big Five Profile (WB5P)*. Charlotte, NC: Centacs.
- Huba, G. J., Segal, B., & Singer, J. L. (1977). Consistency of daydreaming styles across samples of college males and female drug and alcohol users. *Journal of Abnormal Psychology*, 86, 99-102.
- Hülshager, U. R., & Maier, G. W. (2008). Persönlichkeitseigenschaften, Intelligenz und Erfolg im Beruf. Eine Bestandsaufnahme internationaler und nationaler Forschung. *Psychologische Rundschau*, 59 (2), 108-122.
- Humphreys, L. G. (1960). Investigations of the simplex. *Psychometrika*, 25, 313-323.
- Hunthausen, J. M., Truxillo, D. M., Bauer, T. N., & Hammer, L. B. (2003). A field study of frame-of-reference effects on personality test validity. *Journal of Applied Psychology*, 88, 545-551.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72-98.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.
- Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology*, 85, 869-879.
- IHK-Ostwestfalen (2012). *Industrie- und Handelskammer Ostwestfalen zu Bielefeld*. Online-Dokument. Verfügbar unter: www.ostwestfalen.ihk.de [Datum des Zugriffs: 12.04.2012].
- Jackson, D. N. (1976). *Jackson Personality Inventory Manual*. Port Huron, MI: Research Psychologists Press.

- Jackson, D. N. (1984). *Personality Research Form Manual*. Port Huron, MI: Research Psychologists Press.
- Jackson, D. N., Ashton, M. C., & Tomes, J. L. (1996). The six-factor model of personality: facets from the big five. *Personality and Individual Differences*, 21, 391-402.
- Jackson, D. N., Paunonen, S. V., & Tremblay, P. F. (2000). *Six Factor Personality Questionnaire Manual*. Port Huron, MI: Sigma Assessment Systems.
- Jackson, D. N., Paunonen, S. V., Fraboni, M., & Goffin, R. D. (1996). A five-factor versus a six-factor model of personality structure. *Personality and Individual Differences*, 20 (1), 33-45.
- Jackson, J. J., Walton, K. E., Harms, P. D., Bogg, T., Wood, D., Lodi-Smith, J., Edmonds, G. W., & Roberts, B. W. (2009). Not all conscientiousness scales change alike: A multi-method, multisample study of age differences on the facets of conscientiousness. *Journal of Personality and Social Psychology*, 96 (2), 446-459.
- Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the Indicators of Conscientiousness (BIC). *Journal of Research in Personality*, 44 (4), 501-511.
- John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In: Pervin, L. A. (Hrsg.), *Handbook of personality: Theory and research*. New York: Guilford.
- John, O. P., Donahue, E. M., Kentle, R. L., (1991). *The "Big Five" Inventory – Versions 4a and 54* (Technical Report). Berkley: University of California, Institute of Personality Assessment and Research.
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In: Pervin, L.A., & John, O. P. (Hrsg.), *Handbook of personality: Theory and research* (S. 102-138). New York: Guilford.
- Judge, T. A., & Bono, E. B. (2001). Relationship of Core Self-Evaluations traits – self-esteem, generalized self-efficacy, locus of control, and emotional stability – with job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology*, 86 (1), 80-92.
- Judge, T. A., Bono, J. E., Erez, A., & Locke, E. A. (2005). Core self-evaluations and job and life satisfaction: The role of self-concordance and goal attainment. *Journal of Applied Psychology*, 90, 257-268.
- Judge, T. A., Bono, J. E., & Locke, E. A. (2000). Personality and job satisfaction: The mediating role of job characteristics. *Journal of Applied Psychology*, 85, 237-249.
- Judge, T. A., Erez, A., & Bono, E. B. (1998). The power of being positive – the relation between positive self-concept and job-performance. *Human Performance*, 11, 167-187.
- Judge, T. A., Jackson, C. L., Chaw, J. C., Scott, B. A., & Rich, B. L. (2007). Self-efficacy and work-related performance: The integral role of individual differences. *Journal of Applied Psychology*, 92 (1), 107-127.
- Judge, T.A., Heller, D., & Mount, M.K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology*, 87, 530-541.
- Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, 52 (3), 621-652.

- Judge, T. A., & Illies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology*, 87, 797-807.
- Kanfer, R. (1992). Work motivation: New directions in theory and research. In: Cooper, C. L., & Robertson, I. T. (Hrsg.), *International review of industrial and organizational psychology* (7. Aufl., S. 1-53). Chichester, England: Wiley.
- Kanfer, R., & Heggstad, E. D. (1997). Motivational traits and skills: A person-centered approach to work motivation. *Research in Organizational Behavior*, 19, 1-56.
- Kanning, U. P. (2011). Selbstdarstellung in der Personalauswahl: Mythen, Fakten und Perspektiven. *Wirtschaftspsychologie aktuell*, 4/ 2011, 33-36.
- Kaplan, A., & Midgley, C. (1997). The effect of achievement goals: Does level of perceived academic competence make a difference? *Contemporary Educational Psychology*, 22, 415-435.
- Kasper, S. (in Vorbereitung). *Validierung eines Fragebogens zur berufsbezogenen Erfassung der Gewissenhaftigkeitsaspekte Ordnung und Fleiß*. Unveröffentlichte Diplomarbeit, Friedrich-Schiller-Universität, Jena.
- Keith, T. Z., Diamond-Hallam, C., & Fine, J. G. (2004). Longitudinal effects of in-school and out-of-school homework on high school grades. *School Psychology Quarterly*, 19 (3), 187-211.
- Kersting, M. (2011). Schlaraffenland ist abgebrannt – warum sich die Bewerberauswahl ändern muss. *Wirtschaftspsychologie aktuell*, 4/2011, 23-26.
- König, C. J., Jansen, A., & Lüscher Mathieu, P. (2011). *Was, wenn Bewerbende wissen, wie einfach Persönlichkeitstests ausgewertet werden?* Vortrag auf der 7. Tagung der Fachgruppe Arbeits- und Organisationspsychologie der Deutschen Gesellschaft für Psychologie, Rostock, 09. September 2011.
- König, C. J., Klehe, U-C., Berchtold, M., & Kleinmann, U. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment*, 18 (1), 17-27.
- Korinteli, E. (2009). *The relation between orderliness, industriousness, coping and job performance*. Unveröffentlichte Diplomarbeit, Christian-Albrechts-Universität zu Kiel, Kiel.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, 127 (1), 162-181.
- Landers, R. N. (2009). A quantitative examination of trends in I-O-Psychology 2001-2005. *The Industrial-Organizational Psychologist*, 46 (4), 15-23.
- Lang, F. R., & Lüdtke, O. (2005). Der Big Five Ansatz in der Persönlichkeitsforschung: Instrumente und Vorgehen. In S. Schumann (Hrsg.), *Persönlichkeit: Eine vergessene Größe der empirischen Sozialforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften/ GWV Fachverlage GmbH.
- Lang, F. R., Lüdtke, O., & Asendorpf, J. B. (2001). Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen [Psychometric properties and equivalence of the German version of the Big Five Inventory (BFI) for young, middle-aged, and old adults]. *Diagnostica*, 47, 111-121.
- Larson, L. M., & Borgen, F. H. (2006). Do personality traits contribute to vocational self-efficacy? *Journal of Career Assessment*, 14 (3), 295-311.

- Lazarus, R. S. (1974). Psychological stress and coping in adaptation and illness. *International Journal of Psychiatry in Medicine*, 5, 321-333.
- Lazarus, R. S., & Folkman, S. (1984). *Psychological Stress and the coping process*. New York: Springer.
- Le, H., Oh, I.-S., Robbins, S. B., Ilies, R., Holland, E., & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology*, 96 (1), 113-133.
- Lee, S., & Klein, H. J. (2002). Relationships between conscientiousness, self-efficacy, self-deception, and learning over time. *Journal of Applied Psychology*, 87 (6), 1175-1182.
- Lee, L. M., Piccone, K. A., Isaacson, J. A., Trejo, B. C., & Griffith, R. L. (2010). *Profiling the faker: The individual differences behind applicant faking behavior*. Poster präsentiert auf der "25th Annual Meeting of the Society for Industrial and Organizational Psychology", Atlanta, GA, April 2010.
- Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance [Monograph]. *Journal of Vocational Behavior*, 30, 347-382.
- Lewin, K. (1936) Principles of topological psychology. New York: McGraw-Hill.
- Lienert, G. A., & Raatz, U. (Hrsg.) (1994). *Testaufbau und Testanalyse* (5. Aufl.). Weinheim: Psychologie Verlags Union.
- Lievens, P., De Corte, W., & Schollaert, E. (2008). A closer look at the Frame-of-Reference Effect in personality scale scores and validity. *Journal of Applied Psychology*, 93 (2), 268-279.
- Lievens, F., Ones, D. S., & Dilchert, S. (2009). Personality scale validities increase throughout medical school. *Journal of Applied Psychology*, 94 (6), 1514-1535.
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57, 705-717.
- Lord, W. (Hrsg.) (2011). *Das NEO-Persönlichkeitsinventar in der berufsbezogenen Anwendung. Interpretation und Feedback*. Göttingen: Hogrefe.
- Lounsbury, J. W., Sundstrom, E., Loveland, J. M., & Gibson, L. W. (2003). Intelligence, "Big Five" personality traits, and work drive as predictors of course grade. *Personality and Individual Differences*, 35, 1231-1239.
- Lounsbury, J. W., Gibson, L. W., Sundstrom, E., Wilburn, D., & Loveland, J. M. (2004). An empirical investigation of the proposition that 'School is Work': A comparison of personality-performance correlations in school and work settings. *Journal of Education and Work*, 17 (1), 119-131.
- MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences*, 19 (4), 451-458.
- Maddux, J. E. (2001). Self-efficacy: The power of believing you can. In: Snyder, C. R., & Lopez, S. J. (Hrsg.). *Handbook of positive psychology*. New York: Oxford University Press.
- Maddux, J. E., & Meier, L. J. (1995). Self-efficacy and depression. In: Maddux, J. E. (Hrsg.), *Self-efficacy, adaptation, and adjustment: Theory, research and application* (S. 143-169). New York: Plenum.

- Martocchio, J. J., & Judge, T. A. (1997). Relationship between conscientiousness and learning in employee training: Mediating influences of self-deception and self-efficacy. *Journal of Applied Psychology*, 82, 764-773.
- Maties, M., & Wottawa, H. (2011). Nicht immer erlaubt: Wann Eignungstests künftig zulässig sind. *Wirtschaftspsychologie aktuell*, 4/2011, 28-32.
- Marcus, B. (2003). Persönlichkeitstests in der Personalauswahl: Sind „sozial erwünschte“ Antworten wirklich nicht wünschenswert? *Zeitschrift für Psychologie*, 211 (3), 138-148.
- Marcus, B., & Schuler, H. (2006). Leistungsbeurteilung. In: Schuler, H. (Hrsg.), *Lehrbuch der Personalpsychologie* (2. Aufl.) (S. 434-469). Göttingen: Hogrefe.
- Mayer, J. D. (2005). A tale of two visions. Can a new view of personality help integrate psychology? *American Psychologist*, 60 (4), 294-307.
- McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the Five-Factor Model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52, 81-90.
- Miller, R. B., Behrens, J. T., Greene, B. A., & Newman, D. (1993). Goals and perceived ability: Impact on student valuing, self-regulation, and persistence. *Contemporary Educational Psychology*, 18, 2-14.
- Minbashian, A., Wood, R. E., & Beckmann, N. (2010). Task-contingent conscientiousness as a unit of personality at work. *Journal of Applied Psychology*, 95 (5), 793-806.
- Mischel, W. (1968). *Personality and assessment*. New York: John Wiley.
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102, 246-268.
- Moldzio & Partner – Institut für Personalauswahl (2009). *Personalbarometer Sommer 2009. Gewissenhaftigkeit im Arbeitskontext - Eine Praxisanalyse*. Online-Dokument. Verfügbar unter: http://www.moldzio.com/fileadmin/redaktion/Personalbarometer_Sommer_2009.pdf [Datum des Zugriffs: 25.09.2009].
- Mount, M. K., & Barrick, M. R. (1995). The Big Five personality dimensions: Implications for research and practice in human resources management. *Research in Personnel and Human Resources Management*, 13, 153-200.
- Moutafi, J., Furnham, A., & Paltiel, L. (2004). Why is conscientiousness negatively correlated with intelligence? *Personality and Individual Differences*, 37, 1013-1022.
- Müller, D. (2007). Moderatoren und Mediatoren in Regressionen. In S. Albers, D. Klapper, U. Konradt, A. Walter, & J. Wolf (Hrsg.), *Methoden der empirischen Forschung*, S. 245-260. Wiesbaden: Gabler.
- Murphy, S. E., & Ensher, E. A. (1999). The effects of leader and subordinate characteristics in the development of leader-member exchange quality. *Journal of Applied Social Psychology*, 29 (7), 1371-1394.
- Nachtwei, J., & Schermuly, C. C. (2009). Acht Mythen über Eignungstests. *Harvard Business Manager*, 4/ 09, 6-10.
- Natanovich, G., & Eden, D. (2001). *Pygmalion effects among outreach supervisors and tutors: Extending gender and ethnic generalizability*. Vortrag auf der 16th annual conference of the Society for Industrial and Organizational Psychology, San Diego, April, 2001.

- Nauta, M. M. (2004). Self-efficacy as a mediator of the relationships between personality factors and career interests. *Journal of Career Assessment*, 12, 381-394.
- Nauta, M. M., Epperson, D. L., & Waggoner, K. M. (1999). Perceived causes of success and failure: Are women's attributions related to satisfaction and persistence in engineering majors? *Journal of Research in Science Teaching*, 36, 663-676.
- Nesselroade, J. R., & Baltes, P. B. (1975). Higher order factor convergence and divergence of two distinct systems of personality: Cattell's HSPQ and Jackson's PRF. *Multivariate Behavioral Research*, 10, 387-408.
- Ng, T., Eby, L. T., Sorensen, L. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. *Personnel Psychology*, 58, 367-408.
- Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, 93 (1), 16-130.
- Nye, C. D., Do, B.-R., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment*, 16 (2), 112-120.
- O'Brien, E., & LaHuis, D. M. (2011). Do applicants and incumbents respond to personality items similarly? A comparison of dominance and ideal point response models. *International Journal of Selection and Assessment*, 19 (2), 109-118.
- O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, 43, 971-990.
- Okech, O. P., & Harrington, R. (2002). The relationships among black consciousness, self-esteem, and academic self-efficacy in african american men. *Journal of Psychology*, 136 (2), 214-24.
- Olkin, I., & Pratt, J. W. (1958). A multivariate tchebycheff inequality. *The Annals of Mathematical Statistics*, 29 (1), 226-234.
- O'Neill, T. A., & Allen, N. A. (2007). *Broad versus narrow traits: Assessing the bandwidth-fidelity tradeoff at the team-level*. Poster präsentiert auf dem „Meeting of the Interdisciplinary Network for Group Research“, Lansing, MI, Juli 2007.
- Ones, D. S., & Viswesvaran, C. (1996). Bandwidth-Fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior*, 17, 609-626.
- Ones, D. S., & Viswesvaran, C. (1997). *Empirical and theoretical considerations in using conscientiousness measures in personnel selection*. Vortrag auf dem 5th European Congress of Psychology, Dublin, 06.-11.07.1997.
- O'Reilly, C., & Chatman, J. (1994). Working smarter and harder: A longitudinal study of managerial success. *Administrative Science Quarterly*, 39, 603-627.
- Ortner, T. M., Horn, R., Kersting, M., Krumm, S., Kubinger, K. D., Proyer, R. T., Schmidt-Atzert, L., Schuhfried, G., Schütz, A., Wagner-Menghin, M. M., & Westhoff, K. (2007). Standortbestimmung und Zukunft Objektiver Persönlichkeitstests. *Report Psychologie*, 32 (2), 60-69.
- Ostendorf, F., & Angleitner (Hrsg.) (2004). *NEO-PI-R. NEO-Persönlichkeitsinventar nach Costa und McCrae: Revidierte Fassung*. Göttingen: Hogrefe.
- Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57, 401-421.

- Pace, V. L., & Brannick, M. T. (2010). Improving prediction of work performance through frame-of-reference consistency: Empirical evidence using openness to experience. *International Journal of Selection and Assessment*, 18 (2), 230-235.
- Paulhus, D. L., Bruce, M. N., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin*, 21, 100-108.
- Pauls, C. A., & Crost, N. W. (2004). Effects of faking on self deception and impression management scales. *Personality and Individual Differences*, 37, 1137-1151.
- Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of behavior. *Journal of Personality and Social Psychology*, 74, 538-556.
- Paunonen, S. V., & Ashton, M. C. (2001a). Big Five predictors of academic achievement. *Journal of Research in Personality*, 35, 78-90.
- Peabody, D., & de Raad, B. (2002). The substantive nature of psycholexical personality factors: A comparison across languages. *Journal of Personality and Social Psychology*, 83, 983-997.
- Perugini, M., & Gallucci, M. (1997). A hierarchical faceted model of the Big Five. *European Journal of Personality*, 11, 279-301.
- Poropat, A. E. (2009). A meta-analysis of the Five-Factor-Model of personality and academic performance. *Psychological Bulletin*, 135 (2), 322-338.
- Poropat, A. E. (2011). The Eysenckian personality factors and their correlations with academic performance. *British Journal of Educational Psychology*, 81, 41-58.
- Rammstedt, B. (2004). *Zur Bestimmung der Gütekriterien von Multi-Item-Skalen: Eine Einführung*. ZUMA How-to-Reihe Nr. 12. Online-Dokument. Verfügbar unter: <http://www.gesis.org/forschung-lehre/gesis-publikationen/gesisreihen/how-to/> [Datum des Zugriffs: 24.11.2010].
- Rammstedt, B., & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K): Entwicklung und Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit. *Diagnostica*, 51 (4), 195-206.
- Rasch, D., & Kubinger, K. D. (Hrsg.) (2006). *Statistik für das Psychologiestudium*. (1. Aufl.) München: Elsevier Spektrum Akademischer Verlag.
- Reddock, C. M., Biderman, M. D., & Nguyen, N. T. (2010). *Increasing the validity of personality questionnaires*. Paper presented at the 25th Annual Conference of The Society for Industrial and Organizational Psychology, Atlanta, GA.
- Reddock, C. M., Biderman, M. D., & Nguyen, N. T. (2011). The relationship of reliability and validity of personality tests to Frame-of-Reference instructions and within-person inconsistency. *International Journal of Selection and Assessment*, 19 (2), 119-131.
- Renner, K.-H. (2011). Selbstdarstellung und Persönlichkeit: Ergebnisse eines anwendungsorientierten Forschungsprogramms. *Report Psychologie*, 36 (6), 260-271.
- Rigotti, T., Schyns, B., & Mohr, G. (2008). A short version of the Occupational Self-Efficacy Scale: Structural and construct validity across five countries. *Journal of Career Assessment*, 16 (2), 238-255.
- Risavy, S. D., & Hausdorf, P. A. (2011). Personality testing in personnel selection: Adverse impact and differential hiring rates. *International Journal of Selection and Assessment*, 19 (1), 18-30.

- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130 (2), 271–288.
- Roberts, B. W., Walton, K. E., & Bogg, T. (2005). Conscientiousness and health across the life course. *Review of General Psychology*, 9 (2), 156–168.
- Roberts, B. W., & Mroczek, D. (2008). Personality trait change in adulthood. *Current Directions in Psychological Science*, 17, 31–35.
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A metaanalysis of longitudinal studies. *Psychological Bulletin*, 132, 1–25.
- Rode, J. C., Arthaud-Day, M. L., Mooney, C. H., Near, J. P., & Baldwin, T. T. (2008). Ability and personality predictors of salary, perceived job success, and perceived career success in the initial career stage. *International Journal of Selection and Assessment*, 16 (3), 292–299.
- Rothstein, M. G., Paunonen, S. V., Rush, J. C., & King, G. A. (1994). Personality and cognitive ability predictors of performance in graduate business school. *Journal of Educational Psychology*, 86, 516–530.
- Rottinghaus, P. J., Lindley, L. D., Green, M. A., & Borgen, F. H. (2002). Educational aspirations: The contributions of personality, self-efficacy, and interests. *Journal of Vocational Behavior*, 60, 1–19.
- Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management*, 26, 565–606.
- Sadri, G., & Robertson, I. T. (1993). Self-efficacy and work-related behaviour: A review and metaanalysis. *Applied Psychology: An International Review*, 42, 139–152.
- Salgado, J. F. (1997). The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology*, 82, 30–43.
- Salgado, J. F. (1998). The Big Five personality dimensions and job performance in army and civil occupations: A European perspective. *Human Performance*, 11, 271–288.
- Salgado, J.F., Anderson, N., Moscoso, S. Bertua, C., & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A european community meta-analysis. *Personnel Psychology*, 56, 573–605.
- Sarges, W. (2010). Trendwenden der Personaldiagnostik: Gestern – heute – morgen. Höchst persönliche Erfahrungen und Wertungen eines unmittelbar Betroffenen. *Report Psychologie*, 35, 172–186.
- Satterwhite, R. C., Fleenor, J. W., Braddy, P. W., Feldman, J., & Hoopes, L. (2009). A case for homogeneity of personality at the occupational level. *International Journal of Selection and Assessment*, 17 (2), 154–164.
- Saucier, G. (2003). An alternative multi-language structure for personality attributes. *European Journal of Personality*, 17, 179–205.
- Saucier, G. (2009). Recurrent personality dimensions in inclusive lexical studies: Indications for a Big Six structure. *Journal of Personality*, 77, 1577–1614.

- Saucier, G., & Ostendorf, F. (1999). Hierarchical subcomponents of the Big Five personality factors: A crosslanguage replication. *Journal of Personality and Social Psychology*, 76, 613-627.
- Schein, E. H. (1998). *Karriereanker. Die verborgenen Muster in Ihrer beruflichen Entwicklung*. (5.Aufl.). Darmstadt: Beratungssozietät Lanzenberger Dr. Looss Stadelmann.
- Schimmack, U. (2010). What multi-method data tell us about construct validity. *European Journal of Personality*, 24, 241-257.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124 (2), 262-274.
- Schmidt, F. L., Shafer, J. A., & Oh, I.-S. (2008). Increased accuracy for range restriction correction: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, 61, 827-868.
- Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-Reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology*, 80 (5), 607-620.
- Schneider, R. J., Hough, L. M., & Dunnette, M. D. (1996). Broad-sided by narrow traits: How to sink science in five dimensions or less. *Journal of Organizational Behavior*, 17, 639-655.
- Schuler, H., Hell, B., Trapmann, S., Schaar, H., & Borarmir, I. (2007). Die Nutzung psychologischer Verfahren der externen Personalauswahl in deutschen Unternehmen. Ein Vergleich über 20 Jahre. *Zeitschrift für Personalpsychologie*, 6 (2), 60-70.
- Schuler, H., & Höft, S. (2001). Konstruktorientierte Verfahren in der Personalauswahl. In: H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 94-133). Göttingen: Hogrefe.
- Schuler, H., & Höft, S. (2006). Konstruktorientierte Verfahren in der Personalauswahl. In: Schuler, H. (Hrsg.), *Lehrbuch der Personalpsychologie* (2. Aufl.) (S. 102-144). Göttingen: Hogrefe.
- Schwarzer, R., Bäßler, J., Kwiatek, P., Schröder, K., & Zhang, J. X. (1997). The assessment of optimistic self-beliefs: Comparison of the German, Spanish, and Chinese versions of the General Self-Efficacy Scale. *Applied Psychology: An International Review*, 46 (1), 69-88.
- Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In: J. Weinman, S. Wright & M. Johnston (Hrsg.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (S. 35-37). Windsor, UK: NFER-NELSON.
- Schyns, B., & Moldzio, T. (2009). *The value of occupational self-efficacy in selection and development*. Vortrag präsentiert auf der "Conference of the British Academy of Management", Brighton, September 2009.
- Schyns, B., & von Collani, G. (2002). A new occupational self-efficacy scale and its relation to personality constructs and organizational variables. *European Journal of Work and Organizational Psychology*, 11 (2), 219-241.
- Sherer, M., & Adams, C. (1983). Construct validity of the self-efficacy scale. *Psychological Reports*, 53, 899-902.
- SHL (2012). *The Occupational Personality Questionnaire*. Online-Dokument. Verfügbar unter: <http://www.shl.com/assets/resources/OPQ-UK.pdf> [Zugriff am 14.9.2012].

- Snow, R. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist*, 27, 5–32.
- Snow, R. (1995). Foreword. In: Saklofske, D. H., & Zeidner, M. (Hrsg.), *International handbook of personality and intelligence* (S. 11-15). New York: Plenum.
- Specht, J., Egloff, B., & Schunkle, S. C. (2011). Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of Personality and Social Psychology*, 101 (4), 862-882.
- Speier, C., & Frese, M. (1997). Generalized self-efficacy as a mediator and moderator between control and complexity at work and personal initiative: A longitudinal field study in East Germany. *Human Performance*, 10, 171-192.
- Spengler, M., & Fintrup, A. (2011). Tests statt Uni-Noten: Bessere Erfolgsprognosen für Wirtschaftsprüfer. *Wirtschaftspsychologie aktuell*, 4/ 2011, 55-57.
- Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Social and Personality Psychology*, 84, 1041-1053.
- Staff, J., Schulenberg, J. E., & Bachman, J. G. (2010). Adolescent work intensity, school performance, and academic engagement. *Sociology of Education* 83 (3), 183–200.
- Stajkovic, A. D., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin*, 124 (2), 240-261.
- Statistisches Bundesamt (Hrsg.) (2011). *Statistisches Jahrbuch 2011 für die Bundesrepublik Deutschland mit >>internationalen Übersichten<<*. Wiesbaden: Statistisches Bundesamt. Online-Dokument. Verfügbar unter: www.destatis.de/jahrbuch [Datum des Zugriffs: 19.07.2012].
- Stöber, J. (1999). Die Soziale.Erwünschtheits-Skala-17 (SES-17): Entwicklung und erste Befunde zur Reliabilität und Validität [The Social Desirability Scale-17 (SDS-17): Development and first findings on reliability and validity]. *Diagnostica*, 45 (4), 173-177.
- Stöber, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment*, 17, 222-232.
- Struckmeier, L. (2009). *The relation between conscientiousness and self-leadership and their impact on job performance*. Unveröffentlichte Diplomarbeit, Christian-Albrechts-Universität zu Kiel, Kiel.
- Stumpf, H., Angleitner, A., Wieck, T., Jackson, D. N., & Beloch-Till, H. (1985). *Deutsche Personality Research Form (PRF)*. Göttingen: Hogrefe.
- Sudman, S., Bradburn, N. M., & Schwarz, N. (Hrsg.) (1996). *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass.
- Sutherland, R., de Bruin, P. P., & Crous, F. (2007). The relation between conscientiousness, empowerment and performance. *Journal of Human Resource Management*, 5 (2), 60-67.
- Sutin, A. R., & Costa, P. (2010). Reciprocal influences of personality and job characteristics across middle adulthood. *Journal of Personality*, 78 (1), 257-288.
- Sutin, A. R., Costa, P. T. Jr., Miech, R., & Eaton, W. W. (2009). Personality and career success: Concurrent and longitudinal relations. *European Journal of Personality*, 23 (2), 71-84.

- Tabachnick, B. G., & Fidell, L. S. (Hrsg.) (2001). *Using Multivariate Statistics* (4. Aufl.). Needham Heights, MA: Allyn & Bacon.
- Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, 72 (2), 271-324.
- Tellegen, A. (1982). *Brief Manual for the Multidimensional Personality Questionnaire*. University of Minnesota; Minneapolis: Unpublished manuscript.
- Terracciano, A., Löckenhoff, C. E., Zonderman, A. B., Ferrucci, L., & Costa, P. T. (2008). Personality predictors of longevity: Activity, emotional stability, and conscientiousness. *Psychosomatic Medicine*, 70, 621-627.
- Terracciano, A., McCrae, R. R., Brant, L. J., & Costa, P. T. (2005). Hierarchical linear modeling analyses of the NEO-PI-R Scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging*, 20 (3), 493-506.
- Tross, S. A., Harper, J. P., Osher, L. W., & Kneidinger, L. M. (2000). Not just the usual cast of characteristics: Using personality to predict college performance and retention. *Journal of College Student Development*, 41, 323-334.
- Trautwein, U., Lüdtke, O., Roberts, B. W., Schnyder, I., & Niggli, A. (2009). Different forces, same consequence: Conscientiousness and competence beliefs are independent predictors of academic effort and achievement. *Journal of Personality and Social Psychology*, 97, 1115-1128.
- Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology*, 98 (2), 438-456.
- Van der Zee, K. I., Zaal, J. N., & Piekstra, J. (2003). Validation of the Multicultural Personality Questionnaire in the context of personnel selection. *European Journal of Personality*, 17, 77-100.
- Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59, 197-210.
- Viswesvaran, C., Deller, J., & Ones, D. S. (2007). Personality measures in personnel selection: Some new contributions. *International Journal of Selection and Assessment*, 15 (3), 354-358.
- von Polier, X. (2010). *Persönlichkeitstests im Test. Reise ins Ich*. Online Dokument. Verfügbar unter: <http://www.spiegel.de/unispiegel/jobundberuf/0,1518,672556,00.html> [Datum des Zugriffs: 10.03.2010].
- Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment Research & Evaluation*, 14 (5), 1-9. Online-Dokument. Verfügbar unter: <http://pareonline.net/getvn.asp?v=14&n=5> [Datum des Zugriffs: 07.05.2012].
- Wiggins, N., Blackburn, M., & Hackman, J. R. (1969). Prediction of first-year graduate success in psychology: Peer ratings. *Journal of Educational Research*, 63 (2), 81-85.
- Wild, E., & Möller, J. (Hrsg.) (2009). *Pädagogische Psychologie*. Heidelberg: Springer Medizin Verlag.
- Wirth, M. (2008). *Zum Einfluss von Persönlichkeit und Intelligenz auf die Ausbildungszufriedenheit. Eine quer- und längsschnittliche Untersuchung*. Unveröffentlichte Dissertation, Freie Universität Berlin, Berlin.

- Wolfe, R. N., & Johnson, S. D. (1995). Personality as a predictor of college performance. *Educational and Psychological Measurement*, 55, 177–185.
- Wright, J. C., & Mischel, W. (1987). A conditional analysis of dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology*, 53, 1159–1177.
- Ziegeler, L. (2011). *Verfälschungstendenzen in Persönlichkeitsinventaren von Bewerbern im Auswahlkontext: Eine Vergleichsstudie*. Unveröffentlichte Diplomarbeit. Leopold-Franzens-Universität, Innsbruck.
- Ziegeler, L., Dreier, K., & Moldzio, T. (2012). *Faking von Persönlichkeitstests in der Praxis der Personalauswahl: Eine Vergleichsstudie*. Best Practice in der Wirtschaftspsychologie. Begleitband zum 9. Kongress für Wirtschaftspsychologie, Potsdam.
- Ziegler, M., Danay, E., Schölmerich, F., & Bühner, M. (2010). Predicting academic success with the Big 5 rated from different points of view: Self-rated, other rated and faked. *European Journal of Personality*, 24 (4), 341-355.

Zusammenfassung der Untersuchung

Persönlichkeitseigenschaften haben sich seitens der Forschung als valide Prädiktoren von schulischem und beruflichem Erfolg erwiesen. Berufsgruppen- und hierarchieübergreifend hat sich die Gewissenhaftigkeitsdimension als validester Prädiktor von schulischem und beruflichem Erfolg erwiesen, der auch inkrementelle Validität gegenüber allgemeinen kognitiven Fähigkeiten besitzt (Barrick et al., 2001; Poropat, 2009). Neuere Forschung deutet an, dass die Facetten der Gewissenhaftigkeitsdimension gegenüber der allgemeinen Gewissenhaftigkeitsdimension höhere prädiktive Validitäten besitzen und Leistung differenzierter und präziser Vorhersagen (Bipp, 2010; Dudley et al., 2006; O'Connor & Paunonen, 2007). Die Substruktur der Gewissenhaftigkeitsdimension ist nach wie vor umstritten, allerdings deutet sich an, dass eine zweifaktorielle Substruktur derzeit die beste Lösung ist, das Bandwidth-Fidelity-Dilemma der Persönlichkeitspsychologie zu lösen (DeYoung et al., 2007; Jackson et al., 2009). Die Gewissenhaftigkeitsdimension wird den Ergebnissen zufolge in einen eher proaktiven Aspekt (*Fleiß*) sowie einen eher passiven Aspekt (*Ordnung*) unterteilt.

Darüber hinaus fehlt in der bestehenden Forschungsliteratur zur Vorhersage von schulischer und beruflicher Leistung bisher ein komplexes Modell, das die Zusammenhänge zwischen Persönlichkeitsmerkmalen und Erfolg erklärt (Dudley et al., 2006; Lord, 2011). Der Einbezug weiterer beeinflussender Variablen bzw. die Integration der Traittheorie in andere Ansätze wurde von verschiedenen Forschern empfohlen. Ähnlich dem Interaktionistischen Ansatz (Funder, 2006) wurde in dieser Untersuchung das Selbstwirksamkeitskonstrukt als Repräsentant der sozial-kognitiven Lerntheorie (Bandura, 1977) als weiterer Prädiktor hinzugenommen, um mittels eines Person x Umwelt-Ansatzes zu einer höheren Varianzaufklärung beitragen zu können. Darüber hinaus wurde zusätzlich das bislang in der Eignungsdiagnostik eher unbekannte Konstrukt der wahrgenommenen eigenen Fähigkeiten (Nofle & Robins, 2007) in die Modellbildung mit aufgenommen. Selbstwirksamkeit sowie die wahrgenommenen eigenen Fähigkeiten wurden in dieser Untersuchung hinsichtlich ihres moderierenden bzw. medierenden Einflusses des Zusammenhangs zwischen den Gewissenhaftigkeitsaspekten Fleiß und Ordnung mit den erhobenen Leistungskriterien überprüft (Barrick & Mount, 2004).

Ergänzend dazu sprechen sich Forscher in der beruflichen Eignungsdiagnostik zunehmend dafür aus, die verwendeten Instrumente zur Personalauswahl und -entwicklung mit einem Arbeitsbezug zu versehen, um eine höhere Kriteriumsvalidität sowie Akzeptanz der psychologischen Verfahren zu erreichen (Lievens et al., 2008; Pace & Brannick, 2010).

Ziel der Arbeit war es, die prognostische Validität der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie des Selbstwirksamkeitskonstruktes im Rahmen eines komplexen Modells zu überprüfen. Die verwendeten Instrumente sollten einen Arbeitsbezug aufweisen, um eine höhere Kriteriumsvalidität zu erreichen.

Da in der beruflichen Eignungsdiagnostik in Deutschland bislang kein ausreichend validiertes arbeitsbezogen formuliertes Instrument zur Erfassung der Gewissenhaftigkeitsaspekte Fleiß und Ordnung vorliegt, wurde in einem ersten Schritt dieser Untersuchung die Arbeitsbezogenen Gewissenhaftigkeitsaspekte (AGS; Dreier & Moldzio, 2010) hinsichtlich ihrer Konstruktvalidität und Reliabilität überprüft.

Um die Gütekriterien der AGS zu überprüfen, wurde die Skala in einer Grundgesamtheit zweier Kohorten von Ausbildungsbewerbern im kaufmännischen und technischen Bereich ($N=740$) eingesetzt und hinsichtlich verschiedener Gütekriterien untersucht. Die Ergebnisse deuteten auf eine befriedigende bis gute interne Konsistenz der beiden Gewissenhaftigkeitsaspekte hin. Mittels MTMM-Analysen sowie konfirmatorischer Faktorenanalysen konnte die inhaltliche Validität der Skalen bestätigt werden. Die AGS bilden valide die beiden latenten Konstrukte Fleiß und Ordnung ab.

Der zweite Teil der Untersuchung betraf die Überprüfung der prädiktiven Validität der Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie des Selbstwirksamkeitskonstruktes in der Feldstichprobe der Auszubildenden ($N_R = 70$). Die Feldstichprobe wurde in technische und kaufmännische Auszubildende unterteilt und mit einer Kontrollgruppe von Schülern in Abschlussklassen aus den drei gängigen Schulformen in Deutschland verglichen ($N_{KG} = 97$). Die prognostische Validität wurde mittels Kriteriumskorrelationen mit den verschiedenen Leistungskriterien berechnet. Für die Feldstichprobe standen eine Vorgesetztenbeurteilung der Ausbildungsleiter sowie die allgemeine Berufsschulleistung als Kriteriumsmaße zur Verfügung. In der technischen Subpopulation konnten ebenfalls die Ergebnisse der IHK-Zwischenprüfungen als Kriterien herangezogen werden. In der Kontrollgruppe dienten die Abschlussnoten der Schüler als abhängige Variable.

Aufgrund mangelnder Teststärken in den beiden Subgruppen der Feldstichprobe kam es zu keinen signifikanten Ergebnissen in der Berechnung der prognostischen Validität. Es ergaben sich aber nicht-signifikante Hinweise auf die differentielle Vorhersagekraft der AGS. In der Kontrollgruppe konnte die Vorhersagekraft der Fleiß- und Ordnungsaspekte sowie der Selbstwirksamkeit signifikant nachgewiesen werden.

Die Überprüfung der Vorhersagemodelle ergab aufgrund der geringen Teststärke weitestgehend keine signifikanten Ergebnisse, deuteten aber auf eine partielle Mediation der Zusammenhänge durch die Selbstwirksamkeit und wahrgenommenen eigenen Fähigkeiten in Bezug auf den Zusammenhang zwischen den Gewissenhaftigkeitsaspekten Fleiß und Ordnung mit den verschiedenen Leistungskriterien für Ausbildungserfolg hin. Insgesamt scheint es aber, dass das postulierte Modell immer noch eine zu starke Vereinfachung der Zusammenhänge darstellte, auch wenn wie von Funder (2006) gefordert, Mediations- und Moderationsanalysen berechnet wurden. Die Beziehung zwischen Persönlichkeit und Arbeit ist nicht einfach. Zukünftige Studien sollten weitere Einflussvariablen und auch nicht-lineare Zusammenhänge überprüfen (Le et al., 2011; Viswesvaran et al., 2007).

Die Überprüfung der inkrementellen Validität der Gewissenhaftigkeitsaspekte bildete den letzten Teil dieser Untersuchung. Die Berechnungen ergaben für den Ordnungsaspekt tendenziell eine zusätzliche Varianzaufklärung über ein allgemeines Gewissenhaftigkeitsmaß hinweg. Die Ergebnisse der Überprüfung der intrapersonellen Variabilität der arbeitsbezogenen Gewissenhaftigkeitsaspekte im Vergleich mit dem generischen Gewissenhaftigkeitsmaß deuteten zudem teilweise an, dass der Arbeitsbezug der Skala zu einem homogenen Antwortverhalten der Probanden geführt hat.

Die Überprüfung der Kontrollvariablen ergab, dass sozial erwünschtes Antwortverhalten einen Einfluss auf die Ergebnisse gehabt haben könnte. Allerdings konnten diesbezüglich keine eindeutigen Ergebnisse ermittelt werden.

Das gewählte Untersuchungsdesign gewährleistet eine hohe externe Validität der Ergebnisse für die eignungsdiagnostische Untersuchung von Ausbildungsbewerbern.

Die vorliegende Untersuchung ist als Brückenschlag zwischen Wissenschaft und Praxis der beruflichen Eignungsdiagnostik zu interpretieren. Mit der Überprüfung der AGS stellt sie ein reliables und inhaltsvalides Verfahren vor, das in der Zukunft seitens der Forschung weiter hinsichtlich seiner prognostischen Validität und Gültigkeit auch in Studien mit Berufstätigen verschiedener bzw. höherer Altersstufen überprüft werden sollte. Die AGS könnten zukünftig in der Praxis der Eignungsdiagnostik eine sinnvolle Ergänzung der bestehenden Auswahlinstrumente darstellen. Seitens der Wissenschaft sollte darüber hinaus weiter an der Entwicklung und Überprüfung komplexer theoretischer Vorhersagemodelle für den Kontext der Leistungsvorhersage gearbeitet werden. Die Ergebnisse dieser Forschungsarbeiten sollten der Praxis anschaulich vermittelt werden, um insbesondere auch für die Praktiker in der beruflichen Eignungsdiagnostik einen hohen Zugewinn zu erreichen.

Summary

Personality traits have proven to be valid predictors of academic and occupational success. Independent from occupations and hierarchical levels, especially the conscientiousness trait, turned out to be the most valid predictor of scholastic and occupational achievement. Conscientiousness also exhibits incremental validity over general mental abilities (Barrick, Mount & Judge, 2001; Poropat, 2009; Schmidt & Hunter, 1998). Current research suggests that the facets of conscientiousness have higher predictive validities and are better predictors of achievement than the general conscientiousness trait (Bipp, 2010; Dudley et al., 2006; O'Connor & Paunonen, 2007). The underlying structure of the conscientiousness trait is still controversial, but currently a two-factor termination seems to be the best solution to solve the bandwidth-fidelity-dilemma within the personality research (DeYoung, Quilty & Peterson, 2007; Jackson et al., 2009). Concerning these results, the conscientiousness trait should be subdivided into a more proactive aspect, named *industriousness*, and a more passive aspect, named *orderliness*.

Furthermore, the existing research concerning the prediction of scholastic and occupational achievement still lacks a complex model which is able to explain the relationship between personality traits and success (Dudley et al., 2006; Lord, 2011). Various scientists recommend the integration of other influential variables and constructs, i.e. integrating the trait theory with other theories. Similar to the interactionistic approach (Funder, 2006) the self-efficacy construct as representative of the social-cognitive theory (Bandura, 1977) was implemented as additional predictor in this investigation. The aim was to reach a higher amount of explained variance due to this person x environment approach. Additional to the self-efficacy-construct, the still rather unknown construct of perceived abilities (Noftle & Robins, 2007) was integrated into the predictive model. Self-efficacy and perceived abilities were investigated concerning their moderating and mediating effects on the relationship between the conscientiousness aspects of industriousness and orderliness and achievement at school or at work.

Thus, one aim of this study was to test the predictive validity of the conscientiousness aspects of industriousness and orderliness as well as the self-efficacy construct within a complex predictive model.

The methods used should have a work-related frame of reference in order to achieve a better predictive validity, as researchers increasingly recommend adding an occupational frame of reference to the methods used for selecting staff, because work-related questionnaires and tests have proven to have higher predictive validities as well as being better accepted (Lievens et al., 2008; Pace & Brannick, 2010).

As there is not a sufficiently validated and work-related worded questionnaire to assess industriousness and orderliness, the first step of the study at hand was to test the construct validity and reliability of the Arbeitsbezogene Gewissenhaftigkeitsskalen (AGS; Dreier & Moldzio, 2010). For this purpose, the AGS were applied to two cohorts of applicants for either a commercial or technical apprenticeship ($N=740$). The results show satisfactory to good internal consistencies for both scales. Using MTMM- and confirmatory factor analyses (CFA) the content validity could be confirmed as well. Thus, the AGS has proven to be a valid scale which correctly represents the latent constructs of industriousness and orderliness.

The second step of the study at hand was to test the predictive validity of the conscientiousness aspects of industriousness and orderliness as well as the self-efficacy construct in a field sample of apprentices ($N=70$). The field sample was divided into commercial and technical apprentices and compared to a control group consisting of different students of the three common school forms in Germany ($N=97$). Individual assessments of the training managers concerning the achievement of the apprentices as well as the individual vocational school

GPA were used as criteria in the field sample. For the subpopulation of technical apprentices, grades from the intermediate examination were available and used as additional criteria. In the control group, the GPA of their final examination was used as the criterion variable.

Correlation analyses were used to assess the predictive validities. Due to small sample sizes within the field sample no significant predictions for industriousness, orderliness or self-efficacy were found. But the data showed tendencies indicating differential predictive power of the independent variables. Also, no significant correlations for the control group could be found. The examination of the predictive model resulted in mainly non-significant results due to a lack of power within the study at hand. However, the results indicated partial mediational effects of the constructs of self-efficacy and perceived abilities.

Altogether it seems that the recommended predictive model is still an oversimplification of the relationship between the two conscientiousness aspects and the investigated criteria, although mediation and moderation analyses were conducted as suggested by Funder (2006). The relationship between personality and success at work seems to be neither unidirectional nor linear. Future research should add and investigate additional variables as well as consider non-linear relationships between the variables (Le et al., 2011; Viswesvaran et al., 2007).

The last part of the study at hand was the examination of the incremental validity of the AGS. The work-related orderliness aspect revealed a tendency to explain more and additional variance when compared to a general conscientiousness measure. The industriousness aspect did not show such a tendency. Additionally, the comparison of the intrapersonal variability of the two conscientiousness aspects with the general conscientiousness scale partially indicated that the frame of reference of the AGS-items led to a more homogenous answering style within subjects. This result can be seen as evidence that contextualizing items, i.e. adding a work-related frame of reference, is useful.

The study at hand also integrated several control variables in the analyses, namely age, gender, former occupational experiences, education, and faking. Of these control variables age, gender, former occupational experiences, and education had no impact on the data. Faking however might have had an impact on the results, although it was not possible to identify any systematic influence of the construct.

The experimental design of the field study warranted an extremely high external validity of the results for the context of occupational aptitude testing, especially in the field of apprenticeships/ beginners.

This field study aimed at bridging the existing gap between theory and field. With the examination of the AGS, a reliable and content valid questionnaire was introduced. Future research should further investigate this scale, especially its predictive validity. Scientists should also continue to identify the variables of a complex predictive model in order to explain the relationship between personality and achievement. Researchers should also increase the communication of their results to the practitioners. This should result in a surplus for both sides. In the applied field of personnel selection the AGS could become a useful addition to the existing occupational selection methods.

Appendix

A1. Ergebnisse der MVA und Mittelwertimputation

Insgesamt fiel der Anteil der fehlenden Werte eher gering aus. So gab es in der Bewerberstichprobe ($N=760$) insgesamt nur sieben Personen, die mehr als 5 % fehlende Werte aufwiesen bzw. mindestens zwei Items nicht beantworteten. Im Vergleich zur Gesamtgröße der Bewerberstichprobe fiel dieser Anteil jedoch nicht merklich ins Gewicht (0,1 %). 45 Bewerber (5,9 %) wiesen ein nicht beantwortetes Item (= 3,3 % Missing in den AGS) auf. In der Kontrollgruppe ($N=97$) gab es keine Person, die mehr als 5 % Missings in den AGS aufwies. Insgesamt gab es in dieser Gruppe nur sieben Schüler (7,2 %), die ein Item der AGS nicht beantworteten, für den Rest (= 92,8 %) lagen vollständig bearbeitete Fragebögen vor. Dieser Anteil vollständig bearbeiteter Fragebögen war vergleichbar mit der Bewerberstichprobe (93,2 %).

Tabelle A-1: Fehlende Werte in den AGS

Anzahl fehlender Werte pro Fall	Anteil fehlender Werte pro Fall	n (absolute Häufigkeit)	Kumuliertes n	Gültige Prozente	Kumulierte Prozente
9	30,0 %	1	1	0,1 %	0,1 %
3	10,0 %	3	4	0,4 %	0,5 %
2	6,7 %	3	7	0,4 %	0,9 %
1	3,3 %	45 (7)	52 (7)	5,9 % (7,2 %)	6,8 % (7,2 %)
0	0,0 %	708 (90)	760 (97)	93,2 % (92,8 %)	100,0 % (100,0 %)

Anmerkungen: Die Werte der Schüler-Kontrollgruppe stehen in Klammern. Die AGS hat 30 Items.

Für den NEO-FFI mit einbezogener Selbstwirksamkeitsskala (T1) sind in Tabelle A-2 die fehlenden Werte dargestellt.

Tabelle A-2: Fehlende Werte im NEO-FFI und Selbstwirksamkeitsskala

Anzahl fehlender Werte pro Fall	Anteil fehlender Werte pro Fall	n (absolute Häufigkeit)	Kumuliertes n	Gültige Prozente	Kumulierte Prozente
32	40,5 %	4	4	0,5 %	0,5 %
31	39,2 %	1	5	0,1 %	0,6 %
17	21,5 %	1	6	0,1 %	0,7 %
16	20,3 %	6	12	0,8 %	1,5 %
14	17,7 %	8	20	1,1 %	2,6 %
5	6,3 %	1	21	0,1 %	2,7 %
3	3,8 %	2	23	0,3 %	3,0 %
2	2,5 %	3	26	0,4 %	3,4 %
1	1,3 %	32	58	4,2 %	7,6 %
0	0,0 %	702	760	92,4 %	100,0 %

Anmerkungen: Der NEO-FFI und die Selbstwirksamkeitsskala haben gemeinsam 79 Items.

Auffällig war die Diskrepanz zwischen technischen und kaufmännischen Bewerbern bezüglich der fehlenden Werte. So wiesen die technischen Bewerber ($N=358$) mit 33 (AGS) bzw. 37 (NEO-FFI + SE) Personen mit fehlerhaften Selbstbeschreibungsinventaren vergleichsweise mehr Missings (9,2 % bzw. 10,2 %) auf. In der Subgruppe der kaufmännischen Bewerber ($N=402$) beantworteten 17 (AGS; 4,2 %) bzw. 21 Personen (NEO-FFI + SE; 5,2 %) die Fragebogenverfahren nicht vollständig. Der Anteil der Missings lag in der Kontrollgruppe ($N=97$) mit sieben unvollständigen AGS (7,2 %) zwischen diesen Werten.

Tabelle A-3 beinhaltet die fehlenden Werte der Auszubildenden ($N=65$), die im Rahmen der Nachbefragung noch einmal getestet wurden. Insgesamt waren fünf unvollständige Fragebögen (7,7 %) enthalten, in denen jeweils ein Item nicht beantwortet wurde.

Tabelle A-3: Fehlende Werte Retest AGS (Auszubildende)

Anzahl fehlender Werte pro Fall	Anteil fehlender Werte pro Fall	n (absolute Häufigkeit)	Kumuliertes n	Gültige Prozente	Kumulierte Prozente
1	3,3 %	5	5	7,7 %	7,7 %
0	0,0 %	60	65	92,3 %	100,0 %

Anmerkungen: In die Berechnung gingen nur die Fragebögen ein, die eindeutig zuzuordnen waren.

In den Tabellen A-4 und A-5 sind die fehlenden Werte zur Selbstwirksamkeitsskala sowie den neun Items zur Erfassung der wahrgenommenen eigenen Fähigkeiten dargestellt. Die Daten der Stichprobe der Auszubildenden ($N=65$) ist mit denen der Kontrollgruppe der Schüler ($N=97$) relativ vergleichbar. In beiden Stichproben gibt es maximal ein bis zwei nicht beantwortete Items pro Skala.

Tabelle A-4: Fehlende Werte Selbstwirksamkeit T2 Azubis und Kontrollgruppe

Anzahl fehlender Werte pro Fall	Anteil fehlender Werte pro Fall	n (absolute Häufigkeit)	Kumuliertes n	Gültige Prozente	Kumulierte Prozente
2	10,5 %	1	1	1,5 %	1,5 %
1	5,3 %	3 (4)	4 (4)	4,6 % (4,1 %)	6,1 % (4,1 %)
0	0,0 %	61 (93)	65 (97)	93,9 % (95,9 %)	100,0 % (100,0 %)

Anmerkungen: Die Daten der Auszubildenden ($N=65$) beziehen sich auf die Nachtestung. Die Daten der Kontrollgruppe ($N=97$) stehen in Klammern.

Tabelle A-5: Fehlende Werte wahrgenommene eigene Fähigkeiten

Anzahl fehlender Werte pro Fall	Anteil fehlender Werte pro Fall	n (absolute Häufigkeit)	Kumuliertes n	Gültige Prozente	Kumulierte Prozente
2	22,2 %	1	1	1,5 %	1,5 %
1	11,1 %	2 (6)	3 (6)	3,1 % (6,2 %)	4,6 % (6,2 %)
0	0,0 %	62 (91)	65 (97)	95,4 % (93,8 %)	100,0 % (100,0 %)

Anmerkungen: Die Daten der Auszubildenden ($N=65$) beziehen sich auf die Nachtestung. Die Daten der Kontrollgruppe ($N=97$) stehen in Klammern.

Aufgrund der Kürze der verwendeten Instrumente sind die Anteile der fehlenden Werte innerhalb der einzelnen Skalen recht hoch. Wenn man alle Variablen zusammen analysiert, ergeben sich weniger hohe Fehlendraten.

Für die in den Stichproben verbleibenden Probanden wurden die fehlenden Werte durch eine Mittelwertimputation ersetzt. Diese sog. Mittelwertimputation hat den positiven Effekt, dass der geschätzte Randmittelwert gleich bleibt, birgt jedoch die Gefahr, dass es zu verzerrten Schätzungen von Varianzen, Kovarianzen und Korrelationen kommen kann. Ob es eine signifikante Veränderung der Mittelwerte und Standardabweichungen nach der Mittelwertimputation gegeben hat, wurde mittels eines t -Tests für gepaarte Stichproben überprüft.

Tabelle A-6 dokumentiert, welche Fälle aus der Bewerberstichprobe aufgrund eines zu hohen Anteils fehlender Werte aus der Untersuchung ausgeschlossen wurden.

Tabelle A-6: Ausgeschlossene Fälle aufgrund zu hohem Anteil fehlender Werte

Fall-Nr.	Anzahl fehlender Werte	Anteil fehlender Werte	Ausbildungsrichtung
142	14	12,7 %	kaufmännisch
222	14	12,7 %	kaufmännisch
371	31	28,2 %	kaufmännisch
331	32	29,1 %	kaufmännisch
119	32	29,1 %	kaufmännisch
25	14	12,7 %	technisch
186	14	12,7 %	technisch
245	14	12,7 %	technisch
85	14	12,7 %	technisch
262	14	12,7 %	technisch
19	16	14,4 %	technisch
221	16	14,4 %	technisch
143	16	14,4 %	technisch
333	17	15,3 %	technisch
311	17	15,3 %	technisch
62	32	29,1 %	technisch
200	16	14,4 %	technisch
237	16	14,4 %	technisch
13	17	15,3 %	technisch
292	32	29,1 %	technisch

In der Bewerberstichprobe ($N_{Bew}=740$) mussten 95 Mittelwertimputationen vorgenommen werden. Tabelle A-7 enthält die Mittelwerte und Zusammenhänge der Fleiß- und Ordnungsskala der AGS sowie der Selbstwirksamkeitsskala vor und nach der Mittelwertimputation für diese Stichprobe.

Tabelle A-7: Mittelwerte vor und nach der Mittelwertimputation in der Bewerberstichprobe

Skala	<i>M</i>	<i>M</i> imputiert
AGS Fleiß	3,07 (<i>SD</i> = 0,36)	3,07 (<i>SD</i> = 0,36)
AGS Ordnung	3,12 (<i>SD</i> = 0,37)	3,12 (<i>SD</i> = 0,37)
Selbstwirksamkeitsskala	2,98 (<i>SD</i> = 0,34)	2,98 (<i>SD</i> = 0,34)

Anmerkungen: $N_{Bew}=740$. AGS= Arbeitsbezogene Gewissenhaftigkeitsskalen; *M*= Skalenmittelwert; *SD*= Standardabweichung; ** $p < 0,01$ (zweiseitig).

In der Bewerberstichprobe veränderten sich die Mittelwerte und Standardabweichungen durch die Mittelwertimputation nicht. Eine Berechnung der Veränderungen für die Gewissenhaftigkeitsaspekte Fleiß und Ordnung sowie die Selbstwirksamkeitsskala war daher nicht notwendig, da sich die Werte nicht unterschieden bzw. der Standardfehler der Differenz gleich null war. Somit kann davon ausgegangen werden, dass das Imputationsverfahren in der Bewerberstichprobe zu keinen gravierenden Messverzerrungen geführt hat.

In der Kontrollgruppe ($N_{KG}=97$) wurden 17 Mittelwertimputationen durchgeführt. Tabelle A-8 enthält die Mittelwerte und Zusammenhänge der Fleiß- und Ordnungsskala der AGS sowie der Selbstwirksamkeitsskala vor und nach der Mittelwertimputation für diese Stichprobe.

Tabelle A-8: Mittelwerte vor und nach der Mittelwertimputation in der Kontrollgruppe

Skala	<i>M</i>	<i>M</i> imputiert
AGS Fleiß	2,67 (<i>SD</i> = 0,42)	2,67 (<i>SD</i> = 0,42)
AGS Ordnung	2,76 (<i>SD</i> = 0,47)	2,77 (<i>SD</i> = 0,47)
Selbstwirksamkeitsskala	2,63 (<i>SD</i> = 0,49)	2,63 (<i>SD</i> = 0,49)

Anmerkungen: N_{KG} = 97; AGS= Arbeitsbezogene Gewissenhaftigkeitsskalen; *M*= Skalenmittelwert; *SD*= Standardabweichung; ** $p < 0,01$ (zweiseitig).

In der Kontrollgruppe veränderten sich die Mittelwerte und Standardabweichungen durch die Mittelwertimputation nicht bzw. nur geringfügig im erweiterten Nachkommastellenbereich. Der Test auf Vergleich der Mittelwerte wurde daher sowohl für den Fleißaspekt ($p = .81$) als auch für den Ordnungsaspekt ($p = .11$) nicht signifikant. Eine Berechnung der Veränderung für die Selbstwirksamkeitsskala war nicht möglich, da sich die Werte nicht unterschieden bzw. der Standardfehler der Differenz gleich null war. Somit kann davon ausgegangen werden, dass das Imputationsverfahren in der Kontrollgruppe zu keinen gravierenden Messverzerrungen führte.

In der Reteststichprobe ($N_R = 63$) mussten zwölf Mittelwertimputationen durchgeführt werden. Tabelle A-9 enthält die Mittelwerte und Zusammenhänge der Fleiß- und Ordnungsskalen der AGS sowie der Selbstwirksamkeitsskala vor und nach der Mittelwertimputation für die Reteststichprobe.

Tabelle A-9: Mittelwerte vor und nach der Mittelwertimputation in der Reteststichprobe

Skala	<i>M</i>	<i>M</i> imputiert
AGS Fleiß	3,07 (<i>SD</i> = 0,32)	3,07 (<i>SD</i> = 0,32)
AGS Ordnung	3,08 (<i>SD</i> = 0,35)	3,08 (<i>SD</i> = 0,35)
Selbstwirksamkeitsskala	2,86 (<i>SD</i> = 0,34)	2,86 (<i>SD</i> = 0,34)

Anmerkungen: $N_R = 63$; AGS= Arbeitsbezogene Gewissenhaftigkeitsskalen; *M*= Skalenmittelwert; *SD*= Standardabweichung; ** $p < 0,01$ (zweiseitig).

In der Reteststichprobe veränderten sich durch die Mittelwerte und Standardabweichungen durch die Mittelwertimputation nur geringfügig im erweiterten Nachkommastellenbereich. Der Test auf Vergleich der Mittelwerte wurde sowohl für die beiden Gewissenhaftigkeitsaspekte Fleiß ($p = .32$) und Ordnung ($p = .56$) als auch für das Konstrukt der Selbstwirksamkeit ($p = .40$) nicht signifikant. Somit kann davon ausgegangen werden, dass das Imputationsverfahren in der Kontrollgruppe zu keinen gravierenden Messverzerrungen geführt hat.

Insgesamt ist davon auszugehen, dass die Ersetzung der fehlenden Werte mittels Mittelwertimputation keinen Einfluss auf die Ergebnislänge dieser Studie hatte. Daher wurde von weiteren Überprüfungen eventueller Einflüsse fehlender Werte abgesehen.

A2. Ergebnisse der Konsistenzanalysen

Tabelle A-10 können die Kennwerte der internen Konsistenz (Cronbachs Alpha) entnommen werden, die resultieren würden, wenn das entsprechende Item weggelassen worden wäre.

In der Bewerberstichprobe ($N_{Bew} = 740$) hätte eine Itemlöschung nicht zu einer höheren Konsistenz der Fleißskala geführt ($\alpha = .86$). In der Kontrollgruppe ($N_{KG} = 97$) hätte lediglich eine Löschung von F4 zu einer Erhöhung der internen Konsistenz von $\alpha = .80$ auf $\alpha = .82$ geführt.

Bezüglich des Ordnungsaspekts ergibt sich ein ähnliches Bild. Hier hätte in der Bewerberstichprobe die Löschung von O19 zu einer leichten Erhöhung der internen Konsistenz von $\alpha = .78$ auf $\alpha = .79$ geführt. In der Kontrollgruppe gab es zwei Items (O7 und O9), deren Löschung eine Erhöhung der internen Konsistenz von $\alpha = .75$ auf $\alpha = .76$ erwirkt hätte.

Insgesamt deuteten die Konsistenzwerte auf homogene Skalen hin.

Tabelle A-10: Steigerungsmöglichkeiten von Cronbachs Alpha

Fleiß			Ordnung		
Cronbachs Alpha, wenn Item weggelassen			Cronbachs Alpha, wenn Item weggelassen		
Item	BEW	KG	Item	BEW	KG
Fleiß 2	.85	.79	Ordnung 1	.77	.74
Fleiß 4	.86	.82	Ordnung 3	.75	.71
Fleiß 6	.86	.80	Ordnung 5	.77	.74
Fleiß 8	.85	.79	Ordnung 7	.78	.76
Fleiß 10	.86	.79	Ordnung 9	.77	.76
Fleiß 12	.85	.78	Ordnung 11	.77	.74
Fleiß 14	.85	.80	Ordnung 13	.75	.72
Fleiß 16	.85	.79	Ordnung 15	.75	.72
Fleiß 18	.85	.79	Ordnung 17	.76	.75
Fleiß 20	.85	.79	Ordnung 19	.79	.75
Fleiß 22	.85	.78	Ordnung 21	.77	.75
Fleiß 24	.85	.79	Ordnung 23	.76	.72
Fleiß 26	.86	.80	Ordnung 25	.76	.73
Fleiß 28	.85	.80	Ordnung 27	.76	.73
Fleiß 30	.85	.78	Ordnung 29	.78	.75

Anmerkungen: BEW= Bewerberstichprobe ($N_{Bew} = 740$); KG= Schülerkontrollgruppe ($N_{KG} = 97$).

A3. Ergebnisse der Multikollinearitätsüberprüfung der AGS

Die Ergebnisse der Überprüfung der multivariaten Normalverteilung auf Item-Ebene sind in Tabelle A-11 enthalten. Weisen die Items eine Schiefe (Skewness) kleiner als 2,00 und einen Exzess (Kurtosis) kleiner als 7,00 auf, ist von multivariater Normalverteilung auszugehen (Bühner, 2011). Dies trifft für alle Items in beiden Stichproben zu (Ausnahme: O7-Bewerberstichprobe). Eine Verletzung der Verteilungsannahme ist also auszuschließen.

Tabelle A-11: Überprüfung der multivariaten Normalverteilung auf Itemebene

Items	Schiefe	Exzess
Ordnung 1	-0,480 (-0,380)	0,925 (-0,101)
Ordnung 3	-0,717 (-0,646)	0,906 (-0,220)
Ordnung 5	-0,259 (-0,141)	-0,322 (-0,317)
Ordnung 7	-2,303 (-1,180)	5,487 (0,304)
Ordnung 9*	-1,068 (-0,296)	1,297 (-0,628)
Ordnung 11	-0,117 (-0,165)	0,277 (-0,117)
Ordnung 13	-0,589 (-0,366)	0,562 (-0,547)
Ordnung 15*	-1,078 (-0,847)	1,326 (-0,386)
Ordnung 17*	-1,454 (-0,658)	2,887 (-0,348)
Ordnung 19	-1,153 (-0,470)	1,092 (-0,618)
Ordnung 21	-0,383 (-0,176)	0,258 (-0,196)
Ordnung 23	-0,408 (-0,835)	0,972 (0,359)
Ordnung 25*	-1,157 (-0,603)	1,399 (-0,565)
Ordnung 27	-0,258 (-0,088)	-0,010 (-0,811)
Ordnung 29	-0,265 (-0,675)	-0,224 (-0,153)
Fleiß 2	-0,423 (0,157)	2,273 (-0,800)
Fleiß 4	-0,328 (-0,282)	0,529 (-0,373)
Fleiß 6	-0,148 (0,177)	0,520 (0,308)
Fleiß 8	-0,174 (-0,202)	-0,765 (-0,712)
Fleiß 10	-0,474 (-0,116)	0,738 (-0,183)
Fleiß 12	-0,293 (-0,234)	0,209 (-0,519)
Fleiß 14	-0,207 (-0,899)	-0,412 (0,858)
Fleiß 16	-1,027 (-0,460)	2,953 (0,092)
Fleiß 18	-0,018 (-0,278)	-0,193 (-0,433)
Fleiß 20	-0,265 (-0,593)	0,270 (0,378)
Fleiß 22	-0,155 (0,068)	0,408 (-0,308)
Fleiß 24	-0,004 (-0,102)	-0,134 (-0,714)
Fleiß 26	-0,275 (-0,646)	0,317 (0,275)
Fleiß 28	-0,345 (-0,337)	0,285 (0,466)
Fleiß 30	-0,256 (0,117)	-0,414 (-0,469)

Anmerkungen: *c.r.*= critical ratio; N_{Bew} = 740 Ausbildungsbewerber bzw. N_{KG} = 97 Kontrollgruppe; Werte der Kontrollgruppe stehen in Klammern; * umgepolte Items wurden vor der Analyse rekodiert.

Den nachfolgenden Tabellen A-12 und A-13 können die Interitemkorrelationen der Kollinearitätsanalysen entnommen werden. Sowohl in der Bewerberstichprobe (N_{Bew} = 740) als auch in der Kontrollgruppe (N_{KG} = 97) waren sämtliche Korrelationen kleiner als $r= 0,85$, weshalb davon ausgegangen wurde, dass keine Multikollinearität in den Daten vorlag.

Tabelle A-12: Iteminterkorrelationen der Skala Fleiß

	F2	F4	F6	F8	F10	F12	F14	F16	F18	F20	F22	F24	F26	F28	F30
F2	1														
F4	.27** (-.11)	1													
F6	.23** (.14)	.23** (.14)	1												
F8	.25** (.25*)	.23** (.03)	.20** (.19)	1											
F10	.25** (.38**)	.28** (-.04)	.30** (.29**)	.29** (.21*)	1										
F12	.30** (.36**)	.26** (.27*)	.24** (.22*)	.36** (.20*)	.29** (.38**)	1									
F14	.25** (.27**)	.24** (-.03)	.19** (.07)	.40** (.37**)	.22** (.12)	.32** (.12)	1								
F16	.26** (.34**)	.22** (.05)	.25** (.15)	.30** (.02)	.31** (.32**)	.45** (.42**)	.26** (.23)	1							
F18	.30** (.35**)	.25** (-.06)	.23** (.19)	.35** (.09)	.24** (.30)	.47** (.38**)	.39** (.26**)	.34** (.31**)	1						
F20	.33** (.11)	.27** (.18)	.21** (.22*)	.33** (.18)	.26** (.12)	.34** (.33**)	.26** (.24*)	.25** (.15)	.31** (.30**)	1					
F22	.32** (.14)	.38** (.18)	.26** (.31**)	.29** (.24*)	.32** (.28**)	.32** (.39**)	.28** (.21*)	.29** (.33**)	.32** (.29**)	.33** (.41**)	1				
F24	.29** (.13)	.29** (.00)	.24** (.22*)	.53** (.44**)	.33** (.21*)	.43** (.23*)	.46** (.30**)	.37** (.17)	.45** (.31**)	.28** (.28**)	.38** (.34**)	1			
F26	.17** (.24*)	.19** (.03)	.09* (.12)	.30** (.24**)	.16** (.35**)	.24** (.27**)	.24** (-.05)	.13** (.09)	.23** (.33**)	.19** (.23*)	.23** (.20)	.33** (.19)	1		
F28	.36** (.28**)	.34** (.03)	.27** (.05)	.34** (.09)	.28** (.22*)	.44** (.31**)	.35** (.11)	.33** (.24*)	.40** (.31**)	.31** (.14)	.36** (.12)	.43** (.16)	.29** (.18)	1	
F30	.45** (.58**)	.34** (-.00)	.20** (.19)	.29** (.22*)	.23** (.39**)	.37** (.43**)	.27** (.38**)	.31** (.34**)	.33** (.31**)	.37** (.26*)	.36** (.33**)	.37** (.35**)	.19** (.14)	.37** (.17)	1

Anmerkungen: F2= Item 2, der Fleiß-Skala zugeordnet; F4= Item 4, der Fleiß-Skala zugeordnet; ...; F30= Item 30, der Fleiß-Skala zugeordnet; ** $p < .01$ (zweiseitig); * $p < .05$ (zweiseitig); $N_{Bew} = 740$ (Bewerberstichprobe) bzw. $N_{KG} = 97$ (Kontrollgruppe); die Werte der Kontrollgruppe stehen in Klammern.

Tabelle A-13: Iteminterkorrelationen der Skala Ordnung

	O1	O3	O5	O7	O9 ^a	O11	O13	O15 ^a	O17 ^a	O19	O21	O23	O25 ^a	O27	O29
O1	1														
O3	.15** (.23*)	1													
O5	.42** (.27**)	.19** (.29**)	1												
O7	.04 (.08)	.19** (.05)	.08* (.16)	1											
O9^a	.04 (.01)	.19** (.01)	.08* (.09)	.15** (-.06)	1										
O11	.12** (-.02)	.35** (.31**)	.15** (.10)	.14** (.01)	.16** (.13)	1									
O13	.18** (.22*)	.53** (.57**)	.23** (.18)	.15** (.01)	.21** (.19)	.30** (.26*)	1								
O15^a	.10** (.11)	.64** (.57**)	.21** (.20)	.22** (-.02)	.23** (.18)	.33** (.27**)	.55** (.52**)	1							
O17^a	.14** (.11)	.34** (.16)	.10** (-.14)	.09* (-.06)	.25** (.17)	.23** (.17)	.31** (.19)	.42** (.18)	1						
O19	.08* (.13)	.24** (.19)	.09* (-.04)	.10** (.11)	.01 (-) (.05)	.13** (.22*)	.25** (.15)	.23** (.27**)	.09* (.08)	1					
O21	.27** (.03)	.20** (.23*)	.16** (.10)	.13** (-.01)	.25** (.09)	.19** (.23*)	.19** (.20)	.16** (.22*)	.17** (.10)	.00 (.02)	1				
O23	.16** (.11)	.30** (.56**)	.19** (.38**)	.20** (.17)	.32** (.13)	.26** (.23*)	.26** (.33**)	.31** (.37**)	.19** (.19)	.08* (.10)	.41** (.24*)	1			
O25^a	.16** (.16)	.39** (.52**)	.14** (.14)	.08* (.05)	.19** (-.00)	.22** (.24**)	.37** (.31**)	.46** (.37**)	.34** (.21*)	.17** (.18)	.16** (-.01)	.23** (.35**)	1		
O27	.38** (.23*)	.22** (.40**)	.39** (.24*)	.05 (-) (.04)	.11** (.13)	.25** (.16)	.27** (.22*)	.22** (.34**)	.15** (.18)	.13** (.09)	.25** (.21*)	.24** (.28**)	.23** (.33**)	1	
O29	.15** (.43**)	.09* (.01)	.12** (.18)	.09* (.17)	.08* (.19)	.02 (-) (.13)	.10** (.01)	.09* (-.12)	.02 (.21*)	.01 (.06)	.23** (.19)	.16** (.06)	.10** (-.01)	.04 (.20*)	1

Anmerkungen: O1= Item 1, der Ordnungs-Skala zugeordnet; O3= Item 3, der Ordnungs-Skala zugeordnet; ...; O29= Item 29, der Ordnungs-Skala zugeordnet; ^a = umcodierte Items; ** $p < .01$ (zweiseitig); * $p < .05$ (zweiseitig); $N_{Bew} = 740$ (Bewerberstichprobe) bzw. $N_{KG} = 97$ (Kontrollgruppe); die Werte der Kontrollgruppe stehen in Klammern.

A4. Ergebnisse der konfirmatorischen Faktorenanalyse zur Skalenüberprüfung der AGS

Die Abbildungen A-1 bis A-4 zeigen komplett standardisierten Lösungen der einfaktoriellen CFA-Modelle für die Gewissenhaftigkeitsaspekte Fleiß und Ordnung für die Stichprobe der Bewerber ($N_{Bew} = 740$) sowie der Schüler-Kontrollgruppe ($N_{KG} = 97$). Auf den Pfeilen sind die Angaben zu den Ladungen bzw. partiell standardisierten Regressionsgewichten abgetragen. Da jedes Item nur auf einem Faktor lädt, entsprechen die Ladungen in diesen Modellen Korrelationen.

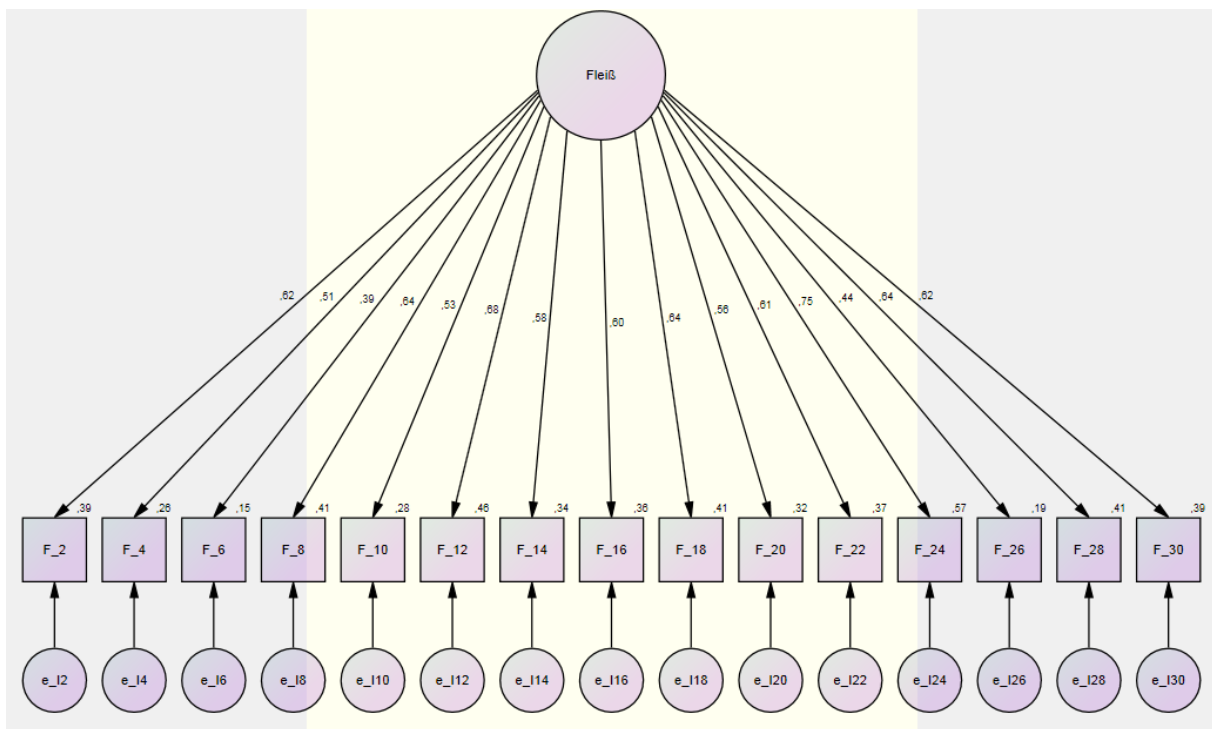


Abbildung A-1: AMOS-Ausgabe des einfaktoriellen CFA-Modells des Fleißaspekts für die Bewerberstichprobe ($N_{Bew} = 740$).

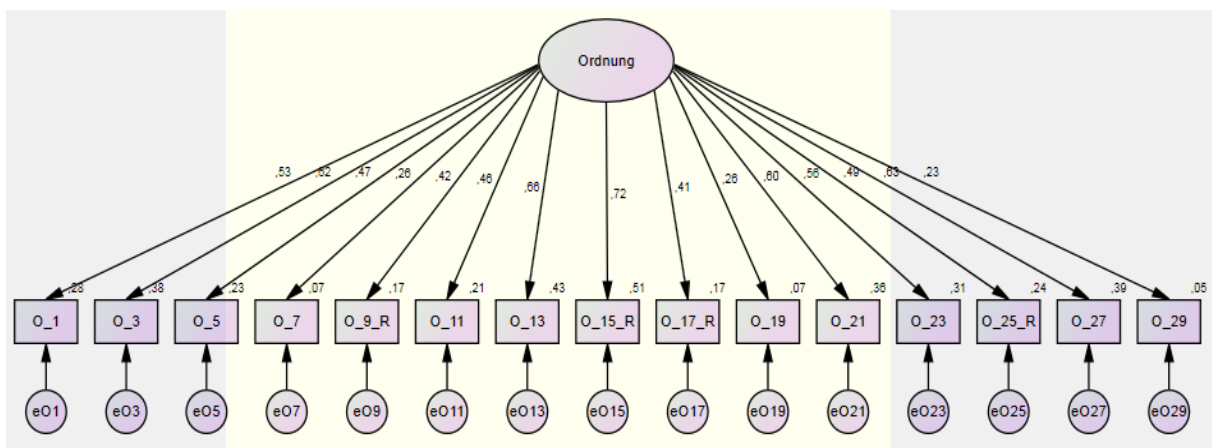


Abbildung A-2: AMOS-Ausgabe des einfaktoriellen CFA-Modells des Ordnungsaspekts für die Bewerberstichprobe ($N_{Bew} = 740$).

Die versetzt auf den beobachteten Variablen ($\square$) stehenden Zahlen (z.B. .39 bei F_2) sind die multiplen quadrierten Korrelationskoeffizienten (R^2), die anzeigen, wie viel Varianz eines Items durch die latente Variable erklärt werden kann. Diese Werte entsprechen in diesem Falle somit auch den Kommunalitäten (h^2) der Items, die eine Mindestschätzung der

Itemreliabilität darstellen. Wenn man diese von 1 abzieht, erhält man somit die standardisierte Fehlervarianz. Die Kommunalitätsschätzungen sind allerdings deutlich niedriger als man es von einer annehmbaren Reliabilität erwarten würde. Dies ist Bühner (2011) zufolge jedoch der Tatsache geschuldet, dass die Reliabilität eines einzelnen Items nicht so hoch ausfallen kann wie die eines ganzen Tests oder einer Subskala eines Tests. Auch die Schiefe des jeweiligen Items spielt hier eine Rolle. Da in jedem Testverfahren nicht nur normalverteilte Items, sondern auch rechts- oder linkssteile Items enthalten sind, hat dies ebenfalls Einfluss auf die Itemreliabilität.

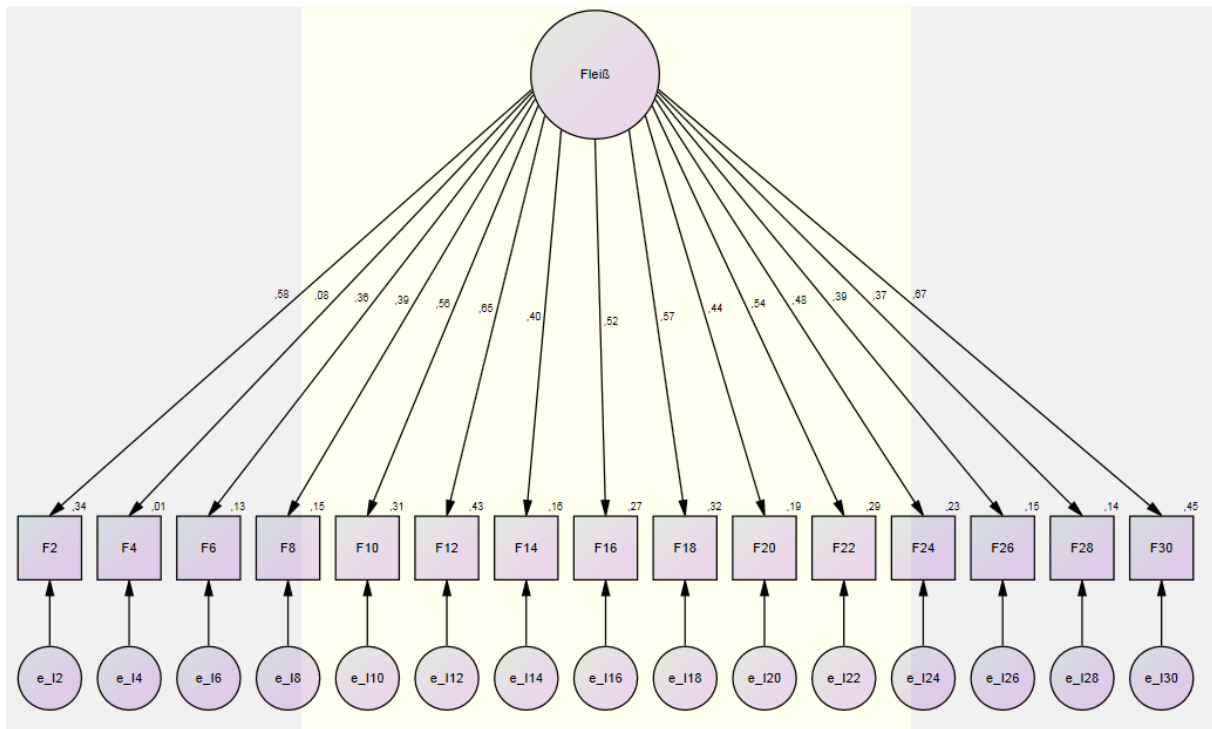


Abbildung A-3: AMOS-Ausgabe des einfaktoriellen CFA-Modells des Fleißaspekts für die Kontrollgruppe ($N_{KG}= 97$).

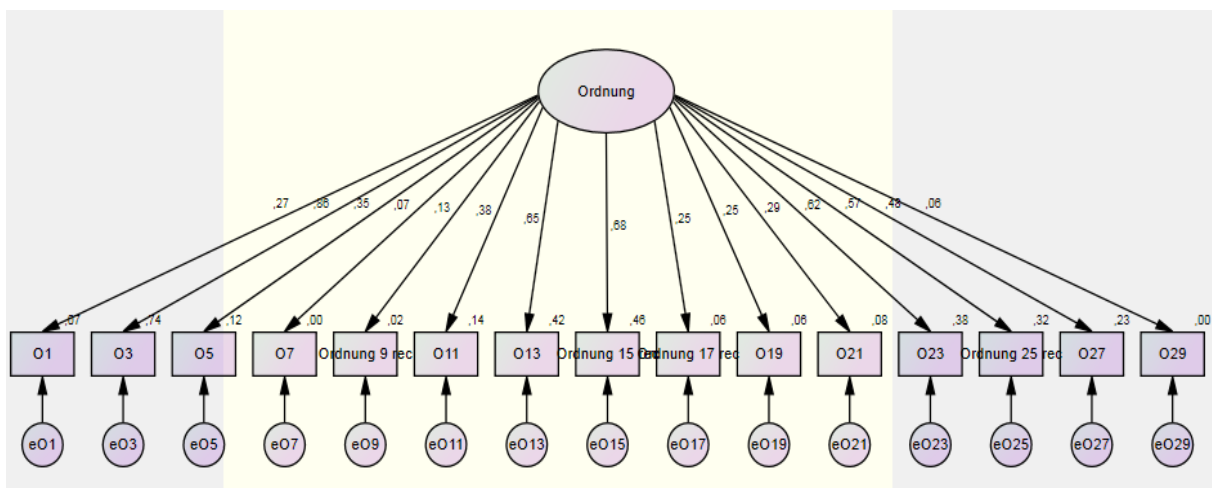


Abbildung 9: AMOS-Ausgabe des einfaktoriellen CFA-Modells des Ordnungsaspekts für die Kontrollgruppe ($N_{KG}= 97$).

Tabelle A-14 gibt nachfolgend einen Überblick über die standardisierten Ladungen bzw. standardisierten semipartiellen Regressionsgewichte und Fehlervarianzen der AGS für beide Stichproben.

Keine der standardisierten Ladungen liegt über 1, was sowohl für die Bewerberstichprobe ($N_{Bew}=740$) als auch für die Kontrollgruppe ($N_{KG}=97$) zutrifft. Ladungen über 1 würden auf Schätzprobleme hindeuten. Aufgrund der vorliegenden Ergebnisse können Schätzprobleme jedoch ausgeschlossen werden. Insgesamt sind die Ladungen in der Kontrollgruppe etwas niedriger ausgeprägt als in der Bewerberstichprobe. Da hier eine einfaktorielle Lösung vorliegt, entsprechen die standardisierten Ladungen den Korrelationen zwischen Item und Faktor. Die Mustermatrix ist also identisch zur Strukturmatrix.

Tabelle A-14: Standardisierte Ladungen und Fehlervarianzen der AGS

Itemzuordnung	Ladungen	Estimate	S.E.	C.R.	P
Ordnung	-	0,030 (0,003)	0,011 (0,011)	2,669 (0,261)	0,008 (0,794)
Ordnung 29	0,234 (0,057)	0,520 (0,904)	0,026 (0,131)	20,172 (6,925)	*** (***)
Ordnung 27	0,626 (0,484)	0,245 (0,633)	0,018 (0,096)	13,334 (6,593)	*** (***)
Ordnung 25*	0,493 (0,569)	0,560 (0,930)	0,056 (0,145)	9,954 (6,397)	*** (***)
Ordnung 23	0,556 (0,620)	0,240 (0,521)	0,017 (0,084)	13,914 (6,233)	*** (***)
Ordnung 21	0,599 (0,290)	0,258 (0,471)	0,019 (0,069)	13,266 (6,828)	*** (***)
Ordnung 19	0,256 (0,248)	0,961 (1,114)	0,070 (0,162)	13,747 (6,857)	*** (***)
Ordnung 17*	0,412 (0,250)	0,446 (1,184)	0,049 (0,173)	9,175 (6,856)	*** (***)
Ordnung 15*	0,717 (0,679)	0,137 (0,817)	0,015 (0,137)	9,426 (5,958)	*** (***)
Ordnung 13	0,659 (0,647)	0,211 (0,548)	0,017 (0,090)	12,098 (6,120)	*** (***)
Ordnung 11	0,464 (0,380)	0,214 (0,400)	0,011 (0,059)	19,313 (6,744)	*** (***)
Ordnung 9*	0,417 (0,131)	0,404 (1,033)	0,034 (0,149)	12,018 (6,909)	*** (***)
Ordnung 7	0,257 (0,069)	0,148 (0,836)	0,012 (0,121)	12,153 (6,923)	*** (***)
Ordnung 5	0,475 (0,348)	0,394 (0,777)	0,024 (0,115)	16,226 (6,778)	*** (***)
Ordnung 3	0,617 (0,858)	0,206 (0,317)	0,019 (0,082)	10,783 (3,846)	*** (***)
Ordnung 1	0,533 (0,272)	0,233 (0,542)	0,017 (0,079)	14,053 (6,842)	*** (***)
Fleiß	-	0,128 (0,236)	0,012 (0,069)	10,269 (3,412)	*** (***)
Fleiß 30	0,624 (0,667)	0,200 (0,294)	0,011 (0,050)	18,821 (5,846)	*** (***)
Fleiß 28	0,644 (0,374)	0,247 (0,518)	0,015 (0,077)	16,610 (6,710)	*** (***)
Fleiß 26	0,438 (0,386)	0,294 (0,645)	0,020 (0,096)	14,837 (6,692)	*** (***)
Fleiß 24	0,752 (0,480)	0,128 (0,513)	0,009 (0,079)	14,960 (6,525)	*** (***)
Fleiß 22	0,610 (0,541)	0,204 (0,380)	0,012 (0,060)	16,982 (6,370)	*** (***)
Fleiß 20	0,562 (0,439)	0,234 (0,432)	0,015 (0,065)	15,202 (6,607)	*** (***)
Fleiß 18	0,642 (0,567)	0,166 (0,339)	0,011 (0,054)	14,917 (6,289)	*** (***)
Fleiß 16	0,599 (0,515)	0,254 (0,629)	0,028 (0,098)	9,163 (6,442)	*** (***)
Fleiß 14	0,582 (0,402)	0,210 (0,619)	0,012 (0,093)	17,490 (6,669)	*** (***)
Fleiß 12	0,677 (0,653)	0,201 (0,393)	0,015 (0,066)	13,567 (5,924)	*** (***)
Fleiß 10	0,526 (0,561)	0,313 (0,463)	0,021 (0,073)	15,144 (6,311)	*** (***)
Fleiß 8	0,644 (0,393)	0,186 (0,580)	0,011 (0,087)	17,583 (6,683)	*** (***)
Fleiß 6	0,388 (0,358)	0,374 (0,526)	0,022 (0,078)	16,772 (6,731)	*** (***)
Fleiß 4	0,514 (0,082)	0,252 (0,694)	0,014 (0,100)	18,046 (6,919)	*** (***)
Fleiß 2	0,623 (0,581)	0,174 (0,350)	0,014 (0,056)	12,700 (6,242)	*** (***)

Anmerkungen: Estimate= nichtstandardisierte Fehlervarianz; S.E.= Standardfehler; C.R.= entspricht z-Wert; P= Signifikanz; *** $P < 0,000$; $N_{Bew}=740$ Ausbildungsbewerber bzw. $N_{KG}=97$ Kontrollgruppe; Werte der Kontrollgruppe stehen in Klammern; * umgepolte Items wurden vor der Analyse rekodiert.

Die nichtstandardisierten Fehlervarianzen sind sowohl in der Bewerberstichprobe ($N_{Bew}=740$) als auch in der Kontrollgruppe ($N_{KG}=97$) signifikant von null verschieden. Das heißt, ein beträchtlicher Varianzanteil der Items lässt sich entweder itemspezifisch oder durch andere Faktoren erklären. Möglich ist auch, dass diese unsystematisch sind. Zusätzlich hat der Aspekt Fleiß in beiden Stichproben eine signifikante Varianz, was für lokalen Modell-Fit spricht. Für den Ordnungsaspekt zeigt sich nur in der Bewerberstichprobe ($N_{Bew}=740$) eine signifikante Varianz. In der Kontrollgruppe ($N_{KG}=97$) wird diese nicht signifikant.

Tabelle A-15: Quadrierte multiple Korrelationen

Itemzuordnung	R^2
Ordnung 1	0,284 (0,074)
Ordnung 3	0,380 (0,736)
Ordnung 5	0,225 (0,121)
Ordnung 7	0,066 (0,005)
Ordnung 9*	0,174 (0,017)
Ordnung 11	0,215 (0,145)
Ordnung 13	0,435 (0,419)
Ordnung 15*	0,515 (0,462)
Ordnung 17*	0,170 (0,062)
Ordnung 19	0,066 (0,061)
Ordnung 21	0,359 (0,084)
Ordnung 23	0,310 (0,384)
Ordnung 25*	0,243 (0,324)
Ordnung 27	0,392 (0,234)
Ordnung 29	0,055 (0,003)
Fleiß 2	0,388 (0,338)
Fleiß 4	0,264 (0,007)
Fleiß 6	0,151 (0,128)
Fleiß 8	0,414 (0,154)
Fleiß 10	0,277 (0,314)
Fleiß 12	0,458 (0,427)
Fleiß 14	0,338 (0,162)
Fleiß 16	0,359 (0,265)
Fleiß 18	0,412 (0,322)
Fleiß 20	0,316 (0,193)
Fleiß 22	0,372 (0,293)
Fleiß 24	0,565 (0,231)
Fleiß 26	0,192 (0,149)
Fleiß 28	0,415 (0,140)
Fleiß 30	0,390 (0,445)

Anmerkungen: R^2 = quadrierte multiple Korrelation; N_{Bew} = 740 Ausbildungsbe-
werber bzw. N_{KG} = 97 Kontroll-
gruppe; Werte der Kontrollgrup-
pe stehen in Klammern; * umge-
polte Items wurden vor der Ana-
lyse rekodiert.

Tabelle A-15 enthält für beide Stichproben die Varianzanteile der Items, die durch die latente Variable erklärt werden, was auch als Kommunalität bezeichnet werden kann. Bühner (2011) zufolge spricht man von einer niedrigen Kommunalität, wenn diese kleiner als 13 % ist. Dies traf in der Bewerberstichprobe (N_{Bew} = 740) auf die Items O7, O19 und O29 des Ordnungsaspektes zu. In der Kontrollgruppe der Schüler (N_{KG} = 97) waren hiervon die Items O1, O5, O9, O17, O19, O21 und O29 des Ordnungsaspektes sowie F4 und F6 des Fleißaspektes betroffen. Dies deutete v.a. für den Fleißaspekt darauf hin, dass es sich um einen eher homogenen Faktor handelte. Bezüglich der Subskala Ordnung gab es in der Bewerberstichprobe vergleichsweise höhere Kommunalitäten. Aber auch in der Kontrollgruppe verfügte die Mehrheit der Items über Kommunalitäten < 13 %. Daher ist davon auszugehen, dass es sich auch bei dieser Subskala um einen homogenen Faktor handelte.

Allerdings wiesen die Items auch relativ hohe Fehlervarianzen ($1-R^2$) auf. Das heißt, der Anteil der unerklärten Varianz, die nicht durch die latente Variable erklärt wurde, war vergleichsweise hoch.

Tabelle A-16: Interkorrelationstabelle Retest-Stichprobe gesamt

	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Prädiktoren T1																
1 T1 Fleiß	3,10	.35														
2 T1 Ordnung	3,13	.39	.47**													
			(.22)													
3 T1 Selbstwirksamkeit	3,06	.35	.79**	.46**												
			(.62)	(.22)												
4 T1 GMA	61,60	4.55	-.12	-.05	-.01											
			(.01)	(.00)	(.00)											
Prädiktoren T2																
5 T2 Fleiß	3,07	.32	.51**	.19	.41**	-.05										
			(.26)	(.03)	(.17)	(.00)										
6 T2 Ordnung	3,08	.35	.17	.46**	.25*	.04	.30*									
			(.03)	(.22)	(.06)	(.00)	(.09)									
7 T2 Selbstwirksamkeit	2,86	.34	.33**	.02	.50**	.06	.65**	.22								
			(.11)	(.00)	(.25)	(.00)	(.42)	(.05)								
8 T2 PA	1,51	.33	-.41**	-.27	-.45**	-.20	-.32*	-.21	-.43**							
			(.17)	(.07)	(.21)	(.04)	(.10)	(.04)	(.18)							
Kriterien																
9 Schriftliche Prüfung ^a	81,69	8.72	.16	.07	.07	-.01	.09	-.23	.04	-.04						
			(.03)	(.00)	(.00)	(.00)	(.01)	(.05)	(.00)	(.00)						
10 Praktische Prüfung ^a	79,77	11.69	.27	-.04	.01	-.29	.26	.18	.15	-.46*	.36					
			(.07)	(.00)	(.00)	(.08)	(.07)	(.03)	(.02)	(.21)	(.13)					
11 Gesamtnote ^b	81,07	8.60	.27	-.04	.05	-.30	.29	-.06	.16	-.27	.78**	.85**				
			(.07)	(.00)	(.00)	(.09)	(.08)	(.00)	(.03)	(.07)	(.60)	(.73)				
12 Allg. Berufsschulnote ^c	2,07	.48	-.11	-.07	-.14	-.14	.02	-.05	-.04	.37**	-.66**	-.51**	-.69**			
			(.01)	(.00)	(.02)	(.02)	(.00)	(.00)	(.00)	(.14)	(.43)	(.26)	(.48)			
13 VG-Beurteilung ^c	2,92	.89	-.27*	.01	-.21	-.34**	-.07	.11	-.14	.38**	-.60**	-.56**	-.64**	.52**		
			(.07)	(.00)	(.04)	(.11)	(.00)	(.01)	(.02)	(.15)	(.36)	(.32)	(.43)	(.27)		
Kontrollvariablen																
14 Soziale Erwünschtheit	12,46	2.75	.29*	.35**	.40**	-.14	.43**	.38**	.34**	-.09	.24	-.07	-.02	-.02	.10	
			(.09)	(.12)	(.16)	(.02)	(.19)	(.15)	(.11)	(.01)	(.06)	(.00)	(.00)	(.00)	(.01)	
15 Alter	17,31	1.95	.14	.07	.29*	.18	.01	-.08	.12	-.20	.25	-.25	.09	-.43**	-.39**	.10
			(.02)	(.00)	(.09)	(.03)	(.00)	(.01)	(.01)	(.04)	(.06)	(.06)	(.01)	(.18)	(.15)	(.01)

Anmerkungen: $N_R=63$; $n_{IHK}=26$ Zwischenprüfungsnoten; $n_{IHK}=30$ Zwischenprüfungsgesamtnoten; $n_{BS,VG}=62$ Berufsschulnoten und Vorgesetztenbeurteilungen; *M*= Mittelwert; *SD*= Standardabweichung; PA= wahrgenommene eigene Fähigkeiten; GMA= Allgemeine kognitive Fähigkeiten; VG-Beurteilung= Vorgesetztenbeurteilung; ** $p < 0,01$ (zweiseitig); * $p < 0,05$ (zweiseitig); Werte in Klammern entsprechen r^2 .

A5. Weitere Ergebnisse zur Überprüfung des Vorhersagemodells

Auch ein möglicher medierender Effekt des Ordnungsaspekts zwischen Fleiß und den verschiedenen Kriteriumsvariablen wurde überprüft.

In der Subgruppe der kaufmännischen Auszubildenden interkorrelierten die Gewissenhaftigkeitsaspekte signifikant miteinander ($r = .38, p < .05$). Obwohl der Fleißaspekt keine signifikanten Korrelationen mit den Kriteriumsvariablen Vorgesetztenbeurteilung ($r = -.11, n.s.$) und allgemeine Berufsschulleistung ($r = -.03, n.s.$) zeigte, wurden die Mediationsanalysen dennoch berechnet. Die Ergebnisse der Mediationsanalyse für das Kriterium der Vorgesetztenbeurteilung sind in Tabelle A-17 dargestellt. Es zeigte sich, dass Fleiß allein kein guter Prädiktor der Vorgesetztenbeurteilung war ($\beta = -.252, t(28) = -0,576, p = .57$). Ein medierender Effekt des Ordnungsaspektes konnte nicht gefunden werden ($\beta = .285, t(28) = 1,456, p = .16$), auch wenn das Mediationsmodell tendenziell besser auf die Daten passte. Dieses Ergebnis wird durch den ebenfalls nicht-signifikanten Sobel-Test unterstrichen ($z = -0,54$). Bezüglich des

Tabelle A-17: Ergebnisse der Mediationsanalyse für Fleiß und Ordnung zur Vorhersage der Vorgesetztenbeurteilung bei den kaufmännischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-.252	.437	-.106	-0,576	-.505	.462	-.213	-1,092
Ordnung					.631	.434	.285	1,456

Anmerkungen: $N_{kf} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; $^{\dagger} p < .10$; * $p < .05$; ** $p < .01$.

Kriteriums der allgemeinen Berufsschulnote erwies sich der Fleißaspekt allein mit $\beta = -.025, t(24) = -0,124, p = .90$ wiederum nicht als guter Prädiktor. Die Modellpassung des Mediationsmodells deutete demgegenüber tendenziell auf einen partiell medierenden Effekt hin ($\beta = -.384, t(24) = -1,719, p < .10$). Der Sobel-Test wurde nicht signifikant ($z = 0,12$). Die Ergebnisse in der kaufmännischen Substichprobe deuten tendenziell auf eine partielle Mediation des Ordnungsaspektes hin.

In der Subgruppe der technischen Auszubildenden zeigte sich ein gegensätzlicher Effekt. Fleiß und Ordnung korrelierten auch in dieser Teilstichprobe signifikant miteinander ($r = .52, p < .01$). Fleiß erwies sich jedoch nicht als signifikanter Prädiktor der Kriteriumsvariablen Vorgesetztenbeurteilung ($r = -.34, n.s.$), allgemeine Berufsschulleistung ($r = -.12, n.s.$), schriftliche ($r = .16, n.s.$) oder praktische Zwischenprüfungsergebnisse ($r = .27, n.s.$). Aufgrund der Tendenzen wurden die Mediationsanalysen dennoch berechnet. Für das Kriterium der Vorgesetztenbeurteilung erwies sich Ordnung nicht als signifikanter Mediator des Zusammenhangs zwischen Fleiß und dem Kriterium (s. Tabelle A-18). Vielmehr erschien eher Fleiß allein

Tabelle A-18: Ergebnisse der Regressionsanalyse für Fleiß und Ordnung zur Vorhersage der Vorgesetztenbeurteilung bei den technischen Auszubildenden

Variable	Model 1				Model 2			
	B	SE B	β	t	B	SE B	β	t
Fleiß	-0,790	-.406	-.340	-1,945 [†]	-1,050	.465	-.451	-2,260*
Ordnung					0,450	.396	.227	1,135

Anmerkungen: $N_{techn} = 30$; B= Regressionskoeffizient; SE B= Standardfehler des Koeffizienten; β = standardisiertes Regressionsgewicht; $^{\dagger} p < .10$; * $p < .05$; ** $p < .01$.

ein guter Prädiktor zu sein. Der Sobel-Test bestätigte dieses Ergebnis ($z = -0,98$). Für das Kriterium der allgemeinen Berufsschulnote führte die Hinzunahme von Ordnung als Mediator ebenfalls zu keiner signifikant besseren Modellpassung des Mediationsmodells ($\beta = .317$,

$t(32) = 1,487$, $p = .15$) gegenüber Model 1 ($\beta = -.124$, $t(32) = -0,720$, $p = .48$). Allerdings passte das Mediationsmodell etwas besser. Dennoch wurde die Mediation laut Sobel-Test nicht signifikant ($z = -0,65$). Hinsichtlich der IHK-Zwischenprüfungsergebnisse zeigte sich mit $\beta = -.363$, $t(23) = -1,448$, $p = .16$ und $z = -0,24$ sowohl für den praktischen Prüfungsteil als auch mit $\beta = -.067$, $t(23) = -0,251$, $p = .80$ und $z = -0,99$ für den schriftlichen Prüfungsteil kein medierender Effekt des Ordnungsaspektes. Insgesamt konnte somit in der Substichprobe der technischen Auszubildenden davon ausgegangen werden, dass Ordnung nicht als Mediator des Zusammenhangs zwischen Fleiß und den Leistungskriterien fungierte.

Für die Kontrollgruppe ergab sich mit $\beta = -.244$, $t(66) = -2,000$, $p < .05$ eine bessere Modellpassung für das Mediationsmodell im Vergleich zum Alternativmodell mit $\beta = -.243$, $t(66) = -2,048$, $p < .05$. Für die Kontrollgruppe musste somit ein partiell medierender Effekt des Ordnungsaspekts angenommen werden. Der Sobel-Test wurde knapp nicht signifikant ($z = -1,76$).

B1. Anschreiben an die Retest Stichprobe



Universität Hamburg
Fakultät für Erziehungswissenschaft,
Psychologie und Bewegungswissenschaft



MOLDZIO & PARTNER
INSTITUT FÜR PERSONALAUSWAHL

MOLDZIO & PARTNER · STOLBERGSTR. 13 · 22967 TREMSBÜTTEL

An die Teilnehmer/-innen
des Ausbildungswochenendes

THOMAS MOLDZIO · DIPL.-PSYCHOLOGE
DR. BERND SOBOTKA · DIPL.-PSYCHOLOGE
MATTHIAS MCKEILEIT · DIPL.-PSYCHOLOGE

STOLBERGSTR. 13 · 22967 TREMSBÜTTEL
TELEFON (0 45 32) 26 61 - 0

TELEFAX (0 45 32) 26 61 - 21

WWW.MOLDZIO.COM

KONTAKT@MOLDZIO.COM

Tremsbüttel, September 2010
KD

Erwartungen an die Ausbildung bei [REDACTED]

Sehr geehrte Auszubildende, sehr geehrter Auszubildender,

vor ungefähr einem Jahr durchliefen Sie den Auswahlprozess bei [REDACTED]. Im Rahmen des Bewerbungstages erhielten Sie von Moldzio & Partner – Institut für Personalauswahl eine Rückmeldung Ihrer Fragebogen- und Testergebnisse des Testtags, an dem Sie im Zuge des Auswahlverfahrens bei [REDACTED] teilnahmen.

[REDACTED] verfolgt im Rahmen dieses Auswahlprozesses das Ziel, mittels wissenschaftlich überprüfter Auswahlinstrumente zuverlässige Prognosen über den zukünftigen Ausbildungserfolg seiner Bewerber zu treffen. Getreu dem Unternehmensmotto [REDACTED] wird auch das Auswahlverfahren bei [REDACTED] stetig weiterentwickelt. Deshalb führen wir als beratendes Institut in Verbindung mit der Universität Hamburg eine wissenschaftliche Untersuchung durch, um einen neu entwickelten Fragebogen hinsichtlich seiner Prognosekraft zu überprüfen.

Es handelt sich bei dem zu überprüfenden Instrument um die "Arbeitsbezogenen Gewissenhaftigkeitsskalen" (AGS), die Sie bereits am Testtag vor einem Jahr bearbeitet haben. Heute möchten wir Sie bitten, die AGS erneut sowie einige weitere, kurze Fragebögen zu Ihren Erwartungen an die Ausbildung bei [REDACTED] auszufüllen.

Hierzu folgen Sie bitte den schriftlichen Anweisungen, die auf den folgenden Seiten enthalten sind. Beantworten Sie alle Fragen vollständig und ehrlich. Es gibt keine richtigen oder falschen Antworten. Nachdem Sie alle Seiten der Befragung vollständig ausgefüllt haben, stecken diese in den dafür vorgesehenen Umschlag, den Sie bitte zugeklebt in den dafür vorgesehenen Karton legen.

Die gesetzlichen Datenschutzbestimmungen (§ 203 StGB) werden von uns selbstverständlich eingehalten. Rückschlüsse auf Einzelpersonen sind nicht möglich und nicht beabsichtigt. Eine gesonderte Rückmeldung über einzelne Ergebnisse an [REDACTED] wird es nicht geben. Die Ergebnisse dieser Studie fließen in die Datenanalyse zur Gesamtevaluation der AGS ein.

Die Untersuchung wird nur dann erfolgreich sein, wenn möglichst alle eingestellten Bewerber teilnehmen. Deshalb bitten wir auch Sie um Ihre Unterstützung! Für die Beantwortung des beiliegenden Fragebogenpakets benötigen Sie ca. 15 Minuten Zeit.

Für eventuelle Fragen stehen wir Ihnen unter der o.g. Telefonnummer gern zur Verfügung und bedanken uns schon vorab für Ihre Mitarbeit.

Mit freundlichen Grüßen

Kirsten Dreier

COMMERZBANK HAMBURG
BLZ 200 400 00 · KTO.-NR. 495 14 14
SPARKASSE DETMOLD
BLZ 476 501 30 · KTO.-NR. 460 666 84
AMTSGERICHT KIEL PR 211 KI

B2. Beispiel-Anschreiben zur Rekrutierung der Kontrollstichprobe

MOLDZIO & PARTNER · STOLBERGSTR. 13 · 22967 TREMSBÜTTEL

Herrn
[REDACTED]
Schulleiter
[REDACTED]
[REDACTED]
[REDACTED]

Wissenschaftliche Untersuchung am [REDACTED]

Sehr geehrter Herr [REDACTED],

wie Herr [REDACTED] Ihnen bereits berichtete, suche ich den Kontakt zu Ihnen, da ich an einer wissenschaftlichen Untersuchung des aktuellen Abiturjahrganges am Gymnasium [REDACTED] interessiert bin.

Ich promoviere in Kooperation mit der Universität Hamburg und Moldzio & Partner – Institut für Personalauswahl zum Thema Gewissenhaftigkeit im Arbeitskontext. Das Ziel meiner Dissertation ist unter anderem die Überprüfung einer arbeitsbezogenen Gewissenhaftigkeitsskala sowie eines theoretischen Modells zur Vorhersage von Ausbildungserfolg. Hierzu kooperiere ich bereits mit einem [REDACTED], dessen Ausbildungsbewerber ich durch den Auswahl- und Berufsschulprozess begleite. Da es sich hierbei jedoch um eine hochselektive Stichprobe handelt, benötige ich eine vergleichbare Kontrollgruppe, um die Aussagekraft meiner Ergebnisse herstellen zu können. Diese Kriterien erfüllt Ihr aktueller Abiturjahrgang, den ich diesbezüglich gern untersuchen würde.

Hauptberuflich bin ich als wissenschaftliche Mitarbeiterin und Assistentin der Geschäftsleitung bei Moldzio & Partner angestellt, wo ich neben meiner Promotion u.a. für die Koordination der wissenschaftlichen Arbeiten verantwortlich bin. Der Fokus der Unternehmensberatung liegt insbesondere auf der Entwicklung und Durchführung von Verfahren zur Personalauswahl und Potenzialerkennung. Zudem sind Qualifizierungen von Führungskräften wie beispielsweise Interviewtrainings in der Palette der Dienstleistungen enthalten. Erste Eindrücke über das Unternehmen können Sie sich durch die beiliegende Broschüre und einen Blick ins Internet unter www.moldzio.com verschaffen. Moldzio & Partner arbeitet für renommierte Unternehmen wie beispielsweise BMW, Miele, Hannover Rückversicherung oder Smurfit Kappa.

Über ein persönliches Gespräch mit Ihnen würde ich mich sehr freuen. Hierzu werde ich Sie in der nächsten Woche anrufen.

Mit freundlichen Grüßen

Kirsten Dreier
Kirsten Dreier



MOLDZIO & PARTNER
INSTITUT FÜR PERSONALAUSWAHL

THOMAS MOLDZIO · DIPL.-PSYCHOLOGE
DR. BERND SOBOTTKA · DIPL.-PSYCHOLOGE
MATTHIAS MICELEIT · DIPL.-PSYCHOLOGE

STOLBERGSTR. 13 · 22967 TREMSBÜTTEL
TELEFON (0 45 32) 26 61 - 0
TELEFAX (0 45 32) 26 61 - 21

WWW.MOLDZIO.COM
KONTAKT@MOLDZIO.COM

Tremsbüttel, 19.08.2010
KD/ mwe

COMMERZBANK HAMBURG
BLZ 200 400 00 · KTO.-NR. 495 14 14

SPARKASSE DIETMOLD
BLZ 476 501 30 · KTO.-NR. 460 666 84

AMTSGERICHT KIEL PR 211 KI

B3. Weitergehende Informationen zur Rekrutierung der Kontrollstichprobe (2 Seiten)



Universität Hamburg
Fakultät für Erziehungswissenschaft,
Psychologie und Bewegungswissenschaft



MOLDZIO & PARTNER
INSTITUT FÜR PERSONALAUSWAHL

Geplante Kontrollgruppenuntersuchung (Abschlussjahrgang 2011)

Untersuchungsplan:

Hintergrund der Untersuchung ist die geplante Überprüfung der Arbeitsbezogenen Gewissenhaftigkeitsskalen (AGS) sowie die empirische Bestätigung des theoretischen Modells zur Vorhersage von schulischen Leistungskriterien (hier: Abschlussnoten) durch spezifische (berufsbezogene) Persönlichkeitseigenschaften.

Eine ähnliche Studie wird bereits bei einem [REDACTED] im Rahmen einer Feldstudie durchgeführt. Im Rahmen der Auszubildendenauswahl des Unternehmens werden in diesem über einen Zeitraum von drei Jahren ca. 1200 Bewerber systematisch durch den Bewerbungs- und Berufsschulprozess begleitet. Die Schülerstichproben fungieren hierzu als unabhängige Vergleichsstichprobe.

Untersuchungsparadigma:

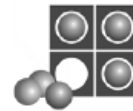
- Hoch gewissenhafte Personen sind schulisch/ beruflich erfolgreicher als wenig zuverlässige und ordentliche Menschen.
- Neuere Erkenntnisse belegen eine zweidimensionale Substruktur der Persönlichkeitseigenschaft Gewissenhaftigkeit. Diese Subfacetten – Ordnung und Fleiß – eignen sich besser zur Vorhersage von schulischen/ beruflichen Erfolgs- und Zufriedenheitsmaßen als der globale, übergeordnete Faktor Gewissenhaftigkeit.
- Schulisch/ beruflich erfolgreiche Personen blicken zuversichtlicher in die Zukunft und schätzen die eigenen Fähigkeiten besser ein bzw. haben eine höhere Selbstwirksamkeitserwartung.
- ABER:
 - Ein integratives Vorhersagemodell mit den Gewissenhaftigkeitsaspekten sowie Selbstwirksamkeitserwartungen zur Vorhersage von schulischen/ beruflichen Erfolgskriterien fehlt in der Theorie bislang.
 - Ergebnisse aus Längsschnittuntersuchungen sowie von Feldstichproben lassen sich in der bestehenden Literatur nur schwer finden.
 - Der Einfluss sozial erwünschten Antwortverhaltens insbesondere im Kontext von Auswahl-situationen ist bisher noch ungeklärt.

Ziele der Kontrollgruppenuntersuchung:

- Überprüfung des theoretischen Modells zur Vorhersage von schulischen Leistungskriterien (=Abschlussnoten)
- Überprüfung der Skaleneigenschaften der AGS hinsichtlich der Gütekriterien Reliabilität und Validität
- Daten als Basis für Normwertberechnungen
- Überprüfung der Tendenz zur positiven Selbstdarstellung (sozial erwünschtes Antwortverhalten) bei Selbstauskünften
- Herstellung der Vergleichbarkeit/ Generalisierbarkeit von Selbstauskünften in neutralen vs. Auswahl-situationen



Universität Hamburg
Fakultät für Erziehungswissenschaft,
Psychologie und Bewegungswissenschaft



MOLDZIO & PARTNER
INSTITUT FÜR PERSONALAUSWAHL

Ablauf der Testung:

Dauer: ca. 25 Minuten

Inhalte:

1. *Einweisung* (Einverständniserklärung der Schüler, Erläuterung des Untersuchungshintergrundes)
2. *Testung*: individuelles Bearbeiten des Testpakets bestehend aus verschiedenen Selbstbeschreibungsinventaren zur Erfassung individueller Persönlichkeitseigenschaften (Ordnung, Fleiß, Impression Management, Selbstwirksamkeit) und Leistungserwartungen (Schulnoten, eigene wahrgenommene Fähigkeiten) sowie einigen Fragen zur Erhebung demografischer Daten (Alter, Geschlecht, evt. Berufswunsch)
3. *Abschluss*: Klärung evt. noch offener Fragen
4. *Frühjahr 2011*: Abfrage der tatsächlich erreichten Durchschnittsnoten der Abschlusszeugnisse mittels anonymisierter Listen

Den Schulen entsteht durch die kurze Befragung kein Mehraufwand:

Die Testvorbereitung, -durchführung, -auswertung sowie -nachbereitung wird komplett durch Kirsten Dreier (Moldzio & Partner – Institut für Personalauswahl) übernommen.

Vorteile durch die Teilnahme an der Untersuchung:

- Die Schulen bzw. Schüler leisten durch die Teilnahme an der Untersuchung einen Beitrag zur Erhöhung der Qualität der Entscheidungen im Rahmen von Personalauswahlverfahren. Insbesondere auch die Auswahl von Azubis wird hierdurch verbessert.
- Die Schüler erhalten einen Einblick in die Praxis der Personalauswahl und können aktiv verschiedene eignungsdiagnostische Auswahlinstrumente erkunden.
- Die Schulen leisten einen Beitrag zur Forschung.
- Die Schüler werden zur Reflektion ihrer schulischen und ggf. auch beruflichen Ziele angeregt, was motivationsförderlich wirken kann.
- Als Dankeschön für die Möglichkeit der Schülerbefragung kann ein Projekttag/ eine Unterrichtseinheit zum Thema "Assessment-Center & Co. – Was erwartet mich in einem Personalauswahlverfahren?" durch Moldzio & Partner an der Schule gestaltet werden.

Anmerkung:

Eine Rückmeldung individueller Testergebnisse an die Schulen ist nicht vorgesehen (Vertraulichkeit), um eine wahrheitsgemäße Bearbeitung der Fragebögen zu gewährleisten und die Akzeptanz der Untersuchung zu erhöhen. Die Schulen können jedoch eine Gesamtauswertung der Studienergebnisse erhalten.

B4. Instruktionen der Kontrollgruppentestung

Liebe Schülerin,
Lieber Schüler,

vielen Dank für Ihre Mithilfe bei der Datengewinnung für meine Doktorarbeit! Ich erforsche, ob Fleiß und Ordnung Schul- bzw. Ausbildungserfolg vorhersagen können. Hierzu halten Sie einen Fragebogen in Ihren Händen.

Auf der ersten Seite möchte ich einige allgemeine Informationen wie beispielsweise Alter und Geschlecht von Ihnen wissen. Diese Angaben benötige ich z.B. für den Vergleich von Mädchen und Jungen.

Danach kommen verschiedene Fragenblöcke. Es handelt sich vor allem um Fragen zu Ihrem Arbeits- bzw. Lernverhalten. Aber auch Ihre Erwartungen an Ihre Leistungen im Abschlusszeugnis interessieren mich.

Bitte beantworten Sie **alle** Fragen sorgfältig und wahrheitsgemäß. Es gibt **keine richtigen oder falschen Antworten!** Lediglich Ihre **persönliche Einschätzung** ist wichtig!

Alle Angaben werden gemäß § 203 StGB streng vertraulich behandelt. Keine Daten oder Testergebnisse werden außerhalb der allgemeinen Auswertungen an andere weitergegeben. Dies ist aufgrund der Anonymisierung der Daten auch nicht möglich.

Falls Sie jedoch an den allgemeinen Ergebnissen meiner Doktorarbeit interessiert sind, schicke ich Ihnen diese nach Auswertung und Bearbeitung aller Daten selbstverständlich gern zu. Bitte kreuzen Sie hierzu das folgende Kästchen an und geben Sie Ihre E-Mail-Adresse an.

☐ Ja, ich möchte über die Ergebnisse der Studie informiert werden:

E-Mail-Adresse: _____

Einverständniserklärung:

Ich nehme zur Kenntnis und erkläre hiermit mein Einverständnis, dass Moldzio & Partner – Institut für Personalauswahl meine Daten auswerten dürfen. Ferner dürfen meine anonymisierten Daten und Auswertungen im Rahmen der wissenschaftlichen Forschung in Zusammenarbeit mit der Universität Hamburg verwendet werden.

Die Bestätigung erfolgt durch meine Unterschrift.

Ort

Datum

Unterschrift

Stadtteilschule

09.12.1963

Martina

09Mar38c

38c

[illegible]