



Entwicklung und Erprobung eines
Beobachtungsinstruments zur Erfassung
fachspezifischer Unterrichtsqualität im
Mathematikunterricht

Publikationsbasierte Dissertation
zur Erlangung des akademischen Grades einer Doktorin
der Philosophie (Dr. phil.)
an der Fakultät für Erziehungswissenschaft
der Universität Hamburg

vorgelegt von
Lena Schlesinger
bei Prof. Dr. Gabriele Kaiser

Hamburg, Mai 2018

Gutachterinnen und Gutachter:

1. Prof. Dr. Gabriele Kaiser, Universität Hamburg
2. Prof. Dr. Johannes König, Universität zu Köln
3. Prof. Dr. Jörg Doll, Universität Hamburg

Ort und Datum der mündlichen Disputation:

Universität Hamburg, 17. Juli 2018

Inhaltsverzeichnis

1	Einleitung	1
2	Ziel und Anlage der Dissertation	2
3	Theoretischer Rahmen und Forschungsstand	5
3.1	Konzeptualisierung von Unterrichtsqualität	5
3.2	Definition: Fachspezifische Qualitätsmerkmale	8
3.3	Beobachtungsinstrumente zur Erfassung der Unterrichtsqualität	9
3.4	Konzeptualisierung fachspezifischer Unterrichtsqualität	10
4	Unterrichtsbeobachtungen als Methode der Datenerhebung	13
4.1	In-vivo Beobachtungen	13
4.2	Hoch-inferente Ratings	15
5	Darstellung des Beobachtungsinstruments der Studie TEDS-Unterricht	17
5.1	Die TEDS-Studien	17
5.2	Entwicklung des Beobachtungsinstruments	19
5.2.1	Items der drei Basisdimensionen	19
5.2.2	Items der zwei fachspezifischen Dimensionen	21
5.3	Desiderata des Beobachtungsinstruments	25
5.4	Analyseeinheit und Training der Rater	26
6	Publikation I	28
6.1	Darlegung des eigenen Anteils an der Publikation	29
6.2	Abdruck der Publikation I	29
7	Publikation II	42
7.1	Darlegung des eigenen Anteils an der Publikation	42
7.2	Abdruck der Publikation II	43
8	Publikation III	67
8.1	Darlegung des eigenen Anteils an der Publikation	67
8.2	Abdruck der Publikation III	68
9	Ergebnisse	85
9.1	Theoretische Ergebnisse zur Entwicklung des Beobachtungsinstruments	85
9.2	Empirische Ergebnisse	86
9.3	Validierungsergebnisse	87

10 Diskussion	90
10.1 Stärken und Schwächen des Beobachtungsinstruments	90
10.2 Weiterentwicklungspotenziale	92
10.3 Fazit und Ausblick	94
Literaturverzeichnis	96
A Zusammenfassung	103
B English summary	105
C Beobachtungsinstrument TEDS-Unterricht	107
D Kodiermanual	111
E Auflistung der Einzelarbeiten der Dissertation	156
F Liste der Publikationen von Lena Schlesinger	157

1 Einleitung

Die Erforschung von Unterrichtsqualität sowie deren Konzeptualisierung und Messung sind in den letzten Jahren intensiv diskutiert worden. Insbesondere die Qualität des Mathematikunterrichts ist in Deutschland aufgrund des schlechten Abschneidens deutscher Schülerinnen und Schüler bei internationalen Vergleichsuntersuchungen wie TIMSS oder PISA in den Fokus der Bildungspolitik und der empirischen Bildungsforschung gerückt. Hierbei wurden zuletzt Zusammenhänge zwischen dem Professionswissen der Lehrpersonen, der Qualität ihres Unterrichts und den Leistungsfortschritten der Schülerinnen und Schüler genauer untersucht, um darauf aufbauend auch im Bereich der Lehrerbildung neue Impulse setzen zu können (Baumert et al., 2010; Hiebert & Grouws, 2007; Hill, Rowan & Ball, 2005; König, Blömeke, Paine, Schmidt & Hsieh, 2014; König et al., 2017).

Hinsichtlich der Konzeptualisierung von Unterrichtsqualität hat sich im deutschsprachigen Raum der Ansatz von Klieme, Lipowsky, Rakoczy und Ratzka (2006) etabliert. Diese beschreiben auf fachübergreifender Ebene Unterrichtsqualität mittels der drei Basisdimensionen *Klassenführung*, *konstruktive Unterstützung* der Lernenden und *kognitive Aktivierung*. Diese werden als maßgeblich für den Leistungszuwachs und die Motivationsunterstützung der Schülerinnen und Schüler angesehen (Baumert et al., 2010; Helmke, 2012; Klieme et al., 2006; Lipowsky et al., 2009). Bei der genauen Erforschung der Qualität von Mathematikunterricht stellt sich jedoch die Frage, welche fachspezifischen Qualitätsaspekte ebenfalls von Bedeutung sind. So argumentieren Brunner (2017) und Blum, Drüke-Noe, Hartung und Köller (2012), dass die *fachliche Richtigkeit* sowie der *fachliche Gehalt* des Unterrichts eine entscheidende Rolle bei der Einschätzung der Unterrichtsqualität spielen. Diese Aspekte werden jedoch in dem Modell der drei Basisdimensionen von Unterrichtsqualität nicht explizit berücksichtigt.

Zur Messung der Unterrichtsqualität wurden in den letzten Jahren viele verschiedene Beobachtungsinstrumente entwickelt, von denen manche fachübergreifend einsetzbar sind, während andere speziell für einzelne Fächer entwickelt wurden. Insbesondere für den Mathematikunterricht gibt es eine Reihe von fachspezifischen Beobachtungsinstrumenten, die sich jedoch zum Teil deutlich unterscheiden und sehr verschiedene Qualitätsaspekte fokussieren (Schlesinger & Jentsch, 2016).

Diese Diversität im Bereich der Konzeptualisierung und Messung von Unterrichtsqualität im Mathematikunterricht bildet den Ausgangspunkt der vorliegenden Dissertation. Sie greift damit sowohl fachübergreifende als auch fachspezifische Forschungsansätze zur Konzeptualisierung und Messung von Unterrichtsqualität auf und versucht, diese Bereiche miteinander zu verbinden.

2 Ziel und Anlage der Dissertation

Ziel der vorliegenden publikationsbasierten Dissertation ist es, Unterrichtsqualität unter einer mathematikdidaktischen Perspektive genauer zu untersuchen. Hierzu werden Möglichkeiten aufgezeigt, fachspezifische Qualitätsaspekte mit externen Beobachtungsverfahren im Mathematikunterricht in-vivo zu erfassen. Da es bisher kein Beobachtungsinstrument gibt, das die Basisdimensionen von Unterrichtsqualität berücksichtigt und gleichzeitig relevante fachspezifische Qualitätsaspekte des Mathematikunterrichts simultan erfasst, wurde im Rahmen dieser Dissertation ein solches Beobachtungsinstrument zur Erfassung einer (mathematikdidaktischen) Unterrichtsqualität entwickelt und erprobt.

Neben der Darstellung empirischer Ergebnisse hinsichtlich der Reliabilität der Messung fand auch eine erste Validierung des Instruments über die professionellen Kompetenzen von Mathematiklehrkräften statt. Die Ergebnisse hinsichtlich der Zuverlässigkeit der Daten des Beobachtungsinstruments sind eine entscheidende Voraussetzung, um Aussagen und Ergebnisse, die mit dem vorliegenden Beobachtungsinstrument gewonnen werden, angemessen interpretieren und bewerten zu können. Die Ergebnisse bieten darüber hinaus das Potenzial einer theoretischen Weiterentwicklung bisheriger Modelle zur Unterrichtsqualität im Mathematikunterricht.

Die publikationsbasierte Dissertation wurde im Rahmen der Studie TEDS-Unterricht verfasst. TEDS-Unterricht ist ein Nachfolgeprojekt der Studien TEDS-M (*Teacher Education and Development Study: Learning to Teach Mathematics*) (Blömeke, Kaiser & Lehmann, 2010) und TEDS-FU (*Follow-Up*) (Blömeke et al., 2014). Dieses Forschungsprojekt wurde 2014-2016 in Hamburg durchgeführt mit dem Ziel, die professionelle Kompetenz von Mathematiklehrkräften der unteren Sekundarstufe und ihres über die Unterrichtsqualität vermittelten Einflusses auf den Leistungszuwachs von Schülerinnen und Schülern zu untersuchen. Dabei stellte die Konzeptualisierung von Unterrichtsqualität durch fachdidaktische Merkmale des Mathematikunterrichts einen wichtigen Teilaspekt dar.

Für die Instrumententwicklung waren die folgenden Fragestellungen leitend:

1. *Wie können Aspekte mathematikdidaktischer Unterrichtsqualität theoretisch konzeptualisiert werden?*
2. *Welche theoretischen und methodologischen Herausforderungen gibt es bei der Messung einer mathematikdidaktischen Unterrichtsqualität durch externe Beobachterinnen und Beobachter?*

Zur Erprobung des Beobachtungsinstruments wurden bei 37 Lehrpersonen Unterrichtsbeobachtungen durchgeführt und ausgewertet. Mithilfe der dabei gewonnenen Daten lassen sich des Weiteren die folgenden empirischen Forschungsfragen genauer untersuchen:

3. *Lässt sich Unterrichtsqualität im Mathematikunterricht aufbauend auf theoretischen Konzeptualisierungen mit mehreren Dimensionen reliabel erfassen?*

-
4. *Welche Zusammenhänge weisen fachspezifische Qualitätsdimensionen mit den Basisdimensionen von Unterrichtsqualität auf?*

Hinsichtlich einer ersten Validierung des Beobachtungsinstruments werden im Rahmen dieser Dissertation abschließend zwei Forschungsfragen in den Fokus genommen, die einerseits Zusammenhänge zu den professionellen Kompetenzen von Lehrpersonen untersuchen und andererseits eine inhaltliche Validierung unter Bezug auf weitere fachspezifische Beobachtungsinstrumente vornehmen:

5. *Sind im Mathematikunterricht fachdidaktische Unterrichtsqualität und domänenspezifische Kompetenzfacetten von Lehrpersonen miteinander korreliert?*
6. *Welche Stärken und Schwächen weist das entwickelte Beobachtungsinstrument zur Erfassung mathematikdidaktischer Unterrichtsqualität in Bezug zu anderen fachspezifischen Beobachtungsinstrumenten auf?*

In dieser Dissertation werden insgesamt drei Zeitschriftenartikel präsentiert, die sich mit den genannten Forschungsfragen intensiv auseinandersetzen. Der erste Artikel „*Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations*“, der 2016 in der Zeitschrift *ZDM Mathematics Education* erschienen ist, behandelt die Fragestellungen 1) und 2) und bildet die Grundlage für die Entwicklung des Beobachtungsinstruments. Der zweite Artikel „*Unterrichtsqualität unter einer mathematikdidaktischen Perspektive – Konzeptualisierung, Messung und Validierung*“, der voraussichtlich 2018 im *Journal für Mathematik-Didaktik* erscheint, befasst sich mit den Forschungsfragen 3), 4) und 5), womit einerseits empirische Ergebnisse in den Fokus genommen werden und andererseits erste Validierungen präsentiert werden. Der dritte Artikel „*Subject-Specific Characteristics of Instructional Quality in Mathematics Education*“, der wiederum 2018 in der Zeitschrift *ZDM Mathematics Education* erschienen ist, thematisiert die letzte Forschungsfrage 6) und analysiert mithilfe detaillierter Unterrichtsbeschreibungen Stärken und Schwächen des Beobachtungsinstruments.

Die vorliegende Dissertation gliedert sich damit wie folgt: Zunächst werden in Kapitel 3 der theoretische Rahmen und der Forschungsstand dargelegt, indem eine Konzeptualisierung von Unterrichtsqualität über die drei Basisdimensionen genauer dargestellt und eine Definition für fachspezifische Qualitätsaspekte entwickelt wird. Darauf aufbauend werden bisherige Beobachtungsinstrumente überblicksartig dargestellt und es wird eine Konzeptualisierung fachspezifischer Unterrichtsqualität vorgestellt. In Kapitel 4 wird die Methode der externen Beobachtungen mittels hoch-inferenter in-vivo Ratings erläutert. Daran anschließend wird in Kapitel 5 das Beobachtungsinstrument detailliert dargestellt, indem zunächst auf die TEDS-Studien eingegangen wird, in deren Rahmen das Instrument entwickelt und erprobt wurde. Die drei Zeitschriftenartikel werden in den Kapiteln 6, 7

und 8 genauer eingeführt und vorgestellt¹. In Kapitel 9 und 10 werden die Ergebnisse im Hinblick auf die eingangs genannten Forschungsfragen zusammenfassend präsentiert und es werden Stärken und Schwächen sowie Weiterentwicklungspotenziale des Beobachtungsinstruments diskutiert. Die Dissertation schließt mit einem Fazit und einem Ausblick.

¹Eine Übersicht über weitere Publikationen, die im Rahmen der Dissertation entstanden sind, befindet sich im Anhang in Abschnitt F.

3 Theoretischer Rahmen und Forschungsstand

Unterrichtsqualität wird in der Unterrichtsforschung als ein mehrdimensionales und facettenreiches Konstrukt verstanden, das nicht einfach zu beschreiben und zu erfassen ist. Daher werden die Dimensionalität und die berücksichtigten Aspekte von Unterrichtsqualität je nach Forschungsziel und Untersuchungsfokus oftmals kontrovers diskutiert (siehe Taut & Rakoczy, 2016).

Einerseits werden aus einer normativen Perspektive heraus Aspekte benannt, die einen guten Unterricht aus Sicht bestehender Standards und Prinzipien ausmachen (Brunner, 2017; Kunter & Ewald, 2016), andererseits ist aus einer empirischen Perspektive heraus Unterricht dann gut, wenn er effektiv bezüglich bestimmter festgelegter Ziele (wie beispielsweise Schülerleistungen oder Motivationsförderung) ist. Darauf aufbauend hat sich der Begriff „Quality teaching“ nach Berliner (2005) als Verknüpfung eines guten sowie effektiven Unterrichts etabliert (Brunner, 2017; Kunter & Ewald, 2016). Das Konstrukt der Unterrichtsqualität lässt sich damit insgesamt als eine Verbindung normativer sowie empirischer Aspekte verstehen (Brunner, 2017).

Aspekte von Unterrichtsqualität lassen sich darüber hinaus auch dahingehend unterscheiden, inwieweit sie sich auf Sicht- oder aber Tiefenstrukturen des Unterrichts beziehen. Sichtstrukturen umfassen Unterrichtsaspekte, die direkt sichtbar und damit leicht erschließbar sind (Kunter & Ewald, 2016). Beispielsweise zählen dazu Organisations- oder Sozialformen des Unterrichts oder aber auch unter einer fachspezifischen Perspektive die behandelten Unterrichtsinhalte. Tiefenstrukturen des Unterrichts sind hingegen Aspekte, die nicht direkt sichtbar sind, da sie sich auf die verschiedenen Interaktionsebenen im Unterricht zwischen der Lehrperson, den Lernenden und dem Inhalt beziehen (Kunter & Ewald, 2016). Hierzu zählt beispielsweise die Unterstützung der Lernenden von Seiten der Lehrperson oder auch die fachliche Tiefe des Unterrichts. Es hat sich gezeigt, dass insbesondere die Tiefenstrukturen des Unterrichts einen bedeutsamen Einfluss auf Zielvariablen des Unterrichts (insbesondere Leistungszuwächse der Schülerinnen und Schüler) haben, weshalb sie auch als Dimensionen der Unterrichtsqualität bezeichnet werden (Clausen, 2002; Helmke, 2012; Kunter & Voss, 2011).

3.1 Konzeptualisierung von Unterrichtsqualität

In den letzten 15 Jahren hat sich im deutschsprachigen Raum ein Modell mit drei Basisdimensionen von Unterrichtsqualität etabliert, das ausgehend von den TIMSS Videostudien (Hiebert, Gallimore, Garnier & Stigler, 2003) mittels hoch-inferenter Beobachterratings empirisch entwickelt wurde und das einen positiven Einfluss auf Schülerleistungsfortschritte sowie deren Motivation zeigt (Baumert et al., 2010; Klieme, Pauli & Reusser, 2009; Lipowsky et al., 2009). Die drei Basisdimensionen lassen sich den Tiefenstrukturen von Unterrichtsqualität zuordnen. Die theoretische Basis dieser drei Dimensionen bilden dabei

Erkenntnisse aus der Lernpsychologie, wonach das Lernen besonders dann erfolgreich ist, wenn angemessen viel Lernzeit zur Verfügung steht, das Lernen selbstbestimmt geschieht und die Lernenden kognitiv herausgefordert werden. Das Modell der drei Basisdimensionen wurde damit fächerübergreifend konzeptualisiert, da diese Komponenten für alle Fächer relevant sind (Praetorius, Klieme, Herbert & Pinger, 2018).

Klassenführung

Effektive Klassenführung durch die Lehrperson stellt eine wichtige Komponente für erfolgreiches Lernen von Schülerinnen und Schülern dar, d.h. es wird eine möglichst störungsfreie und optimale Nutzung der zur Verfügung stehenden Lernzeit gewährleistet. Zur effektiven Klassenführung zählen sinnvolle Routinen und Regeln im Unterricht, ein störungspräventives Verhalten sowie ein transparenter Lernprozess und eine strukturierte Unterrichtsgestaltung (Brophy, 2000; Helmke, 2012; Klieme et al., 2006; König & Kramer, 2016; Kounin, 1970).

Konstruktive Unterstützung

Die zweite Dimension, eine konstruktive Unterstützung der Lernenden, beinhaltet eine unterstützende Lernumgebung, welche von der Lehrperson eine aufmerksame Diagnose von Schwierigkeiten und Fortschritten, individuelle Hilfestellungen und Rückmeldungen im Lernprozess sowie eine Individualisierung und Differenzierung des Unterrichtsangebots erfordert. Es zeigt sich, dass eine solche Unterstützung für die Motivation sowie das Kompetenz- und Autonomieerleben des einzelnen Lernenden von großer Bedeutung ist (Deci & Ryan, 1985; Praetorius, 2014; Rakoczy & Pauli, 2006). Die Begrifflichkeit innerhalb der Literatur ist für diese Dimension nicht einheitlich: So verwenden Klieme, Schümer und Knoll (2001) im Rahmen von TIMSS zunächst den Begriff „Schülerorientierung“, weitere Bezeichnungen sind „Motivationsunterstützung“ oder auch „unterstützendes Unterrichtsklima“ (Klieme & Rakoczy, 2008; Lipowsky et al., 2009). Im Folgenden wird für diese Qualitätsdimension der Begriff „konstruktive Unterstützung“ verwendet, da die Unterstützung der Lernenden durch individuelle Hilfestellungen, Differenzierung im Unterricht und detaillierte Rückmeldungen durch die Lehrperson im Fokus stehen. Auch im Rahmen des Projektes COACTIV (Kunter & Voss, 2011) wird der Begriff konstruktive Unterstützung verwendet.

Kognitive Aktivierung

Die dritte Qualitätsdimension, kognitive Aktivierung, beinhaltet Unterrichtsangebote, die anspruchsvolle Denk- und Lernprozesse bei den Schülerinnen und Schülern initiieren, und basiert auf der Annahme, dass Wissen das Ergebnis aktiver und individueller Konstruktionsprozesse ist, die intern, selbstgesteuert und subjektiv ablaufen, aber im Unterricht durch soziale Kontexte unterstützt werden müssen (Aebli, 1985; Baumert et al., 2010; Klieme & Rakoczy, 2008; Lipowsky et al., 2009; Reinmann & Mandl, 2006; Shuell, 1993). Insbesondere diese Dimension wird von verschiedenen Studien unterschiedlich konzeptualisiert, was auch der Tatsache geschuldet ist, dass sich eine kognitive Aktivierung der

Lernenden nicht unmittelbar beobachten lässt (Praetorius, 2014). Darüber hinaus wird die Fachspezifität dieser Dimension wiederholt angemerkt (Praetorius, Pauli, Reusser, Rakoczy & Klieme, 2014; Schlesinger & Jentsch, 2016).

In den letzten Jahren ist mehrfach die Frage aufgeworfen worden, inwieweit das Modell der drei Basisdimensionen fachspezifische Aspekte enthält und ob es um eine Beschreibung fachlicher Unterrichtsqualitätsaspekte erweitert werden müsste (Blum et al., 2012; Brunner, 2017). Es stellt sich damit auch insgesamt die Frage, ob und welche fachspezifischen Merkmale von Unterrichtsqualität sich grundsätzlich identifizieren lassen und inwiefern sie mit den drei fachübergreifenden Basisdimensionen von Unterrichtsqualität zusammenhängen.

Im Kontext der deutschen Bildungsstandards, in denen allgemeine prozessbezogene Kompetenzen für den Mathematikunterricht konkretisiert und vereinheitlicht worden sind, und in denen fachliche und fachübergreifende Ziele zur Verbesserung der Unterrichtsqualität formuliert werden, wird insbesondere von Blum et al. (2012) eine fachlich gehaltvolle Unterrichtsgestaltung als wesentliches Qualitätsmerkmal von Mathematikunterricht neben der kognitiven Aktivierung der Lernenden und einer effektiven und schülerorientierten Unterrichtsführung genannt (siehe auch Blum, Vogel, Drücke-Noe & Roppelt, 2015). Unter einer fachlich gehaltvollen Unterrichtsgestaltung versteht Blum et al. (2012) dabei einerseits permanente Gelegenheiten zum Kompetenzerwerb durch geeignete Aufgaben im Unterricht, andererseits einen langfristigen Kompetenzaufbau mit permanenten integrierten Übungen, Wiederholungen, Vertiefungen und Vernetzungen (Blum et al., 2012). Die allgemeinen, prozessbezogenen Kompetenzen, die hiermit gefördert und aufgebaut werden sollen, sind dabei überwiegend genuin fachspezifisch und lassen sich nicht direkt auf andere Fächer übertragen. Blum et al. (2015) betont darüber hinaus für einen qualitätsvollen Mathematikunterricht aus fachdidaktischer Perspektive die Notwendigkeit des Einsatzes einer Vielfalt geeigneter unterrichtlicher Beispiele und die Förderung individueller Lösungsstrategien. Die verwendeten Aufgaben und deren fachliche Qualität spielen dabei im Mathematikunterricht eine besondere Rolle, da sie sowohl inhaltlich als auch organisatorisch die zentrale Schnittstelle von Schüler- und Lehrertätigkeiten darstellen (Jordan et al., 2008).

Darüber hinaus zeigt Brunner (2017) auf, dass die fachliche Richtigkeit des Unterrichts in dem Modell der drei Basisdimensionen vollständig unberücksichtigt bleibt, jedoch aus einer normativen Perspektive heraus für einen qualitätsvollen Mathematikunterricht notwendig erscheint, da ansonsten kein zielführender Fachunterricht möglich ist. Dieses Argument ist überzeugend, wenn man sich Unterrichtsbeispiele vorstellt, bei denen die Schülerinnen und Schüler zwar sehr engagiert sind, der Unterricht reibungslos läuft und gut organisiert ist, die mathematische Qualität im Unterricht aber fehlt, da beispielsweise keine Verbindungen zu fundamentalen mathematischen Ideen hergestellt werden und die Lehrperson mathematisch unpräzise oder sogar inkorrekt ist.

3.2 Definition: Fachspezifische Qualitätsmerkmale

Bislang gibt es jedoch keine einheitliche Definition dafür, wann ein Qualitätsaspekt im Unterricht als fachspezifisch bezeichnet wird. Stattdessen oder gerade aus diesem Grund wurde die Fachspezifität bestimmter Qualitätsmerkmale in den letzten Jahren oftmals sehr kontrovers diskutiert. Es stellt sich zunächst insgesamt die Frage, was genau fachspezifisch bedeutet. Bedeutet Fachspezifität, dass der Aspekt nur in diesem Fach auftreten kann? Bedeutet Fachspezifität, dass er in diesem Fach relevanter ist als in anderen Fächern? Bedeutet es, dass die Lehrperson ein bestimmtes fachspezifisches Wissen benötigt, um den Aspekt im Unterricht erfolgreich umsetzen zu können? Wenn ein Qualitätsaspekt für ein bestimmtes Fach „spezifisch“ ist, dann müsste man annehmen, dass dieser Aspekt sich in verschiedenen Fächern unterschiedlich manifestiert oder aber, dass eine fachfremde Lehrperson nicht ohne weiteres in der Lage ist, diesen Qualitätsaspekt im Fachunterricht erfolgreich umzusetzen. Um sich also der Frage nach der Fachspezifität anzunähern, könnte man deshalb auch die Frage stellen, was eine fachfremde Lehrperson im Mathematikunterricht gut umsetzen kann, und bei welchen Qualitätsaspekten sie vermutlich Probleme haben wird, da ihr die Expertise fehlt. Brunner (2017) verwendet den Begriff fachspezifisch dazu, um Aspekte zu benennen, die aus der fachdidaktischen Forschung kommen. Die fachliche Richtigkeit, die sie als fachspezifischen Qualitätsaspekt des Unterrichts in den Vordergrund stellt, definiert sie dabei direkt über die fachliche Expertise der Lehrperson (d.h. das Fachwissen sowie das fachdidaktische Wissen der Lehrperson). Im Folgenden soll darauf aufbauend zunächst eine Definition des Begriffs fachspezifisch gegeben werden, die die Grundlage der vorliegenden Dissertation bildet.

Unter fachspezifischen Merkmalen von Unterrichtsqualität werden in der vorliegenden Arbeit solche Merkmale verstanden, deren Sicherstellung notwendigerweise ein gewisses Maß an Fachwissen oder fachdidaktischem Wissen der Lehrperson erfordert. Fachspezifische Merkmale von Unterrichtsqualität sind demnach Merkmale, die besondere fachliche oder fachdidaktische Herausforderungen an die Lehrperson und ihr unterrichtliches Handeln stellen. Unter generischen (auch fachübergreifenden oder fachunspezifischen) Merkmalen von Unterrichtsqualität werden hingegen Aspekte verstanden, die unabhängig vom fachlichen bzw. fachdidaktischen Wissen der Lehrperson sind und für die die Lehrperson lediglich pädagogisches Wissen benötigt. Dementsprechend lassen sich Qualitätsaspekte auch entlang eines Kontinuums anordnen, je nachdem, wie das Verhältnis von pädagogischem und fachspezifischem Wissen ist, das die Lehrperson für die Umsetzung benötigt.

Diese Definition ist im Einklang mit der Beschreibung fachspezifischer Qualitätsmodelle von Charalambous und Praetorius (2018): They are *„informed by subject-specific demands of teaching within a particular discipline“* (S. 2). Auch hier wird ein Kontinuum als Grundlage genommen, entlang dessen Modelle und Instrumente zur Unterrichtsqualität von generisch bis fachspezifisch angeordnet werden können.

Einerseits bietet eine solche Definition eine Grundlage für die Einordnung von Quali-

tätsaspekten des Unterrichts, andererseits ist für viele Qualitätsaspekte nicht eindeutig, ob und wie viel fachspezifisches Wissen der Lehrperson notwendig ist. Einige Qualitätsaspekte können relativ eindeutig als fachunspezifisch bezeichnet werden. Hierzu zählt vor allem eine effiziente Klassenführung, die von der Lehrperson in erster Linie pädagogische Fähigkeiten verlangt. Andere Qualitätsaspekte werden dagegen sehr kontrovers diskutiert, beispielsweise der Aspekt der „Klarheit“ im Unterricht, der sich je nach Konzeptualisierung überwiegend auf organisatorischer oder aber inhaltlicher Ebene verorten lässt und entsprechend eher pädagogisches oder aber auch fachdidaktisches Wissen benötigt (Brunner, 2017; Schlesinger & Jentsch, 2016).

Eine Einordnung von Qualitätsmerkmalen entlang des Kontinuums von fachunspezifisch bis fachspezifisch kann damit einerseits argumentativ vorgenommen werden, andererseits lässt sich auch empirisch genauer untersuchen, inwieweit tatsächlich für die Umsetzung von bestimmten Qualitätsaspekten im Unterricht fachspezifisches oder aber pädagogisches Wissen nötig ist.

3.3 Beobachtungsinstrumente zur Erfassung der Unterrichtsqualität

Zur Messung der Unterrichtsqualität mit externen Beobachterinnen und Beobachtern wurde in den letzten Jahren eine Vielzahl verschiedener Beobachtungsinstrumente entwickelt, die sich in Bezug auf die Dimensionalität der erfassten Aspekte, die Fokussierung des Instruments sowie den Einbezug von normativen gegenüber empirisch wirksamen Aspekten, Aspekten auf Sicht- und Tiefenstrukturebene, aber auch insbesondere hinsichtlich der fachspezifischen Gewichtung deutlich unterscheiden (für eine umfassende Aufarbeitung des Diskussionsstandes siehe Charalambous und Praetorius (2018); eine Aufarbeitung des Diskussionsstandes unter einer stärker mathematikdidaktischen Perspektive findet sich in Schlesinger und Jentsch (2016)).

Charalambous und Praetorius (2018) konstruieren ein Kontinuum von generisch bis fachspezifisch, entlang dessen sie Modellen und Instrumenten zur Unterrichtsqualität einordnen. Auf der generischen Seite des Kontinuums verorten sie beispielsweise das *Classroom Assessment Scoring System* (CLASS, Pianta & Hamre, 2009), das *Framework for Teaching* (FfT, Danielson, 2011), das *Dynamic Model of Educational Effectiveness* (DMEE, Creemers & Kyriakides, 2007) sowie das im Rahmen der Pythagoras-Studie verwendete Modell der drei genannten Basisdimensionen (Klieme et al., 2009), die in erster Linie fachübergreifende Qualitätsaspekte fokussieren.

Auf der fachspezifischen Seite des Kontinuums hingegen befinden sich beispielsweise das *Mathematical Quality of Instruction* (MQI, Hill et al., 2008; Learning Mathematics for Teaching Project, 2011), das *Elementary Mathematics Classroom Observation Form* (Thompson & Davis, 2014), das *Instructional Quality Assessment* (IQA, Matsumura, Garnier, Slater & Boston, 2008) oder das *Mathematics Scan* (M-Scan, Walkowiak, Berry, Meyer, Rimm-Kaufman & Ottmar, 2014), die speziell für den Mathematikunterricht

entwickelt wurden.

Darüber hinaus existieren einige „hybride“ Beobachtungsinstrumente, die in der Mitte des Kontinuums verortet werden können, da sie gleichzeitig fachübergreifende als auch fachspezifische Aspekte simultan fokussieren (Charalambous & Praetorius, 2018). Hierzu zählen beispielsweise das Projekt *Teaching for Robust Understanding* (TRU Math, Schoenfeld, 2013) oder das *UTeach Observation Protocol* (UTOP, Marder & Walkington, 2014). Entweder werden hierbei fachspezifische Qualitätsaspekte additiv zu fachunspezifischen Qualitätsaspekten ergänzt oder aber diese werden ineinander integriert (Brunner, 2017).

Um ein vertieftes Verständnis von Unterrichtsqualität zu erhalten und um genauer zu untersuchen, welche Dimensionalität Unterrichtsqualität aufweist, wurden in den letzten Jahren Studien durchgeführt, bei denen verschiedene Beobachtungsinstrumente miteinander kombiniert wurden (Blazar, Braslow, Charalambous & Hill, 2017; Charalambous & Praetorius, 2018; Kane & Staiger, 2012). Für das CLASS-Instrument, das Ähnlichkeiten zu dem Modell der drei Basisdimensionen aufweist (Pianta & Hamre, 2009), konnte dadurch gezeigt werden, dass sich in Kombination mit dem MQI-Instrument neben den fachübergreifenden Faktoren ein reiner, fachspezifischer Faktor identifizieren lässt, der die fachliche Richtigkeit sowie fachliche Fehler und Ungenauigkeiten im Unterricht beinhaltet (Blazar et al., 2017). Diese Dimensionalitätsanalysen legen die Vermutung nahe, dass auch für das Modell der drei Basisdimensionen von Unterrichtsqualität eine solche fachspezifische Ergänzung sinnvoll erscheint.

3.4 Konzeptualisierung fachspezifischer Unterrichtsqualität

Auf Basis einer systematischen Literaturrecherche zu bestehenden mathematiksspezifischen Beobachtungsinstrumenten (z.B. MQI, M-Scan, TRU Math) konnten zwei Perspektiven auf fachspezifische Qualitätsaspekte herausgearbeitet werden, die sich bei der Analyse der Dimensionen bestehender Beobachtungsinstrumente identifizieren ließen. Eine genauere Beschreibung dieser Literaturrecherche befindet sich auch in der ersten Publikation dieser Dissertation, die 2016 in *ZDM Mathematics Education* erschienen ist.

Einerseits konnte eine *mathematische, stoffbezogene Perspektive* identifiziert werden, die in erster Linie die mathematische, inhaltliche Qualität des Unterrichts fokussiert und auf der Annahme basiert, dass die Schülerinnen und Schüler nur effektiv lernen können, wenn die Inhalte im Unterricht korrekt, bedeutsam und fachlich tiefergehend sind. Hierzu zählen deshalb eine korrekte und präzise Darstellung mathematischer Inhalte (siehe auch Hill et al., 2008) sowie eine angemessene fachliche Tiefe des Unterrichts, die Möglichkeiten zu Verallgemeinerungen, inhaltlichen Vernetzungen sowie begrifflichen Verdichtungen bietet. Hierzu zählt auch, dass der Unterricht die allgemeinen mathematischen Kompetenzen wie Problemlösen, Modellieren oder auch Argumentieren und Beweisen fördert, die in den Bildungsstandards fest verankert sind (Blum et al., 2012).

Andererseits konnte eine *unterrichtsbezogene, mathematikdidaktische Perspektive* identifiziert werden, die fachdidaktische Unterrichtsentscheidungen fokussiert, mit der Annahme, dass ein effektives Lernen sowie eine Motivationsförderung der Schülerinnen und Schüler nur möglich sind, wenn die Unterrichtsinhalte für sie zugänglich und sinnstiftend gestaltet werden. Hierzu zählt deshalb, dass Unterrichtsbeispiele und Aufgaben angemessen ausgewählt werden, dass die Relevanz der vermittelten Inhalte deutlich gemacht wird sowie verschiedene Perspektiven und Darstellungsformen gewählt werden, die den Schülerinnen und Schülern verschiedene Zugänge zum mathematischen Stoff ermöglichen.

Eine solche zweigeteilte Konzeptualisierung der fachdidaktischen Unterrichtsqualität in Form einer stoffbezogenen und einer unterrichtsbezogenen Facette findet sich auch bei neueren Konzeptualisierungen des fachdidaktischen Professionswissens von Mathematiklehrkräften. Buchholtz, Kaiser und Blömeke (2014) konnten aufbauend auf den TEDS-Studien eine Unterteilung des fachdidaktischen Wissens in eine stoffbezogene, mathematikdidaktische und eine unterrichtsbezogene, mathematikdidaktische Subfacette vornehmen. Das stoffbezogene Wissen umfasst hierbei „*stofflich geprägte Fragestellungen des Lehrens von Mathematik*“ sowie eine „*fachlich geprägte Diagnostik von Schülerlösungen*“ (Buchholtz et al., 2014, S.111). Das unterrichtsbezogene Wissen beinhaltet demgegenüber Aspekte, die sich auf unterrichtsmethodische Entscheidungen beziehen. Hierzu gehören ein angemessener Umgang mit Heterogenität im Mathematikunterricht, fachdidaktische Lehr- und Lernformen sowie spezielle fachspezifische Unterrichtsarrangements.

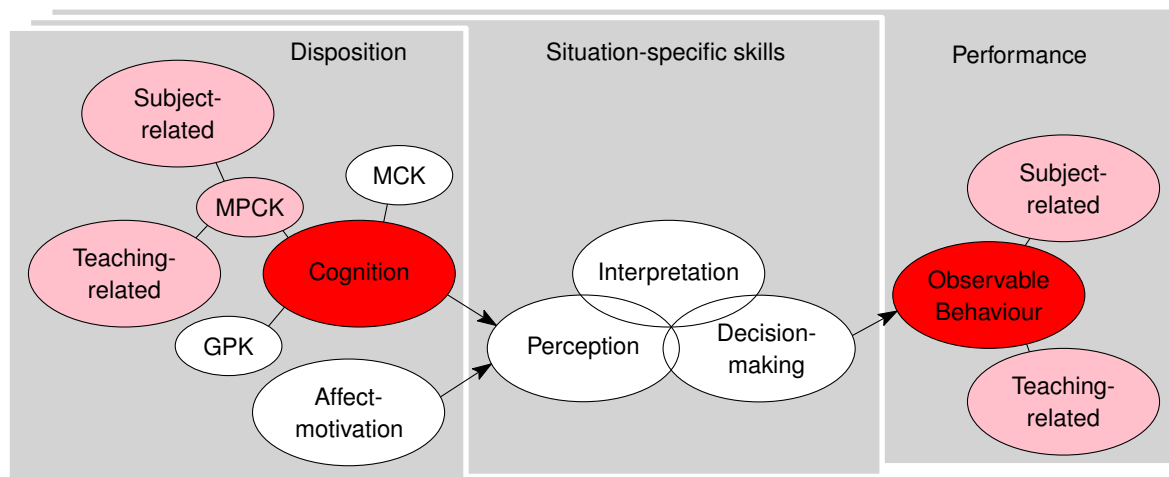


Abbildung 1: Modellierung der Kompetenz als Kontinuum nach Blömeke et al. (2014), adaptiert und weiterentwickelt

Seit der Annahme, dass die professionelle Kompetenz von Mathematiklehrkräften als Kontinuum modelliert werden kann, auf dem sowohl kognitions- als auch performanzähnliche Facetten angeordnet werden, die über die professionelle Unterrichtswahrnehmung (perception, interpretation, decision-making, bezeichnet als PID-Modell) mediiert werden (Blömeke, Gustafsson & Shavelson, 2015), stellt sich damit auch die Frage, inwieweit Domänen des professionellen Wissens der Lehrperson auch in ihrem Handeln im Unterricht

noch rekonstruierbar sind (siehe Abbildung 1).

Insofern eröffnet die zweigeteilte Konzeptualisierung sowohl des fachdidaktischen Wissens der Lehrperson als auch der fachdidaktischen Qualität im Unterricht interessante Möglichkeiten, um auch im Bereich der Kompetenzmodellierung die Diskussion voranzutreiben.

4 Unterrichtsbeobachtungen als Methode der Datenerhebung

Unterrichtsbeobachtungen gelten als zentrale Methode zur direkten Erfassung von Unterrichtsqualität in der schulischen Praxis und werden dazu u.a. auch von Schulinspektionen genutzt (Helmke, 2012; Pietsch & Tosana, 2008). Bei der Analyse von Stundenplanungen, Aufgaben oder anderen Unterrichtsmaterialien ist dagegen lediglich eine indirekte Erfassung von Unterrichtsqualität möglich (Praetorius, 2014). Daher wird im deutschsprachigen Raum von Unterrichtsforschung im engeren Sinne nur dann gesprochen, wenn auch Beobachtungsmethoden eingesetzt werden (Helmke, 2012; Waldis, Grob, Pauli & Reusser, 2010). Hierbei lassen sich drei zentrale Datenquellen unterscheiden: die Lehrkräfte, die Schülerinnen und Schüler und externe Beobachterinnen und Beobachter (im Folgenden auch Rater genannt). Letztere bieten im Vergleich zur Lehrer- oder Schülerperspektive einige Vorteile, welche nachfolgend betrachtet werden (Clausen, 2002; Kunter & Baumert, 2006; Praetorius, 2014).

Sowohl die Lehrperson als auch die Schülerinnen und Schüler sind während der Unterrichtsstunde aktiv am Unterrichtsgeschehen beteiligt und können sich deshalb nicht uneingeschränkt auf das Beobachten konzentrieren. Daher wird externen Beobachterratings eine höhere Validität zugesprochen (Praetorius, 2014; Rakoczy, 2008; Waldis et al., 2010). Schließlich beobachten geschulte Rater in der Regel eine große Anzahl von Unterrichtsstunden verschiedener Lehrkräfte, so dass sie im Gegensatz zu den Lehrkräften oder Schülerinnen und Schülern eine weniger subjektiv behaftete Einschätzung der Unterrichtsqualität abgeben können (Clausen, 2002; Praetorius, 2014; Rakoczy, 2008).

Mögliche Nachteile wie beispielsweise die Tatsache, dass der Beobachtungszeitraum der Rater im Vergleich zur Langzeitperspektive der Lehrperson und der Schülerinnen und Schüler gering ist und dass der Rater im Normalfall nur wenige Informationen über die Klasse und einzelne Schülerinnen und Schüler besitzt, werden als nicht so schwerwiegend eingeschätzt wie die Nachteile, die sich bei der Einschätzung durch die Lehrperson oder die Schülerinnen und Schüler ergeben (Praetorius, 2014). Diese und weitere Schwierigkeiten bei Unterrichtsbeobachtungen sind in den letzten Jahren vermehrt ins Zentrum der Aufmerksamkeit gerückt, darunter auch Fragen der Reliabilität und Validität bei Ratings durch externe Personen (Praetorius, Lenske & Helmke, 2012; Praetorius et al., 2014).

4.1 In-vivo Beobachtungen

Bei der Bewertung des Unterrichts durch externe Beobachterinnen und Beobachter stellt sich eine weitere Frage nach der Entscheidung zwischen in-vivo Beobachtungen im Kontrast zu technisch-unterstützten Beobachtungen mithilfe von Videoaufzeichnungen.

Videoaufzeichnungen stellen eine große organisatorische Herausforderung an ein Forschungsprojekt, da von der Genehmigung der Videographie über die Aufzeichnung bis hin zur Bereitstellung der Videos und Analysen ein enormer zeitlicher Aufwand nötig

ist. Darüber hinaus ist es zunehmend schwieriger, von Seiten der Schulbehörde und der Schulen Genehmigungen für das Aufzeichnen von Unterrichtsstunden zu erhalten, so dass für quantitative Forschungsvorhaben die Datenerhebung zu einer großen Herausforderung werden kann. Aus diesem Grund ist das in-vivo Beobachten und Bewerten des Unterrichts in den letzten Jahren zunehmend attraktiver geworden, insbesondere im Bereich der Schulinspektionen, die eine große Anzahl an Unterrichtsstunden beobachten und bewerten müssen (Pietsch & Tosana, 2008). Darüber hinaus befinden sich die Beobachterinnen und Beobachter bei in-vivo Beobachtungen direkt im Unterrichtsgeschehen und sehen den Klassenraum und den Unterricht nicht nur durch die Perspektive der Videokamera. In letzterem Fall sind Aspekte außerhalb des Kamerafokus für den Rater nicht sichtbar, wenn er lediglich die videographierten Stunden bewertet (Casabianca et al., 2013).

Es gibt jedoch auch eine Reihe von Nachteilen der in-vivo Beobachtungen gegenüber Bewertungen anhand von Videos. Den Beobachterinnen und Beobachtern bei in-vivo Beobachtungen ist es nicht möglich, den Unterrichtsverlauf zu stoppen, um eine Bewertung abzugeben. Das bedeutet, dass Bewertungen der Unterrichtsqualität sehr schnell geschehen müssen, wenn im Verlauf einer Unterrichtsstunde mehrfach Bewertungen abgegeben werden sollen. Darüber hinaus ist es nicht möglich, sich eine Unterrichtsstunde noch einmal anzuschauen. Und auch die Audio-Qualität der Gespräche im Unterricht kann für in-vivo Rater eine Herausforderung sein, denn Gespräche am anderen Ende des Klassenraums zwischen der Lehrperson und einzelnen Schülerinnen und Schülern können unter Umständen für die Rater nur schwer oder gar nicht zu hören sein. Bei videographierten Unterrichtsstunden ist es mittlerweile Standard, dass die Lehrperson ein eigenes Mikrofon trägt, das auch Gespräche mit einzelnen Schülerinnen und Schülern aufzeichnet und für die Beobachterinnen und Beobachter im Nachhinein gut hörbar macht. Dies kann allerdings wiederum auch zu Nachteilen führen, wenn stattdessen andere Gespräche im Klassenraum durch die Platzierung der Mikrophone im Klassenraum verloren gehen (Casabianca et al., 2013). Und schließlich haben videographierte Unterrichtsstunden den Vorteil, dass die Bewertung der Unterrichtsqualität zu jeder beliebigen Tageszeit stattfinden kann, während in-vivo Beobachtungen dann stattfinden müssen, wenn die Unterrichtsstunde ansteht. So können zeitliche Effekte auf das Rating aufgrund der Tageszeit der Unterrichtsbeobachtung nicht vermieden werden.

Bisher gibt es allerdings nur sehr wenige Studien, die sich tatsächlich mit dem Bewertungseffekt zwischen in-vivo Beobachtungen und videographierten Unterrichtsstunden befassen (Casabianca et al., 2013). Casabianca et al. (2013) untersuchten den Effekt der Beobachtungsmethode mithilfe des CLASS-S (Secondary) Instruments in 82 Unterrichtsstunden zum Thema Algebra. Die Ergebnisse zeigen, dass die Methode keinen starken Einfluss auf die Bewertung der Qualität der Lehrperson, über ein Schuljahr betrachtet, hat. Jedoch gab es bezogen auf eine einzelne Unterrichtsstunde Unterschiede in der Reliabilität und Unterrichtsqualität. Die Mittelwerte und Reliabilitäten der gemessenen Un-

terrichtsqualität waren etwas höher für die in-vivo Bewertung gegenüber der Bewertung der Videos. Eine weitere Forschergruppe um Liang (2015) hat darüber hinaus Live-Video-Beobachtungen durchgeführt, d.h. die Bewertung der videographierten Unterrichtsstunde fand live in dem Moment des Unterrichts statt. Sie konnten damit Reaktivitätseffekte der Lehrpersonen sowie subjektive Wertungen der Rater reduzieren (Liang, 2015). Die Frage nach den Effekten dieser Kombination aus Videographie und Live-Beobachtung ist jedoch insgesamt noch schwieriger vorherzusagen.

4.2 Hoch-inferente Ratings

Die Bezeichnung hoch-inferente Ratings (auch Schätzverfahren genannt) wird verwendet, wenn für die Bewertung eines Qualitätsaspekts im Unterricht eine hohe Interpretations- und Abstraktionsleistung (Inferenz) von den Beobachterinnen und Beobachtern notwendig ist. Hierzu zählt beispielsweise die Bewertung der Qualität der Beziehungen zwischen der Lehrperson und den Lernenden, die nicht direkt beobachtbar und erfassbar ist, sondern anhand verschiedener Merkmale und Situationen im Unterricht abstrahiert und bewertet werden muss (Seidel & Prenzel, 2010). Im Kontrast dazu spricht man von niedrig-inferenten Beobachtungen, wenn beispielsweise Unterrichtsaspekte angegeben werden sollen, die direkt sichtbar sind (beispielsweise ob sich mehr als 50% der Lernenden melden oder ob im Unterricht Gruppenarbeit stattfindet).

Hoch-inferente Ratings stellen damit eine hohe Anforderung an die Beobachterinnen und Beobachter, da für die Bewertung der Qualitätsaspekte im Unterricht mehrere Ausprägungen gleichermaßen berücksichtigt werden müssen. In die Bewertung fließt ein

- wie häufig ein bestimmtes Verhalten oder Merkmal zwischen zwei Messzeitpunkten auftritt,
- wie intensiv ein bestimmtes Verhalten oder Merkmal auftritt,
- wie viel Prozent der Schülerinnen und Schüler der Klasse an einem bestimmten Verhalten teilhaben,
- ob ein bestimmtes Merkmal allen Schülerinnen und Schülern grundsätzlich zur Verfügung steht, auch wenn es nicht von allen genutzt wird.

Grundsätzlich gilt für die Bewertung eines Qualitätsaspekts, dass aus diesen Ausprägungen ein Gesamturteil der Beobachterinnen und Beobachter gebildet werden muss. Aus diesem Grund sind eine umfangreiche Schulung sowie eine detaillierte Beschreibung der Items mit Ankerbeispielen und Indikatoren im Vorfeld der Beobachtung notwendig.

Hoch-inferente Einschätzungen haben einerseits den Vorteil, dass Verhaltensweisen und Qualitätsaspekte des Unterrichts qualitativ eingestuft werden, so dass qualitative Abstufungen zwischen verschiedenen Unterrichtsstunden und Lehrpersonen erfasst werden

können, die eine enge Verbindung zu theoretischen Konstrukten und Qualitätsdimensionen aufweisen (Seidel & Prenzel, 2010). Darüber hinaus haben hoch-inferente Einschätzungen oftmals eine höhere Validität, weil sich in der Unterrichtsforschung gezeigt hat, dass niedrig-inferente Indikatoren aufgrund ihrer einfachen Erfassung wenig Erklärungskraft für die Lernprozesse von Schülerinnen und Schülern haben (Baumert et al., 2010). Der Grund dafür scheint zu sein, dass direkt beobachtbare Aspekte (z.B. welche Methode verwendet wird), d.h. die Sichtstrukturen des Unterrichts, manchmal völlig unabhängig von seiner tatsächlichen Qualität, d.h. den Tiefenstrukturen, variieren können (Kunter & Voss, 2011). Andererseits hat dieses Verfahren den Nachteil, dass das Rating keine Aussage über konkrete Verhaltensweisen mehr macht, sondern lediglich ein qualitatives Aggregieren verschiedener Verhaltensweisen im Unterricht ist (Seidel & Prenzel, 2010). Insgesamt steigt mit dem Interpretations- und Abstraktionsgrad der Bewertung auch die Gefahr, dass die Beobachterinnen und Beobachter sich bei der Einschätzung nicht einig sind. Eine Möglichkeit, dieser Gefahr entgegenzuwirken, ist, den Zeitraum der Bewertung zu verkürzen, so dass die Einschätzung nicht für eine gesamte Unterrichtsstunde abgegeben werden muss, sondern lediglich für kurze Abschnitte. Dies kann über die Einführung mehrerer Messzeitpunkte in einer Unterrichtsstunde realisiert werden. Wie bereits oben beschrieben, stellt dies jedoch wiederum erhöhte Anforderungen an die Beobachterinnen und Beobachter, wenn die Einschätzung in-vivo erfolgt.

5 Darstellung des Beobachtungsinstruments der Studie TEDS-Unterricht

Im Folgenden wird das Beobachtungsinstrument zur Erfassung der Unterrichtsqualität im Mathematikunterricht mit externen Beobachterinnen und Beobachtern genauer dargestellt, das im Rahmen der Studie TEDS-Unterricht entstanden ist und einen zentralen Bestandteil der Dissertation darstellt. Hierzu werden zunächst Anlage und Design der TEDS-Studien genauer dargestellt. Anschließend wird die Entwicklung des Beobachtungsinstruments detailliert beschrieben und es werden die verwendeten Items vorgestellt. Das Kapitel schließt mit einer Darstellung von Desiderata des Beobachtungsinstrumentes und mit einer Beschreibung der Analyseeinheit sowie dem Training der Beobachterinnen und Beobachter.

5.1 Die TEDS-Studien

Die Studie TEDS-M (*Teacher Education and Development Study of Mathematics*) wurde 2008/2009 in 17 Ländern (darunter Deutschland) durchgeführt und war eine internationale Vergleichsstudie der „International Association for the Evaluation of Educational Achievement (IEA)“ zur Wirksamkeit der Lehrerbildung im Fach Mathematik. Ziel war die Untersuchung der professionellen Kompetenz von angehenden Mathematiklehrkräften der Primar- und Sekundarstufe I. Hierzu wurden verschiedene Testinstrumente entwickelt und eingesetzt, die das mathematische, mathematikdidaktische sowie pädagogische Wissen erheben. Die Studie TEDS-M erbrachte verschiedene deskriptive und relationale Ergebnisse hinsichtlich der professionellen Kompetenz angehender Mathematiklehrkräfte und verschiedener Wirkungspotenziale in der Mathematiklehrerbildung (Blömeke et al., 2010).

Die Studie TEDS-FU ist eine Follow-Up-Studie von TEDS-M und wurde von 2010 bis 2013 in Deutschland durchgeführt mit dem Ziel, die Entwicklung der Kompetenzen der Mathematiklehrkräfte in der Berufseingangsphase zu untersuchen. Damit konnten erstmals auch längsschnittliche Ergebnisse in der Lehrerbildung entwickelt werden, indem eine Teilstichprobe der Mathematiklehrkräfte, die bereits an TEDS-M teilgenommen hatten, nach einigen Berufsjahren erneut getestet wurde. Unter der Annahme, dass sich die Kompetenzen im Laufe zunehmender Professionalisierung in Bezug auf das Kompetenzniveau und auch die Kompetenzstruktur verändern, wurden neben den Papier-und-Bleistift-Testinstrumenten aus TEDS-M weitere Instrumente u.a. in Form von Video-Vignetten-Tests entwickelt, die die Kompetenzen auch handlungsnah erfassen (Blömeke et al., 2014; Kaiser et al., 2017; Kaiser, Busse, Hoth, König & Blömeke, 2015; König, Blömeke, Klein et al., 2014).

TEDS-Unterricht ist ein Nachfolgeprojekt der Studien TEDS-M und TEDS-FU. Es ist ein Gemeinschaftsprojekt der Universität Hamburg, der Humboldt-Universität zu Berlin

und der Universität zu Köln in Kooperation mit dem Institut für Bildungsmonitoring und Qualitätsentwicklung und mit Unterstützung der Behörde für Schule und Berufsbildung in Hamburg. Die Projektleitung besteht aus Prof. Dr. Gabriele Kaiser, Prof. Dr. Johannes König und Prof. Dr. Sigrid Blömeke. TEDS-Unterricht wurde 2014-2016 in Hamburg durchgeführt mit dem Ziel, die professionelle Kompetenz von Mathematiklehrkräften der unteren Sekundarstufe und ihres über die Unterrichtsqualität vermittelten Einflusses auf den Leistungszuwachs von Schülerinnen und Schülern zu untersuchen. Das Design der Studie TEDS-Unterricht ist in Abbildung 2 dargestellt.

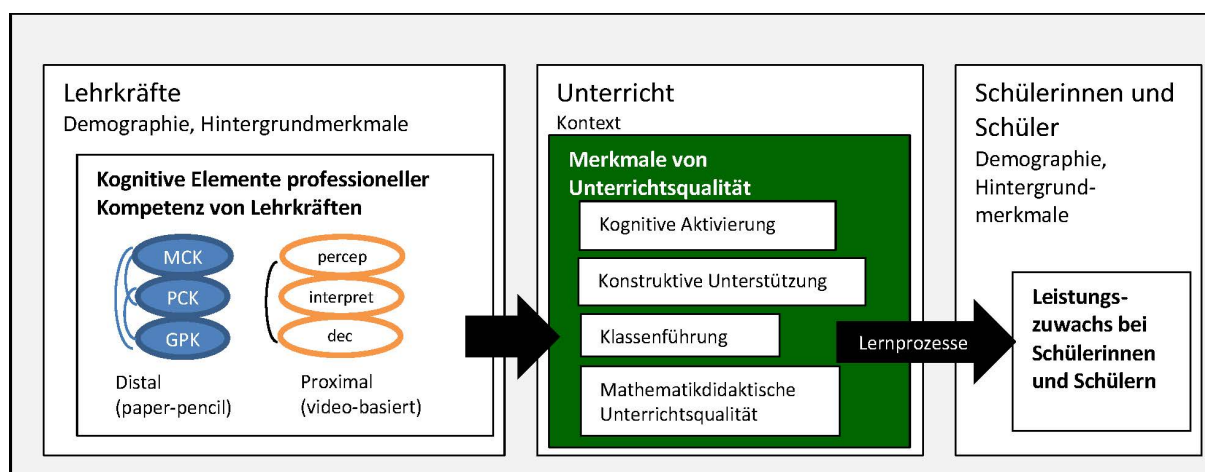


Abbildung 2: Untersuchungsmodell TEDS-Unterricht

Den theoretischen Rahmen stellt ein allgemeines Modell professioneller Kompetenz von Lehrkräften dar, das aus kognitiven und affektiv-motivationalen Facetten besteht. Die kognitive Facette, das Professionswissen, wird weiter untergliedert in die Dimensionen Fachwissen (mathematical content knowledge, MCK), fachdidaktisches Wissen (pedagogical content knowledge, PCK) und pädagogisches Wissen (general pedagogical knowledge, GPK) (Baumert & Kunter, 2006; Blömeke, 2002; Shulman, 1985, 1986). Die professionelle Unterrichtswahrnehmung als proximale Facette wird in die Subfacetten der Wahrnehmung, Interpretation und Entscheidung zentraler unterrichtlicher Ereignisse differenziert (perception, interpretation, decision-making, bezeichnet als PID-Modell) (Sherin, Jacobs & Philipp, 2011; Van Es & Sherin, 2002).

Grundlegende Vorarbeiten für die verwendeten Testinstrumente der Studie TEDS-Unterricht stellen dabei die in TEDS-M und in TEDS-FU entwickelten und überprüften Instrumente zur Erfassung der Kompetenzen von Mathematiklehrkräften dar. Die Studie TEDS-Unterricht verfolgt damit sowohl methodische als auch inhaltliche Ziele. Im Bereich Methodik bildet sie die Vorstudie zur Studie TEDS-Validierung, deren Ziel eine Validierung bisher entwickelter Instrumente sowie eine Replikation zentraler Ergebnisse des bisherigen Forschungsstandes zu Lehrerkompetenzen ist. Im inhaltlichen Bereich sind die Ziele der Studie, konkurrierende Hypothesen zur Struktur professioneller Kompetenz und des Zusammenhangs ihrer Dimensionen mit Unterrichtsqualität und Leistungszu-

wachs der Schülerinnen und Schüler zu prüfen. Bisher wurde das Zusammenspiel der für Lehrerprofessionswissen als unverzichtbar angesehenen Trias von MCK, PCK und GPK auf Unterrichtsqualität und Lernfortschritt der Lernenden nicht simultan modelliert, ebenso wenig wie eine Berücksichtigung situationsspezifischer kognitiver Fähigkeiten (perception, interpretation, decision-making, vgl. Blömeke et al., 2014). Darüber hinaus stellt die Konzeptualisierung von Unterrichtsqualität durch fachdidaktische Merkmale des Mathematikunterrichts einen wichtigen Teilaspekt dar, um ein erweitertes Verständnis von Unterrichtsqualität zu prüfen, das fachdidaktischen Ansprüchen differenzierter als bisher gerecht werden kann. Hierzu wurden bei einer Teilstichprobe in Hamburg neben der Testung zur professionellen Kompetenz außerdem Unterrichtsbeobachtungen im Mathematikunterricht durchgeführt.

5.2 Entwicklung des Beobachtungsinstruments

Zur Erfassung der Unterrichtsqualität von Mathematikunterricht wurde ein Beobachtungsinstrument adaptiert und weiterentwickelt, das einerseits die drei Basisdimensionen von Unterrichtsqualität und andererseits zwei weitere fachspezifische Dimensionen mathematikdidaktischer Unterrichtsqualität, eine unterrichtsbezogene und eine stoffbezogene, zu messen beansprucht. Die Bewertung der Items erfolgt auf einer vierstufigen Skala (1 = „trifft überhaupt nicht zu“ bis 4 = „trifft vollständig zu“). Im Folgenden werden die Items der drei Basisdimensionen sowie der zwei fachspezifischen Dimensionen vorgestellt und genauer erläutert. Das gesamte Beobachtungsinstrument mit Items und Indikatoren sowie das Kodiermanual befinden sich im Anhang in den Abschnitten C und D.

5.2.1 Items der drei Basisdimensionen

Die Items zur Erfassung der drei Basisdimensionen von Unterrichtsqualität wurden teilweise aus dem Beobachtungsinstrument der Pythagoras-Studie inhaltlich übernommen oder adaptiert (Rakoczy & Pauli, 2006). Darüber hinaus wurden weitere Items zu den drei Basisdimensionen ergänzt.

Die Dimension Klassenführung wurde in der Pythagoras-Studie lediglich mit drei Items erfasst, von denen das Item „Klarheit“ jedoch unzureichende Reliabilitätswerte (relativer Generalisierbarkeitskoeffizient) aufwies (Rakoczy & Pauli, 2006). In Anlehnung an aktuelle Instrumente der Schulinspektion (Pietsch & Tosana, 2008) wurden die drei Items deshalb ausdifferenziert und in insgesamt sechs Items unterteilt, so dass die Komplexität jedes Items reduziert werden konnte. Das Item „Klarheit“ wurde in die zwei Items „Transparenz des Lernprozesses“ und „Strukturierung des Unterrichts“ unterteilt, das Item „Disziplinprobleme/Unterrichtsstörungen“ wurde in die zwei Items „Störungspräventives Verhalten“ und „Arbeitsatmosphäre“ unterteilt, und das Item „Classroom Management“ wurde schließlich in die zwei Items „Effektive Lernzeitnutzung“ und „Sinnvolle Routinen“

und Regeln“ unterteilt. Detailliertere Beschreibungen der Items befinden sich im Kodiermanual (siehe Anhang, Abschnitt D). Zusammen mit dem Kodiermanual der Pythagoras-Studie ist es damit auch möglich, die Adaption und Veränderungen der Items detailliert nachzuvollziehen.

Die Dimension „konstruktive Unterstützung“ wurde ebenfalls in Teilen aus dem Pythagoras-Beobachtungsinstrument inhaltlich übernommen oder adaptiert. Items, die bereits bei der Pythagoras-Videostudie unzureichende Reliabilitätswerte aufwiesen, wurden nicht übernommen (beispielsweise „Einforderung von Anstrengung“, „Ängstlichkeit“ oder auch „Engagement der Lehrperson“). Der Grund hierfür war, dass die Herausforderung an die Rater in in-vivo Beobachtungen wie bereits dargestellt noch größer ist, so dass die Reliabilität voraussichtlich eher geringer ausgefallen wäre. Die Items „Rückmeldung der Lehrperson“, „Anerkennung/Wertschätzung der Lehrperson“ und „Kooperation“ aus der Pythagoras-Studie wurden inhaltlich übernommen. Das Item „Notendruck“ sollte zunächst für unser Beobachtungsinstrument adaptiert werden, musste jedoch nach der Pilotierung gestrichen werden, weil das Unterrichtsmerkmal nicht beobachtet werden konnte (siehe 5.3 „Desiderata des Beobachtungsinstruments“). Die Items „Wahlmöglichkeiten“, „Eigenverantwortung“ und „Individualisierung“ aus der Pythagoras-Studie wurden an die aktuelle Diskussion zur Individualisierung angepasst und in zwei Items „Differenzierung“ und „selbstgesteuertes Lernen“ unterteilt. Die beiden Items „Inhaltliche Relevanz“ und „Fehlerkultur“ wurden für das vorliegende Beobachtungsinstrument fachspezifisch konzeptualisiert (siehe Items der zwei fachspezifischen Dimensionen).

Die Dimension „kognitive Aktivierung“ wurde ebenfalls in Teilen aus der Pythagoras-Studie adaptiert. Das Item „Herausfordernde Probleme und Fragen“ wurde dabei inhaltlich direkt übernommen. Bei den Items „Exploration des Wissens“, „Exploration von Denkweisen“ und „Evolutionärer Umgang“ musste in der Pilotierung festgestellt werden, dass sie sich bei den Beobachtungen nicht angemessen trennen ließen, weshalb sie zu einem Item zusammengefasst wurden. Das Item „rezeptives Lernverständnis“ wurde für das Beobachtungsinstrument nicht übernommen, da es sich hierbei überwiegend um das negativ gepolte Item zu „Herausfordernde Probleme und Fragen“ sowie „Evolutionärer Umgang“ handelt. Diese Veränderungen im Vergleich zur Pythagoras-Studie resultierten in lediglich zwei Items, so dass zur breiteren Erfassung des Konstrukts kognitive Aktivierung drei weitere Items „kognitiv aktivierende Qualität der Unterrichtsmethoden“, „Förderung von Metakognition“ und „Wissenssicherung“ in Anlehnung an Helmke (2012) und Lipowsky (2015) hinzugenommen wurden. Detailliertere Beschreibungen der Items befinden sich wiederum im Kodiermanual im Anhang (Abschnitt D).

Insgesamt werden die drei Basisdimensionen in dem vorliegenden Beobachtungsinstrument somit mit 18 hoch-inferent einzuschätzenden Items erfasst.

5.2.2 Items der zwei fachspezifischen Dimensionen

Basierend auf den Vorarbeiten von Buchholtz et al. (2014) zur Konzeptualisierung mathematikdidaktischer Kompetenzen von Lehrpersonen und Analysen der systematischen Literaturrecherche zu bestehenden Unterrichtsbeobachtungsinstrumenten wurden zwei fachspezifische Dimensionen von Unterrichtsqualität in unserem Beobachtungsinstrument operationalisiert.

Die erste Dimension, die *stoffbezogene mathematikdidaktische Dimension*, umfasst dabei Aspekte, die sich auf die mathematische Korrektheit, Präzision und Tiefe der Unterrichtsinhalte beziehen. Hierzu zählen auch die fachliche Analyse von Fehlern sowie die fachliche Angemessenheit und der Fokus auf fundamentale mathematische Ideen. Die zweite *unterrichtsbezogene, mathematikdidaktische Dimension* beinhaltet demgegenüber Aspekte, die sich auf unterrichtsbezogene Entscheidungen beziehen, wie die Verwendung verschiedener Repräsentationsformen und angemessener Beispiele, Unterrichtsarrangements wie intelligente Übungsphasen und pädagogisch-psychologische Aspekte zur Motivation der Schülerinnen und Schüler durch Aufzeigen der Relevanz der Inhalte. In Tabelle 1 sind die fachspezifischen Items dargestellt, die im Folgenden genauer präsentiert werden.

Fachliche Korrektheit: Für einen erfolgreichen Mathematikunterricht ist es notwendig, dass die Lehrperson mathematisch korrekt und präzise agiert und insbesondere inhaltliche Fehler vermeidet (Schoenfeld, 2013). Ohne eine fachlich kompetente Lehrperson ist es den Schülerinnen und Schülern kaum möglich, selbst mathematische Kompetenzen zu erlangen (Drollinger-Vetter & Lipowsky, 2006; Learning Mathematics for Teaching Project, 2011). Aus diesem Grund stellt die fachliche Korrektheit der Aktionen der Lehrperson eine Notwendigkeit für erfolgreiche Lernprozesse der Schülerinnen und Schüler im Unterricht dar, worauf Brunner (2017) explizit hinweist.

Fachliche Erklärungen der Lehrperson: Für einen qualitativ hochwertigen Mathematikunterricht ist es notwendig, dass sich die Lehrperson einerseits verständlich und andererseits mathematisch korrekt ausdrückt (Learning Mathematics for Teaching Project, 2011). Dies bezieht sich auf kurze Erläuterungen der Lehrperson auf Nachfragen der Schülerinnen und Schüler, aber auch auf längere Lehrervorträge, die die Lehrperson ggf. schon im Vorfeld vorbereitet hat. Die Erklärungen der Lehrperson sollten dabei adressatengerecht und inhaltlich gut strukturiert sein (Drollinger-Vetter, 2011). Hierzu gehört, dass es der Lehrperson gelingt, an für das Verständnis substanziellen Stellen besonders präzise und anschaulich zu erklären, wofür sie bestimmtes mathematikdidaktisches Wissen benötigt (Schilcher, Krauss, Rincke & Hilbert, 2017).

Umgang mit Fehlern: Fehler gehören als produktive Bestandteile zum Lernprozess

Tabelle 1: Operationalisierung zweier fachspezifischer Qualitätsdimensionen

Mathematikdidaktische, stoffbezogene Qualität	Mathematikdidaktische, unterrichtsbezogene Qualität
Umgang mit inhaltlichen Fehlern der Schülerinnen und Schüler Die Lehrperson analysiert die Fehler der Schülerinnen und Schüler und verwendet sie als Lern Gelegenheit. Fachliche Korrektheit der Lehrperson Die Lehrperson ist mathematisch korrekt und präzise und macht keine inhaltlichen Fehler. Fachliche Erklärungen der Lehrperson Die fachlichen Erklärungen der Lehrperson sind gut strukturiert, präzise, adressatengerecht und fokussieren die grundlegenden mathematischen Konzepte und Ideen. Fachliche Tiefe Im Unterricht wird innermathematisch vernetzt und verallgemeinert und mathematische Inhalte werden strukturiert und begrifflich verdichtet. Förderung prozessbezogener, mathematischer Kompetenzen Die Lehrperson fördert die allgemeinen mathematischen Kompetenzen (z.B. Problemlösen, Modellieren, Argumentieren und Beweisen, mathematisch Kommunizieren, ...).	Verschiedene Repräsentationsformen und Darstellungswechsel Die Lehrperson verwendet verschiedene Repräsentationsformen im Unterricht und verknüpft diese miteinander, um den Schülerinnen und Schülern den Zugang zum mathematischen Inhalt zu erleichtern. Intelligentes Üben Die Lehrperson stellt Übungsaufgaben zur Verfügung, die sinnstiftend, entdeckungsoffen und selbstdifferenzierend sind sowie reflexive Aspekte enthalten. Angemessenheit der Beispiele Die ausgewählten mathematischen Beispiele fokussieren grundlegende mathematische Konzepte und Ideen. (Inhaltliche) Relevanz Die Lehrperson verdeutlicht die Relevanz der mathematischen Inhalte für die Schülerinnen und Schüler und zeigt Zusammenhänge zum Alltag auf, so dass die Lernenden ihre eigenen Erfahrungen und Interessen einbringen können.

der Schülerinnen und Schüler. Sie können beispielsweise dazu dienen, Verständnisprobleme zu klären (Zech, 1998). Zu einem qualitativ hochwertigen Unterricht gehört somit ein konstruktiver Umgang mit Schülerfehlern (Helmke, 2012). Dies bedeutet, dass auftretende Fehler als Lerngelegenheiten erkannt und genutzt werden, anstatt diese vorschnell zu korrigieren, zu übergehen oder sogar negativ zu bewerten (Learning Mathematics for Teaching Project, 2011). Es ist dabei wichtig, dass in Lernsituationen eine Fehlertoleranz ohne Angst vor Fehlern herrscht. Obwohl dieser Aspekt ebenfalls zur konstruktiven Unterstützung der Lernenden gehört (Rakoczy & Pauli, 2006), ist ein guter Umgang mit auftretenden Fehlern im Unterricht jedoch auch in hohem Maße fachspezifisch (Learning Mathematics for Teaching Project, 2011; Steinweg, 2011). Insbesondere im Mathematikunterricht besteht die Gefahr, dass die Lehrperson falsche Antworten lediglich kurz korrigiert, weil bestimmte Aufgaben nur eine „richtige“ Lösung aufweisen und es damit relativ klar ist, wann ein Fehler in der Mathematik vorliegt (Reiss & Hammer, 2013; Steuer, 2014). Beim konstruktiven Umgang mit Fehlern im Mathematikunterricht wird deshalb auch berücksichtigt, inwieweit Lehrpersonen über diagnostische Kompetenzen verfügen, um einen mathematischen Fehler zu erkennen, zu interpretieren und anschließend adäquate Maßnahmen zu ergreifen (Steuer, 2014).

Fachliche Tiefe: Mathematikunterricht ist kognitiv umso herausfordernder, je fachlich tiefer er gestaltet ist (Blum et al., 2012; Schoenfeld, 2013). Zu dieser fachlichen Tiefe zählen innermathematische Vernetzungen zwischen verschiedenen Inhaltsbereichen und Themengebieten, mathematische Verallgemeinerungen, aber auch die Verdichtung der mathematischen Inhalte durch beispielsweise neue Begrifflichkeiten und inhaltliche Erweiterungen (Sawada et al., 2002). Hierzu müssen die mathematischen Inhalte im Unterricht gut strukturiert und geordnet werden und alle Elemente, die für einen erfolgreichen Strukturaufbau und für das Verständnis erforderlich sind, müssen thematisiert werden, damit die Lernenden in der Lage sind, inhaltliche Lernerfolge zu erzielen und ihr eigenes Wissen zu strukturieren („Verstehenselemente“ nach Drollinger-Vetter, 2011).

Prozessbezogene (allgemeine) mathematische Kompetenzen: Seit der Etablierung der Bildungsstandards (KMK, 2003) als Reaktion auf die ersten PISA-Ergebnisse in Deutschland hat insbesondere für den Mathematikunterricht eine breite Diskussion über Bildungsziele und den Erwerb inhaltlicher Kompetenzen begonnen (Blum et al., 2012). Für den Mathematikunterricht wurden sechs prozessbezogene mathematische Kompetenzen (Mathematisch argumentieren, Probleme mathematisch lösen, Mathematisch modellieren, Mathematische Darstellungen verwenden, mit Mathematik symbolisch/formal/technisch umgehen, Mathematisch kommunizieren) formuliert, von deren Implementation eine Steigerung der Unterrichtsqualität

erwartet wird (Blum et al., 2012; Steinweg, 2011). Die Orientierung des Unterrichts an diesen sechs prozessbezogenen mathematischen Kompetenzen soll dazu führen, dass die Schülerinnen und Schüler vorgegebene Lernziele erreichen, Wissen länger behalten und vernetzen können und damit insgesamt Kompetenzen erwerben, die für ihr späteres Leben wichtig sind (Blum et al., 2012, 2015; Reiss & Hammer, 2013).

Repräsentationsformen und Darstellungswechsel: Die Verwendung verschiedener Repräsentationsformen und der Wechsel zwischen verschiedenen Darstellungen zielt darauf ab, ein vertieftes fachliches Verständnis bei den Schülerinnen und Schülern zu fördern (Bruner, 1974; Kersting, Givvin, Thompson, Santagata & Stigler, 2012). Auf diese Weise wird Lernenden ermöglicht, den mathematischen Unterrichtsgegenstand aus verschiedenen Perspektiven zu erfahren (Drollinger-Vetter & Lipowsky, 2006) (Drollinger-Vetter, 2011). Insbesondere im Mathematikunterricht scheint dies für einen erfolgreichen Lernprozess wichtig zu sein, da neben der ikonischen und der enaktiven Ebene die symbolische Darstellungsebene noch einmal in eine sprachliche und eine formale Ebene unterteilt wird, was die Komplexität der Repräsentationsformen erhöht (Zech, 1998).

Intelligentes Üben: Es ist unstrittig, dass Übungsphasen zum Mathematikunterricht gehören (Helmke, 2012). Für eine hohe Qualität des Mathematikunterrichts ist es deshalb notwendig, dass diese Phasen und die darin verwendeten, fachlichen Übungsaufgaben produktiv sind. In diesem Zuge wurde der Begriff „intelligentes Üben“ oder auch „produktives Üben“ etabliert. In intelligenten Übungsaufgaben werden Vorstellungen gefestigt und mathematische Begriffe und Verfahren reflektiert. Gute Übungsaufgaben zeichnen sich dadurch aus, dass sie sinnstiftend, entdeckungsoffen und selbstdifferenzierend sind sowie reflexive Aspekte enthalten (Winter, 2015; Wittmann & Müller, 1990; Wynands, 2012). Des Weiteren zielt intelligentes Üben nicht auf das lineare Abarbeiten von vorgezeichneten Tätigkeiten ab, sondern es ermöglicht den Schülerinnen und Schülern, auf ihrem individuellen Leistungsstand eigene Lösungswege zu gehen und für sich einen Nutzen aus den Aufgaben zu ziehen (Vygotsky, 1978).

Angemessenheit der Beispiele: Mathematische Gegenstände sind ohne den Einsatz von Beispielen im Unterricht nicht behandelbar (Freudenthal, 1983; Vom Hofe, 1995). Aus diesem Grund ist es für einen guten Mathematikunterricht zentral, dass die verwendeten Beispiele geeignet sind, d. h. dass sie einerseits das Wesen des Inhalts wiedergeben, und dass sie andererseits auch nachvollziehbar und verständlich sind. Dies bedeutet, dass durch ein geeignetes Beispiel fundamentale Einsichten über die zugrunde liegende Theorie gewonnen werden können (Klafki, 1964), was insbesondere dann zum Tragen kommt, wenn neue Begriffe oder Verfahren eingeführt werden. Ein Beispiel für einen solchen mathematischen Gegenstand sollte dabei im besten Fall

nicht nur einen Bezug zum Alltag der Schülerinnen und Schüler aufweisen, sondern auch auf andere mathematische Gebiete hin „ausstrahlen“ (Wagenschein, 1999).

(Inhaltliche) Relevanz: In der psychologischen Forschung, aber auch in der deutschsprachigen Erziehungswissenschaft, ist es Konsens, dass Lernende Inhalte am ehesten lernen und behalten, wenn diese für sie individuell relevant sind (Klafki, 1964; Rakoczy, 2008; Wagenschein, 1999). Dies ist gerade im Mathematikunterricht wichtig, da die Bedeutung der Lerninhalte hier besonders häufig unklar zu sein scheint (Baruk, 1989; Vinner, 2007). Um die mathematischen Inhalte für die Schülerinnen und Schüler verständlicher und besser erlernbar zu machen, ist es deshalb notwendig, die persönliche oder inhaltliche Relevanz der im Unterricht behandelten Inhalte durch Alltagsbezüge, handlungsbezogene Materialien und vertraute Phänomene aufzuzeigen (Schoenfeld, 2013). Hierdurch wird für die Lernenden transparent, „wozu“ bestimmte Inhalte im Unterricht behandelt werden und was sie mit dem Alltag des Einzelnen zu tun haben (Rakoczy & Pauli, 2006). Dies ist möglich, indem die Lernenden zum Beispiel eigene Erfahrungen und Interessen einbringen und diskutieren. Darüber hinaus kann die Lehrperson die Relevanz der Inhalte auch dadurch verdeutlichen, indem sie den Schülerinnen und Schülern die vielfältigen Anwendungsbezüge von Mathematik im Alltag aufzeigt (Kaiser, 1995; Kaiser, Blum, Borromeo Ferri & Greefrath, 2015).

5.3 Desiderata des Beobachtungsinstruments

Aus theoretischen Gründen wurden bei der Konzeption des Instruments neben den bereits beschriebenen Items weitere Aspekte von Unterrichtsqualität als relevant eingestuft, die ebenfalls erfasst werden sollten. Hierzu zählten die Trennung von Lern- und Leistungssituationen und die Adaptivität der Lehrerhandlungen im Unterricht (Rakoczy & Pauli, 2006; Van de Pol, Volman & Beishuizen, 2010; Wood, Bruner & Ross, 1976). Beide Aspekte sind aus normativen und empirischen Gründen relevant für guten Unterricht. Im Verlauf der ersten Datenerhebungen zeigte sich jedoch, dass im Rahmen der Studie die externen Beobachterinnen und Beobachter diese Aspekte im Unterricht weder reliabel noch valide erfassen konnten, da hierzu weitere Informationen notwendig sind. Es ließ sich „von außen“ nicht einschätzen, inwieweit die beobachtete Unterrichtssituation bewertet wird und damit eine Leistungssituation für die Lernenden darstellt. Durch die Vergabe „mündlicher“ Noten über das Schuljahr hinweg ist die Bewertung der Lernenden in vielen Fällen als permanente Hintergrundbedingung vorgegeben. Selbst konkrete Äußerungen der Lehrperson dahingehend, dass die Unterrichtssituation nicht bewertungsrelevant sei, hätten dieses Problem nicht grundsätzlich gelöst, da nicht geklärt wäre, inwieweit die Lehrperson diese Absprache umsetzt und damit für die Lernenden glaubhaft ist. Hinzu kommt, dass stets nur zwei Doppelstunden je Lehrperson beobachtet wurden und der langfristige Umgang

mit Bewertungssituationen auch dahingehend nicht eingeschätzt werden kann. Ebenso ließ sich Adaptivität im Unterricht nicht beobachten, da die Lehrperson aufgrund der Anlage der Studie und der organisatorischen Zwänge, die im normalen Schulalltag herrschen, nicht zu ihren Handlungsentscheidungen in bestimmten Situationen befragt werden konnte (etwa durch nachträgliches lautes Denken (Weidle & Wagner, 1994)). Aus diesen Gründen wurden diese Items aus dem Beobachtungsinstrument entfernt. Ähnliche Probleme zeigten sich auch bei der Entwicklung anderer Instrumente wie dem amerikanischen MQI-Instrument (Learning Mathematics for Teaching Project, 2011).

5.4 Analyseeinheit und Training der Rater

Die Unterrichtsbeobachtungen im Rahmen der Studie TEDS-Unterricht wurden von insgesamt sechs trainierten Ratern durchgeführt. Für jede Unterrichtsbeobachtung wurden zwei Rater zufällig ausgewählt, so dass einer systematischen Konsensbildung durch bestimmte Raterteams vorgebeugt werden konnte. Die Rater hatten alle mindestens ein abgeschlossenes Bachelorstudium im Bereich Lehramt Mathematik, vier der sechs Rater hatten sogar bereits ein abgeschlossenes Masterstudium und strebten eine Promotion in der Mathematikdidaktik an. Im Vorfeld der Unterrichtsbeobachtungen fand eine umfangreiche Schulung der Rater statt, in der die Items des Beobachtungsinstruments mithilfe des Kodiermanuals intensiv besprochen wurden. Darüber hinaus wurde das Raten ebenfalls mehrfach an Videos und in-vivo bei realen Unterrichtsbeobachtungen trainiert. Das Training umfasste damit insgesamt über 30 Stunden in einem Zeitraum von ca. vier Wochen.

Jede Lehrperson wurde insgesamt zweimal im Unterricht besucht. Die Unterrichtsbesuche wurden im Vorfeld mit den Lehrpersonen abgesprochen, und die Unterrichtsstunden hatten jeweils einen Umfang von 90 Minuten. Es handelte sich also um Doppelstunden, die ohne Pausen durchgeführt wurden und die im Folgenden als Unterrichtsstunde bezeichnet werden. In einer Unterrichtsstunde wurde die Qualität mithilfe des Beobachtungsinstruments viermal in annähernd gleichen Zeitintervallen bewertet, so dass im Mittel nach 22,5 Minuten, 45 Minuten, 67,5 Minuten und am Ende der Stunde nach 90 Minuten ein Rating stattfand. Beide Rater bewerteten den Unterricht dabei unabhängig voneinander, jedoch wurde durch Handzeichen sichergestellt, dass das Rating zur gleichen Zeit im Unterricht erfolgte. Das mehrfache Rating im Unterricht zielte darauf ab, einerseits die in-vivo Bewertung für die Rater zu erleichtern und andererseits eine Qualitätsentwicklung im Laufe einer Unterrichtsstunde abzubilden.

Zusätzlich zu den Ratings wurde ein Unterrichtsprotokoll zur beobachteten Stunde angefertigt, in dem die Zeiten, Sozialformen, Lehrer-Schüler-Interaktionen, fachlichen Inhalte und verwendeten Aufgaben und Medien festgehalten wurden.

Nach einer Unterrichtsstunde fand eine ausführliche Nachbesprechung der beiden Rater statt. Bei dieser Nachbesprechung wurde die Stunde anhand des Protokolls noch einmal

rekonstruiert und alle Ratings wurden besprochen. Diese Nachbesprechung diente dazu, den Ratern noch einmal die Möglichkeit zu geben, einzelne Ratings zu korrigieren, da das schnelle Rating im laufenden Unterricht die Gefahr barg, dass Wahrnehmungs- und Bewertungsfehler passieren, die nicht aus inhaltlichen Gründen, sondern aus Zeitgründen entstehen. Die Nachbesprechung zielte nicht darauf ab, eine einheitliche Bewertung zwischen den Ratern zu erreichen. Im Gegenteil war es auch möglich, dass ursprünglich einheitliche Ratings nach der Nachbesprechung divergierten.

Die Nachbesprechung der Ratings zielte damit auch darauf ab, die inhaltliche Validität zu erhöhen, da durch das Fehlen von Videoaufzeichnungen der Unterrichtsstunden keine Möglichkeit bestand, eine „objektive“ Klärung der Geschehnisse im Unterricht zu erreichen, und es auch nicht möglich war, die Unterrichtsstunden bei Bedarf erneut zu bewerten.

6 Publikation I

Die erste Publikation dieser publikationsbasierten Dissertation mit dem Titel „*Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations*“ ist 2016 in der Zeitschrift *ZDM Mathematics Education* im Rahmen des Themenheftes „*Perception, interpretation and decision making: understanding the missing link between competence and performance*“ erschienen, das von Sigrid Blömeke und Jon R. Star herausgegeben wurde. Sie behandelt einerseits die Frage nach einer theoretischen Konzeptualisierung von Aspekten mathematikdidaktischer Unterrichtsqualität und andererseits untersucht sie sowohl theoretische als auch methodologische Herausforderungen bei der Erfassung einer mathematikdidaktischen Unterrichtsqualität mit externen Ratern.

Damit wird in dieser Publikation die theoretische Grundlage für die Entwicklung des vorliegenden Beobachtungsinstruments dargelegt, indem einerseits Herausforderungen bei der theoretischen Konzeptualisierung und Operationalisierung bestehender Beobachtungsinstrumente zur Erfassung von Unterrichtsqualität im Mathematikunterricht dargestellt werden und andererseits methodologische Herausforderungen bei der Datenerhebung (beispielsweise Anzahl der Rater und Stunden) thematisiert werden. Ziel ist es dabei, theoretische und methodologische Gemeinsamkeiten und Unterschiede bisheriger Forschungsprojekte in dem Bereich herauszuarbeiten und auf Basis dieser Analysen erste Ansätze für unser Beobachtungsinstrument aufzuzeigen. Der Artikel gibt damit auch einen Überblick über standardisierte Beobachtungsinstrumente und Forschungsprojekte, um zukünftigen Forschungsvorhaben die Auswahl bestehender Instrumente zur Erfassung der Unterrichtsqualität zu erleichtern. Ziel ist es dabei nicht, ein Beobachtungsinstrument als „bestes Instrument“ herauszustellen, da die Beurteilung immer abhängig von konkreten Forschungsvorhaben und Zielsetzungen ist.

Die Hauptergebnisse der vorliegenden Publikation lassen sich ebenfalls in die zwei Bereiche theoretische Konzeptualisierung und methodologische Herausforderungen unterteilen. Bei der theoretischen Analyse bisheriger standardisierter Beobachtungsinstrumente, die für den Mathematikunterricht entwickelt wurden, konnten zwei Perspektiven auf fachspezifische Qualitätsaspekte herausgearbeitet werden. Es konnte eine mathematische, stoffbezogene Perspektive und eine pädagogische, unterrichtsbezogene Perspektive auf fachspezifische Aspekte identifiziert werden. Darüber hinaus konnten einige Forschungsprojekte hilfreiche Antworten bezüglich methodologischer Entscheidungen geben. Diese beziehen sich jedoch oftmals nur auf das in dem Projekt verwendete Beobachtungsinstrument und die spezifischen Daten und lassen sich nicht direkt auf andere Projekte und Instrumente übertragen, weshalb ein projektübergreifender Konsens bisher aussteht.

6.1 Darlegung des eigenen Anteils an der Publikation

Die vorliegende Publikation ist in enger Zusammenarbeit mit meinem Kollegen Armin Jentsch von der Universität Hamburg entstanden, der ebenfalls im Rahmen des Projektes TEDS-Unterricht promoviert. Während mein Schwerpunkt dabei auf der theoretischen Konzeptualisierung und Entwicklung unseres Beobachtungsinstruments liegt, hat Armin Jentsch sich intensiver mit methodologischen Fragestellungen und Analysen in Bezug auf unser Beobachtungsinstrument befasst. Entsprechend lassen sich auch die Abschnitte der vorliegenden Publikation jeweils stärker meiner Arbeit beziehungsweise der Arbeit von Armin Jentsch zuordnen. Die beiden Abschnitte 1 und 2 wurden zum Großteil von mir verfasst, wohingegen Abschnitt 3 zusammen mit Armin Jentsch entstanden ist. Abschnitt 4 wurde ebenfalls in Zusammenarbeit mit Armin Jentsch verfasst. Insbesondere die Analysen und Zusammenstellungen bisheriger Forschungsprojekte und Beobachtungsinstrumente in den Tabellen 1 und 2 erfolgten von mir. Die zweimalige Überarbeitung des Artikels, die auf Basis des durchgeführten peer-reviewed Begutachtungsprozesses durch drei externe Reviewer nötig war, wurde gemeinsam von Armin Jentsch und mir vorgenommen. Die Einreichung und Betreuung im Editorial Manager der Zeitschrift *ZDM Mathematics Education* erfolgte durch mich.

6.2 Abdruck der Publikation I

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations, *ZDM Mathematics Education*, 48(1), 29-40.

Reprinted by permission from Springer Customer Service Centre GmbH: Springer Nature, ZDM Mathematics Education, Schlesinger & Jentsch, Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations, Copyright © 2016.

Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations

Lena Schlesinger¹ · Armin Jentsch¹

Accepted: 5 February 2016 / Published online: 19 February 2016
© FIZ Karlsruhe 2016

Abstract In this article, we analyze theoretical as well as methodological challenges in measuring instructional quality in mathematics classrooms by examining standardized observational instruments. At the beginning, we describe the results of a systematic literature review for determining subject-specific aspects measured in recent lesson studies in mathematics education. The main results are that there is little or no consistency in the conceptualization and nomination of subject-specific aspects. We therefore structured these different aspects along two perspectives, a mathematical perspective on mathematics educational quality of instruction as well as a pedagogical perspective. Furthermore, referring to the usage of these observational instruments in the field, in this paper we inquire into methodological challenges in measuring instructional quality in mathematics classrooms, e.g., the optimal number of raters and lessons to be observed. The results are twofold: on the one hand, there are recent studies that provide a useful answer to these questions. On the other hand, these results appear to be specific to the given data. Therefore, this problem seems to be unsolved so far.

Keywords Instructional quality · Methodological challenges · Classroom observations

1 Introduction

Within the last decade, research on both teachers' professional competencies and teachers' performance in the classroom has been of major interest in mathematics education (Baumert et al., 2010; Charalambous, & Hill, 2012; Hill et al., 2008; Learning Mathematics for Teaching Project, 2011; Schmidt et al., 2007). As these characteristics have a long tradition of being polarized, a great step forward in the field of educational research has been the framework of Blömeke, Gustafsson, and Shavelson (2015). In this framework, competence is conceptualized as a continuum including teachers' dispositions, their situation-specific skills and their performance, namely the observable behavior in real classroom situations, with the situation-specific skills comprising the competence facets *Perception*, *Interpretation* and *Decision-making* (in short PID-model).

Recent studies focused mainly on the relation between teacher competence and students' achievements, analyzing to what extent competence is directly predictive of students' outcomes. With regard to the processes between these characteristics, instructional quality received increased attention as a variable mediating the relation between teachers' competence and students' achievements (Baumert et al., 2010; Kersting et al., 2012; Hill, Rowan, & Ball, 2005). The question of instructional quality or what good instruction *is* has a long history in educational research (e.g. Oser, Dick, & Patry, 1992). Depending on different research traditions, the approach to this question and possible answers vary. Recent foci emphasized methods of teaching, instructional goals, as well as learning theories, which resulted in educational reforms in several countries (Atweh, Clarkson, & Nebres, 2003; Matsumura, Garnier, Pascal, & Valdés, 2002; Sawada et al., 2002). Particularly in mathematics education and mathematics instruction “scholars, policymakers,

✉ Armin Jentsch
armin.jentsch@uni-hamburg.de

Lena Schlesinger
lena.schlesinger@uni-hamburg.de

¹ University of Hamburg, Hamburg, Germany

and educators have spent decades debating what ‘counts’ in mathematics classrooms” (Learning Mathematics for Teaching Project, 2011, p. 25).

Quality of instruction is discussed as one of the main influential factors on students’ learning and achievement (Hattie, 2009; Hill et al., 2005). Seidel and Shavelson (2007) identified the largest teaching effects on students’ outcomes for domain-specific aspects of teaching. Hiebert and Grouws (2007) claimed that “the nature of classroom mathematics teaching significantly affects the nature and level of students’ learning” (p. 371). However, the most relevant subject-specific aspects of instructional quality in mathematics education have yet to be identified as these aspects are inconsistent in different studies. In addition, depending on learning goals, it is not unambiguous how mathematics teaching can be effective (Hiebert & Grouws, 2007).

Several instruments have recently been developed for measuring instructional quality; amongst others, in Germany this development was influenced by the unsatisfactory achievements of German students in international large scale assessments such as TIMSS or PISA (Beaton, Mullis, Martin, Gonzales, Kelly, & Smith, 1996). In addition, research on instructional quality has a long tradition in the U.S. resulting for example in a large number of instruments on the measurement of educational reforms (Learning Mathematics for Teaching Project, 2011; Smith & Gorard, 2007; Scheerens & Bosker, 1997). Overall, a major goal of these approaches is to describe teachers’ professional competence and classroom practice in order to improve teaching and teacher education (Scheerens, 2004).

However, most of these instruments, which have been developed by researchers and practitioners, do not consider subject-specific aspects of instructional quality as the instruments are used for different subjects, which means that subject-related aspects, for example concerning mathematics education, are not part of the evaluation.

Not unexpectedly, methodological challenges arise when measuring instructional quality reliably, particularly when regarding effects on students’ outcomes (Hill et al., 2010, 2012; Praetorius et al., 2014). However, these challenges have frequently not been discussed in empirical research publications so far, although research literature points out that influences of teacher competence on instructional quality and its effects on students’ achievement depend strongly on methodological considerations (e.g. Hiebert & Grouws, 2007; Praetorius et al., 2012).

The aim of this paper is therefore twofold: we focus on theoretical as well as methodological challenges in measuring instructional quality in mathematics education with classroom observations which means observations of lessons (videos as well as live observations) by external persons using standardized rating instruments. Firstly, the

paper aims at conceptualizing instructional quality and discusses how to integrate subject-specific aspects into a generic model. For this purpose, we provide an overview of important recent instruments measuring instructional quality in mathematics education. Secondly, the paper examines methodological considerations in measuring instructional quality using observer ratings. These analyses finally lead to prospects concerning the future development of the measurement of instructional quality in mathematics education. Our aim is not to identify the “best” or “most adequate” instrument for measuring instructional quality. In contrast, we present theoretical and methodological challenges to researchers in mathematics education, for facilitating the selection of instruments measuring instructional quality in mathematics classrooms.

2 Conceptualization of instructional quality

In the last decades, *teaching effectiveness* research seems to be one of the most prominent strands in the fields of psychology and education (Brophy, 2000; Kersting et al., 2012; Oser et al., 1992). In this strand, instructional quality is understood in a more functional way, i.e., the main goal is to predict students’ achievements at school (Seidel & Shavelson, 2007). These studies are often based on the *process-mediation-product paradigm* (Brophy, 2000, 2006), which emphasizes relations between aspects of instruction as “opportunities to learn” provided by the teacher (process), students’ usage (mediation) and their achievement (product). This framework has been modified to the *utilization of learning opportunities model* (Fend, 1981; Helmke, 2012), which is based on the constructivist idea that students’ learning processes cannot be controlled from outside. The teacher’s task is to provide learning opportunities that have to be used effectively by students in order to develop their achievements. Scholars conducting empirical research tend to describe instructional quality by a setting of characteristics independently from certain instructional designs. As recent studies on instructional quality have created many individual results it has become important to structure or “meta-analyze” the relevant factors of effective instruction (Seidel & Shavelson, 2007). However, these approaches were mainly *inductive*, i.e. they were not based on learning theories (Pianta & Hamre, 2009).

Recently, another framework for instructional quality has gained attention. It has been developed by several studies from German-speaking countries within the TIMSS Video Study and consists of three dimensions which are called *classroom management*, *personal learning support* and *cognitive activation* (see Fig. 1). It treats instructional quality in more detail, but separately

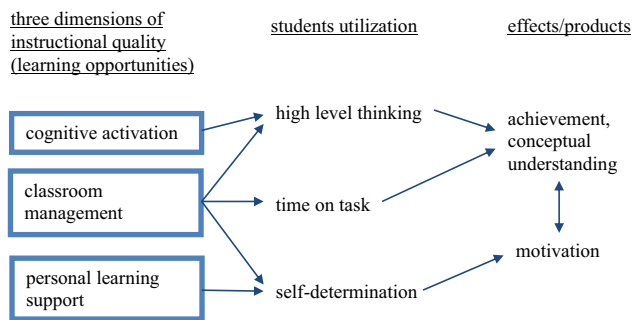


Fig. 1 Three-dimensional model of instructional quality (reproduced from Klieme & Rakoczy, 2008, p. 228)

from other facets like subject, school and grade even though it was firstly developed for mathematics instruction (Baumert et al., 2010; Lipowsky et al., 2009; Lotz, Lipowsky, & Faust, 2013; Helmke, 2012). Several studies have used this or a similar three-dimensional framework as a foundation for further empirical research (Klieme, Pauli, & Reusser, 2009; Pianta & Hamre, 2009). The predictive validity of these three basic dimensions on students' achievement has already been pointed out (Baumert et al., 2010; Klieme et al., 2009; Kunter et al., 2013; Lipowsky et al., 2009).

The first dimension, *classroom management*, focuses on quality-oriented learning time provided for students. Amongst other aspects, this dimension focuses on how effectively the teacher deals with disruptions or disciplinary conflicts (Brophy, 2000; Kunter, Baumert, & Köller, 2007). Effective *classroom management* is characterized by a structured and well-organized lesson with clear rules and routines (Lipowsky et al., 2009; see also Kounin, 1970).

The second dimension, called *personal learning support*, focuses on aspects of self-determination theory (Deci & Ryan, 1985). This dimension includes students' individual support provided by differentiation, the creation of a positive learning climate with a good relationship between students and teacher as well as constructive feedback (Klieme et al., 2009; Lipowsky et al., 2009).

Finally, *cognitive activation* refers to a high level of students' thinking supported by the teacher with problem-solving tasks to activate learning and understanding processes (Brophy, 2000; Hiebert & Grouws, 2007; Klieme et al., 2009; Lipowsky et al., 2009). This dimension includes the activation of previous knowledge, the activity of co-construction beginning with student ideas as well as challenging tasks and questions (Lipowsky et al., 2009; Praetorius et al., 2014). With regard to subject-specific aspects of instructional quality in mathematics education, the conceptualization of the third dimension can differ largely between studies (see Sects. 2.1, 2.3).

2.1 Instructional quality: generic or subject-specific?

Since the three-dimensional framework we just described was developed to hold for many different subjects, it does not sufficiently describe the *mathematics educational* quality of instruction. It becomes obvious that the domain *content* is not completely covered (Klieme & Rakoczy, 2008; Praetorius et al., 2014). However, the importance of subject-specific aspects on instructional quality has been claimed widely from theoretical and empirical perspectives (Drollinger-Vetter, 2011; Hiebert & Grouws, 2007; Klieme & Rakoczy, 2008; Seidel & Shavelson, 2007).

Despite these claims, the *relation* between subject-specific and generic aspects of instructional quality has been treated seldom in mathematics education and it is not clear until now to what extent the generic dimensions include subject-specific aspects especially regarding *cognitive activation*. Furthermore, the dimensionality of subject-specific aspects has rarely been addressed in mathematics education. Various studies aimed at solving this problem by either integrating a few subject-specific aspects into the three basic dimensions of instructional quality or by creating additional dimensions. We explore these approaches first before proposing a more systematic way of including subject-specific aspects in educational quality of instruction in Sect. 2.3.

Whether a dimension could be understood as subject-related depends essentially on its operationalization (Baumert et al. 2010; Drollinger-Vetter 2011). Educational researchers have mostly agreed on what they mean by *classroom management*, whereas other dimensions of instructional quality are not as well-established (Praetorius et al., 2014). *Classroom management* is mainly regarded as a generic dimension, because the aspects of this dimension are important in all subjects (e.g., clear rules and routines and well-organized lessons, see description in Sect. 2). Still, it is possible that reasons for students' disruptions are partly content-related due to domain-specific interests or content-specific mental under- or overload (Drollinger-Vetter, 2011, p. 325). Furthermore, there are different kinds of operationalization of *structural clarity*. Several researchers classify *structural clarity* as a generic sub-dimension of *classroom management* focusing on *organizational* structure and clarity (e.g. Helmke, 2012; Kunter et al., 2007). In contrast, other researchers regard *structural clarity* as a subject-specific dimension focusing amongst others on the development and implementation of mathematical concepts. As an example, in the German-Swiss project *Instructional Quality and Mathematical Understanding in Different Cultures* (so-called "Pythagoras-study"; Klieme et al., 2009) subject-specific aspects were measured separately beyond the three-dimensional generic framework. This subject-specific dimension consists of elements of understanding, the

quality of representations and structural clarity in a content-related operationalization (Drollinger-Vetter & Lipowsky, 2006; Drollinger-Vetter, 2011, p. 179). However, it remains uncertain whether all authors who use the term "structural clarity" refer to the same kind of structure.

The dimension *personal learning support* can also be regarded as generic, focusing on positive learning climate and the relationship between students and teacher (Klieme & Rakoczy, 2008; Pianta & Hamre, 2009). Still, various kinds of operationalization differ between studies as other researchers include content-specific activities that support students' learning (Baumert et al., 2010). In mathematics classes, subject-specific aspects of *personal learning support* could be content-related adaptive support, inner differentiation based on different content foci, as well as a positive approach to students' conceptual errors or misunderstandings.

Especially the operationalization of *cognitive activation* differs largely between various studies and it is ambiguous, to what extent *cognitive activation* contains subject-specific aspects. Beyond the approach to use "challenging tasks" or foster "high-level thinking", it is still not clear what such a dimension could consist of. Several researchers describe *cognitive activation* by aspects like "activating previous knowledge", "building on students' ideas" or "stimulating students to explain their solution methods" (Klieme et al., 2009), whereas others include aspects like "scaffolding" (Schoenfeld, 2013). Especially when focusing on students' mathematical concepts and solution methods, these aspects are mainly subject-specific. However, aspects like "fostering high-level thinking" or "using scaffolding" might occur not only in mathematics education, but in other subjects too and can therefore be described as generic. For instance, researchers in the German study *Professional Competence of Teachers, Cognitively Activating Instruction, and Development of Students' Mathematical Literacy* (in short COACTIV-study; Baumert et al., 2010) developed a framework according to the three basic dimensions of instructional quality with a focus on the *potential for cognitive activation*. By doing so, *cognitive activation* was understood from a content-specific perspective with a focus and an emphasis on cognitively activating and mathematically demanding tasks (Baumert et al., 2010; Klieme et al., 2009; Lipowsky et al., 2009).

So, even if one might use the same framework and the same dimensions of instructional quality, it is possible that the interpretation of a certain dimension and the subject-specific depth may vary greatly. This leads to several empirical problems as analyses between instructional quality and students' achievement can differ largely between different studies due to known conceptual differences (Seidel & Shavelson, 2007).

Overall, there is a shortcoming of frameworks that consist only of generic dimensions, namely, they do not seem

to suffice as a theoretical framework to describe mathematics instruction completely. Even if several aspects of instructional quality are partly subject-specific, be it *cognitive activation* or *structural clarity*, many important details of teaching and learning are not addressed, such as language, representations, or correctness of the results. With regard to mathematics education and instructional quality, Blum, Dürke-Noe, Hartung, and Köller (2006) therefore refer to an educational approach describing mathematics education as orchestration of rich subject-related content beyond the three generic dimensions. In the following we use *mathematics educational quality of instruction* (MEQI) to describe those aspects of instructional quality that are specific to mathematics as a subject.

2.2 Subject-specific aspects of instructional quality

Conceptualizing subject-specific aspects of instructional quality seems yet to be a major theoretical challenge in educational research: "if educators could more satisfactorily describe and measure the MQI [mathematical quality of instruction], they would be in a better position to improve teaching and learning" (Learning Mathematics for Teaching Project 2011, p. 30).

In the following, we shall concentrate on subject-specific aspects of instructional quality in detail and how they have been addressed by recent studies. In order to gain a representative insight into the field we first conducted an explorative survey and looked at a wide range of different instruments that measure instructional quality, although most of these instruments do not focus on subject-specific aspects of instructional quality. For validating our findings we then conducted a systematic literature review within the international databases of Web of Science, ERA and ERIC. In the description of our research we included only articles from social sciences, educational sciences or psychology. Our selection criterion was the publication source of the papers or the instruments, restricting ourselves to papers coming from peer-reviewed journals that have been published within the last 20 years. The following keywords were used within the selection process: *instructional quality*, *quality of instruction*, *teaching quality*, *educational quality*. We crossed each of these keywords with the following terms: *mathematics*, *mathematics education* or *mathematics instruction*. By reviewing the titles and abstracts, we then excluded articles that focus on a specific part of educational science, i.e., distance education, special education or pre-school education. Finally, we excluded papers that do not use classroom observations to assess instructional quality.

In the following table we have not analyzed a given paper itself; the focus was on the standardized observational instrument that is described in this paper. As an example, there has apparently been more than one publication on the

MQI instrument, but in order to provide clarity we have listed the MQI instrument referring only to one publication instead of listing every publication in which the instrument is described. We want once again to highlight that the table does not include classroom observation instruments that do not focus on subject-specific aspects in mathematics education, as for example the CLASS instrument (Pianta & Hamre, 2009). Moreover, due to unpublished descriptions of potential subject-specific aspects it was not possible to find examples for each dimension listed in the table. Generic dimensions are not included in the table, as well as the development of the presented instruments; we are focusing only on the final conceptualizations.

We are aware of the fact that we might have missed important contributions in our study; however, this exploratory approach will give an overview of studies that have been carried out in this field. In order to get an overview of different instruments measuring instructional quality, we included the results of other research in mathematics education based on literature reviews (Learning Mathematics for Teaching, 2011; Schoenfeld, 2013). As in our case, the authors of these publications, too, could find only a small number of studies and instruments.

When looking at the different subject-specific aspects or dimensions, it becomes obvious that there is no consistency in either the nomination or the conceptualization of subject-specific aspects. In addition, it is confusing that some studies use the term *dimension* whereas other studies talk only about *aspects* or even *items*. Therefore, one aspect of a given study could cover a whole dimension within another study which makes it very difficult to compare the instruments described in Table 1. Even more generally, the dimensionality of subject-specific aspects has rarely been studied empirically in mathematics education. In addition, it shows that the relevant aspects of mathematics educational quality have yet to be identified, not to mention the relation between those aspects. On the other hand, it becomes apparent that there are some aspects that are included in many of the above listed instruments (e.g. representations, demanding tasks).

Therefore, a first approach to identify these commonalities can be to find out the aspects that are included in more than one instrument. The aspects covered are:

- Representations
- Mathematical language
- Mathematical content and topics (e.g. problem solving, reasoning)
- Connections, relations, abstractions and generalizations (i.e. mathematical richness)
- Mathematical errors, mathematical correctness
- Elements of understanding
- Instructional practices (classroom practices)
- Implementation of the task
- Students' participation and understanding
- Cognitive demand, cognitive activation, potential of the task
- Materials, manipulatives.

When analyzing these different conceptualizations, it is also important to consider that some instruments were developed for special content areas (in particular introduction to the Pythagorean Theorem), whereas others were not even developed for mathematics instruction solely. We will discuss this fact in Sect. 3.2 when describing methodological challenges. In the following, we develop some suggestions of how the subject-specific aspects in mathematics education could be structured in a more systematic way.

2.3 Structuring generic and subject-specific aspects of instructional quality

Overall, we can state that there is presumably more than one way to identify categories for mathematics educational quality of instruction. However, as we lack theories on the subject-specific depth of instructional quality (as mentioned in Sect. 2.1) we might as well move forward with a rather pragmatic approach. When we ignore the *terms* that were used to name subject-specific aspects in those studies (second column of Table 1), but look mainly at the content of a given aspect (see the examples), we see at least two independent approaches towards a mathematics educational quality of instruction.

Firstly, we find in the studies the attempt to define those facets of instruction that are subject-specific in a narrow way, for instance *mathematical language*, *mathematical errors or mathematical concepts, topics and connections*. By these aspects, one addresses aspects that are not relevant only in classroom practice but also in other situations that deal with mathematics; that is, these aspects might come from a more *mathematical perspective* on mathematics educational quality of instruction. In most cases, it is not viable to study *mathematical concepts* in other than mathematics classes, in contrast to the issue of *classroom practice* or *cognitive activation*, which can be studied in all kinds of classes. Consequently, we argue that these genuine subject-specific aspects of mathematical instruction cannot be integrated into a generic framework that aims at measuring instructional quality in any classroom.

Secondly, subject-specific dimensions can be found which arise when looking at mathematics educational quality of instruction from a *pedagogical perspective*. This can be done either theoretically or in a way that is related to educational practice. In this category, we can put subject-specific aspects like *instructional practices*, *connecting classroom practice to mathematics* or *students' discussions*

Table 1 Overview of subject-specific aspects of instructional quality measured in different studies

Study	Subject-specific dimensions/aspects (<i>examples</i>)
Elementary mathematics classroom observation form (Thompson & Davis, 2014)	Computation and concepts (<i>problem-solving, reasoning, algebra, geometry, measurement, data,...</i>) Technology (<i>calculators, computers, others</i>) Manipulatives (<i>materials, real-world objects, journals, pictures, textbooks</i>)
Inside the classroom observation protocol (Horizon Research, Inc., 2000)	Mathematics content (<i>significance, appropriateness, content information, concepts, abstractions, connections, sense-making</i>)
Instructional quality assessment (IQA) (Matsumura et al., 2002)	Potential of the task Implementation of the task Student discussion following the task Rigor of expectations
Mathematical quality of instruction (MQI) (Learning Mathematics for Teaching Project, 2011)	Richness of the mathematics (<i>representations, explanations, mathematical language, multiple procedures, developing generalizations</i>) Mathematical errors and imprecisions (<i>major mathematical errors, imprecision in language, lack of clarity</i>) Students participating in meaning making and reasoning (<i>providing explanations, posing mathematically motivated questions, engaging in reasoning</i>) Working with students and mathematics Connecting classroom practice to mathematics
PERLE (Lotz, Lipowsky, & Faust, 2013)	Elements of understanding Mathematical language Representations
Pythagoras study (Klieme et al., 2009)	Elements of understanding (<i>occurrence, duration</i>) Representations (<i>enactive, iconic, formal, verbal</i>) Structural clarity
Reformed teaching observation protocol (RTOP) (Sawada et al., 2002)	Propositional knowledge (<i>conceptual understanding, fundamental concepts, abstractions, connections</i>) Procedural knowledge (<i>representations, reflection, intellectual rigor, thought-provoking activity</i>)
TIMSS video study 1999 (Hiebert et al., 2003)	Instructional practices (<i>solution methods, mathematical processes, applications, problem context</i>) Mathematical content (<i>topics, reasoning, complexity, mathematical relations</i>)
TRU Math (Schoenfeld, 2013)	Mathematical focus, coherence and accuracy (<i>richness and centrality, mathematical practices,...</i>) Cognitive demand Equitable access to content
Uteach teacher observation protocol (UTOP) (Marder, & Walkington, 2014)	Lesson structure (<i>organization, important concepts, students' understanding</i>) Implementation (<i>problem-based approach, conceptual understanding</i>) Mathematics content (<i>deep knowledge and fluidity of the teacher, abstractions, connections</i>)
Capturing teacher knowledge (Kersting et al., 2012)	Developing concepts (<i>mathematical concepts or ideas are mathematically correct</i>) Appropriate use of representations to explain algorithms (<i>manipulatives and drawn representations</i>) Connecting concepts and topics

and *students' participation*. Moreover, there are facets of mathematics educational quality of instruction that are strongly related to the three generic dimensions of instructional quality. Apparently, under this category falls the *potential of the task* from IQA, *cognitive demand* from TRU Math and *cognitive activation* from PERLE and the Pythagoras study. We can identify particular learning

theories (e.g., social constructivism) that have apparently been the conceptual foundation for these quality dimensions.

This second, pedagogical perspective of subject-specific aspects is, in our opinion, the more complicated part: due to the pedagogical perspective on mathematics instruction some aspects may be seen as generic. This is of course only

partly true, as we have discussed earlier using the examples of *cognitive activation* and *structural clarity*. Thus, we can understand these mathematics educational aspects of instructional quality as both subject-specific and also generic (or in-between). Instructional quality can be seen in a direct connection to teachers' professional knowledge, as it is then interpreted as teachers' performance in the classroom (Blömeke et al., 2015; see also Kersting et al., 2012). It may hence be possible that instructional quality can be divided into separate subject-specific dimensions, similarly to the well-known division of teachers' knowledge into MCK, PCK and GPK. Broad discussions exist especially on the operationalization of PCK (see e.g. Buchholtz, Kaiser, Blömeke, 2014), which is supposed to have a major impact on students' learning (Shulman, 1986, 1987). Nevertheless, the conceptual issues we just mentioned can be regarded only as a first step towards future developments.

3 Methods and methodological challenges for measuring instructional quality

Beyond the question of how to conceptualize instructional quality, methodological challenges should also be addressed more carefully in order to gain a better understanding of how the research design affects the psychometric quality of a study (Hill, Charalambous, & Kraft, 2012; Hill, Kapitula, & Umland, 2010). The following table gives an overview of methods and methodological issues from several studies in which the instruments from Sect. 2.2 were used (Table 1). It becomes apparent that, regarding methodological decisions, these studies vary largely.

Classroom observation is the most commonly used method to measure instructional quality directly (Clare, Valdés, Pascal, & Steinberg, 2001; Helmke, 2012).¹ These observations can be conducted by either *internal* or *external* observers (or both). Internal observers are teachers or students who attend the class. According to many authors, internal observers have some disadvantages (e.g. Hiebert & Grouws, 2007; Pianta & Hamre, 2009): as teachers and students participate actively in the lesson, they cannot concentrate exclusively on the observation. In addition, self-reports by teachers are influenced by how teachers think they should be teaching. Hence, classroom observations by external observers are awarded higher validity. In addition, external observers generally look at a huge number of lessons of different teachers, so their ratings tend to be less

subjective than the evaluation by teachers or students (Praetorius et al., 2012; Helmke, 2012).

However, there are a few well-known disadvantages that go along with external ratings (Howard et al., 1980; Lüdtke et al., 2009; Reyes et al., 2012). External observers spend little time on the observation compared with the long-term perspective of the teacher or students. In addition, the observers normally have only a little information about the class and the students. Nevertheless, these disadvantages are seen as less serious than those that go along with the evaluation by teachers or students (Clare et al., 2001; Helmke, 2012; Pianta & Hamre, 2009; Praetorius et al., 2014). As an example, Kunter and Baumert (2006) claim that there are only a few aspects of instructional quality that can be observed by teachers and students adequately, for example sampling information about the frequency of different instructional practices in a classroom or students' perceptions regarding the classroom climate or learning environment (see also Praetorius et al., 2014).

3.1 Assessing instructional quality with high-inferent observer ratings

Classroom instruction can be divided into aspects that can be observed directly (so-called *surface structures*, e.g., how many students put their hands up) and aspects that need interpretations (so-called *deep structures*, e.g., how many students participate in high-level thinking), which accounts for the amount of *inference* that is left to the observer. When analyzing an instrument, one can distinguish between items of higher or lower inference, i.e., the aspect to be measured requires more or less interpretation respectively, from the observer (Rosenshine, 1970). Nowadays, instructional quality is described mainly as a latent construct requiring interpretations, because recent research showed that the presence of aspects from surface structures and the quality of the deep structures can vary almost independently from each other (Baumert et al., 2010; Veenman, Kenter & Post, 2000). Hence, a questionnaire that contains only low-inference-items will lead to validity problems, even if the instrument is reliable. Of course, items with less inference are more likely to measure reliably and, therefore, have frequently been recommended in the past (Brophy, 2006; Kounin, 1970; Soar, Medley, & Coker, 1983). However, the main goal in measuring instructional quality must be that the ratings are not only reliable, but also valid (Hill et al., 2010; Kane, 2006; American Educational Research Association/American Psychological Association, 1999). Still, there are also some facets of instructional quality with lower inference, which can hence be evaluated more easily (e.g., students' oral participation in class vs. high-level thinking). In some cases, the indicators for the surface structure are an indication for the quality at deep

¹ Sometimes the word instructional research is used only if classroom observation methods are performed, as for example Helmke puts it: "The silver bullet of the description and assessment of instruction is without doubt observation" (own translation, 2012, p. 288).

structure; for example, mathematical high-level thinking of the students can only occur when the lesson time is connected to the learning of mathematics (i.e. time-on-task). For describing both surface and deep structures of mathematics teaching, an instrument is necessary that contains items at any inference level. The studies presented in Table 2 have measured instructional quality mainly with items of both lower and higher inference.

3.2 Reliability issues of external observer ratings

At a methodological level, the reliability of classroom observations—that is a necessary condition for valid measures—has been a major issue in the last few years (Hill et al., 2010, 2012; Praetorius et al., 2014). More precisely, the following questions are discussed in detail:

- How many lessons are sufficient in order to secure reliable data?
- How many raters are necessary?
- What is the optimal length of a rating period/time interval (i.e. *unit of analysis*)?
- How can different lessons be compared objectively?

In order to analyze these questions, the generalizability theory or, in short, G Theory is a useful psychometric method (for more information see Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). With G Theory it is possible to investigate different sources of error that can occur in classroom situations (Brennan, 2001; Shavelson & Webb, 1991; Hill et al., 2012). For every potential source of error, variance estimations are performed. The resulting G coefficients can be interpreted as reliability coefficients of classical test theory. Decision studies can then be conducted for estimating the reliability under different conditions (e.g. number of lessons and raters per class) and deciding on the optimal conditions for obtaining sufficiently reliable measures. With regard to other theories, reliability issues as mentioned above can be addressed less conditionally than in item response theory and more precisely than in classical test theory, which makes G Theory a great tool for analyzing instruments (Brennan, 2011; Hill et al., 2012). In conclusion, the decision studies enable researchers to address issues such as *rater bias*, i.e., errors caused by observers. Studies showed that errors committed by observers sum up to 41 % of the variance (Praetorius et al., 2012, 2014; for more details see Shavelson & Webb, 1991). However, in most cases, these analyses are limited to a single instrument since it is usually not possible to compare questionnaires validly, because of different concepts or terms (Hill et al., 2010; Seidel & Shavelson, 2007).

In the following, we present selected recent results on these questions and issues. The first question focuses on the variation of instructional quality between lessons and is asking, how many lessons are necessary to obtain sufficiently reliable measures for instructional quality. Until recently in empirical research, this issue has been discussed controversially by different researchers using plausibility arguments for both short periods and longer observation periods without empirical evidence. Praetorius et al. (2014) analyze this problem with data from the Pythagoras study using G theory. The results of their analysis show that the two quality dimensions *classroom management* and *personal learning support* (operationalized as content-independent) show a high stability across lessons, whereas the dimension *cognitive activation* (operationalized partly content-dependent) varied largely between lessons (Praetorius et al., 2014). In detail, one lesson is sufficient to measure *classroom management* and *personal learning support*, whereas nine lessons are needed to measure *cognitive activation* reliably. In addition, results of the MET-study suggest the benefit of observing each teacher during more than one single lesson (Gates Foundation, 2012). However, the questions remain whether these results are generalizable for further studies and how the operationalization (content-dependent vs. content-independent) influences the results.

The second question on the number of raters is discussed for the Pythagoras study by Praetorius et al. (2014). For reliability values greater than $E^2 = .90$ there are theoretically eight observers necessary for the evaluation of the dimension *personal learning support* (for more information see Brennan, 2001, 2011; Shavelson & Webb, 1991). Similar analyses were made for the MQI-instrument focusing on the number of lessons per teacher and the number of raters. The results show that the relatively small numbers of three lessons and two raters seem to be the optimal combination for their research purposes (Hill et al., 2012). Other studies, which were not presented here, found different optimal combinations that in some cases vary largely between the number of raters and lessons, characterized by measurement points (see Praetorius et al., 2014). However, these analyses depend strongly on the study within which observer reliability and rater bias have been measured and therefore a transfer to other studies is not easily possible. As Hill et al. (2012) put it: “there is no optimal number of observations or raters that transcends specific instruments and rater populations and we caution against extrapolating our results to other observational instruments and scoring designs” (p. 62). Even if no optimal number of raters generally exists, for analyzing rater bias, it is necessary that a minimum of two raters observe one lesson.

The third question addressing rating periods is another important methodological issue, which treats the *unit of analysis*. With regard to the relationship between

Table 2 Overview of methods and methodological issues for several studies measuring instructional quality in mathematics education

Study/instrument	Grades/classes	Number of lessons and ratings per lesson	Raters	Standardization of the lessons	Reliability
Algebra teaching study/TRU Math (Schoenfeld, 2013)	Grade 6–10 United States	10-min episodes	n.a.	No standardization	n.a.
Capturing teacher knowledge (Kersting et al., 2012)	Grade 5–7 36 teachers United States	1 lesson, 1 rating per lesson segment (several segments per lesson)	2 (and more)	Introduction to new fraction concept or idea	Matching percentage among raters: $>.80$
Elementary mathematics classroom observation (Thompson & Davis, 2014)	Grade 1–5 250 primary teachers Over 2000 observations United States	Two 15-min observation or 1 h per week, 24 weeks	n.a.	No standardization	n.a.
Horizon research/inside the classroom observation protocol (Horizon Research, 2000)	364 lessons United States	1 rating per lesson	n.a.	No standardization	n.a.
IQA (Matsumura et al., 2008)	Grade 6–7 13 teachers United States	2 observations, 1 rating per observation	1	Problem-solving activity; related discussions	Rater agreement (for 4 raters and 7 observations): $.11 < ICC \leq 1.0$
LMT project/MQI (Hill et al., 2012)	No specific grades Eight middle school teachers United States	6 lessons, 1 rating every 7.5 min (6–8 segments per lesson)	2	No standardization	Generalizability coefficient (for 2 raters and 3 lessons per teacher): $.68 < E\rho^2 \leq .81$ Split-half: $.87 < r \leq 1.0$
TIMSS video study 1999 (Hiebert et al., 2003/Jacobs et al., 2003)	Grade 8 638 classes Seven countries	1 lesson, 1 rating per lesson segment (several segments per lesson)	2	No standardization	
PERLE video study mathematics (Lotz, Lipowsky, & Faust, 2013)	Grade 1–2 36 classes Germany	2 lessons, 1 rating per lesson	2	Introduction to multiplication	Rater agreement: $.60 < \kappa \leq 1.0$ Generalizability coefficient: $.70 < E\rho^2 \leq .99$
Pythagoras study (Klieme et al., 2009)	Grade 8–9 37 classes Switzerland and Germany	5 lessons, 1 rating per lesson	2 (3)	Introduction to Pythagorean theorem	Generalizability coefficient: $.57 < E\rho^2 \leq .88$ Internal consistency: $\alpha = .88$
ACEPT/reformed teaching observation protocol (Sawada et al., 2002)	153 classes Schools, colleges and university United States	3 lessons, 1 rating per lesson	1–2	No standardization	Rater agreement: <i>Pearson's</i> $r = .98$ Internal consistency: $.80 < \alpha \leq .93$
MET-study/Uteach teacher observation protocol (Marder & Walkington, 2014; Gates Foundation, 2012)	Grade 4–8 249 teachers 982 videos United States	2–4 lessons, 1 rating per lesson	1–3	No standardization	Reliability measure: $.44 < ICC \leq .65$

instructional quality and students' achievement it is important to analyze meaningful units that allow descriptions of the processes in class. Hiebert and Grouws (2007) proposed the analysis of units of typical lesson periods. Still, this decision depends on the aspects that are to be measured as there can be large variance in some aspects of instructional quality within one lesson, e.g., whether the lesson time is used efficiently. Therefore, recent observational instruments use shorter and therefore more rating periods (Learning Mathematics for Teaching Project 2011). This leads to other technical challenges, e.g., how many ad-hoc ratings can be performed per lesson if no videos are recorded.

The last question focuses on the comparability of lessons. In order to gain higher comparability, it is possible to standardize content, classes, grades, teacher activities, settings or learning goals within the lesson. Domain-specific teaching studies are mainly conducted under quasi-experimental conditions (Seidel & Shavelson, 2007). This quasi-experimental approach leads to some advantages and, therefore, is used in different recent studies like the Pythagoras study or the PERLE study. In both of these studies the content was standardized, e.g., all classes in the Pythagoras study treated an introduction into the Pythagorean Theorem. The instruments used in standardized content settings allow the deepening of the subject-specific perspective (Lipowsky et al., 2009) and the specification of the subject-specific items. However, this approach may have an influence on other methodological questions; for example, the question remains of how results of these studies can be transferred to other mathematical content and also, more globally, how these results can be transferred to ordinary all-day classroom instruction without any standardization. Therefore, one main goal in future research might be to develop instruments that can be applied in all-day instruction and that are not bound to specific contents or classroom situations.

4 Summary and conclusion

Until now, there seems to be no general framework for instructional quality, especially for the purposes of mathematics education and subject-specific aspects. Moreover, the understanding of what MEQI really *is* varies greatly among studies. This shows that there is apparently no consensus about the essential issues of MEQI, its structure and its relation to generic aspects. In addition, analyzing results of classroom studies with a focus on mathematics education becomes quite difficult as almost every framework uses different terms describing different underlying concepts. Therefore, we argue that agreeing on some definitions could be helpful for further research in mathematics

education and, more generally, empirical research. As a first step, we suggest that MEQI could be studied from both a mathematical and a pedagogical perspective.

At the methodological level, we gave an overview of challenges concerning observer ratings. In detail, we focused on questions such as how many lessons and raters are necessary for securing reliable data and what is the optimal length of a rating period (unit of analysis). In addition, we discussed advantages and disadvantages that go along with quasi-experimental studies standardizing aspects of the observed lessons (e.g., the content). Still, there are other questions which could not be discussed in this paper (e.g., how raters can be trained ideally or the number of items in the observational instrument; see Hill et al., 2012; Praetorius et al., 2012). Consequently, Hill et al., (2012) use the term *observational systems* to highlight the interdependency of raters, instruments and conceptualizations in a given study. However, it is necessary also to think of systematic problems. In particular, one should address not only reliability, but even more important validity issues of instructional quality.

From the perspective of educational practice, the instruments described in Sects. 2.2 and 3.1 carry the problem of not always being easily applicable to large samples. They often require videotaping and, in order to analyze the instructional quality, need multiple inspections of videos, which leads to many organizational challenges (for more information see Casabianca et al., 2013). Hence, for conducting empirical research in mathematics classrooms it may be necessary to develop an observational instrument that can be used in ad-hoc classroom observations. Regarding cost-efficiency, statistical methods such as G theory can provide a useful benefit for practitioners as it is possible to decide on a theoretical basis how many raters and items may be necessary to gain reliable measures (see Hill et al., 2012; Shavelson & Webb, 1991). Thus, for future research it is important to use theory-based frameworks and observation instruments that fulfill both validity and reliability conditions for measuring instructional quality in mathematics education.

Acknowledgments We thank Nils Buchholtz, Andreas Busse and the reviewers for helpful suggestions and comments on earlier versions of this article.

References

- American Educational Research Association/American Psychological Association. (1999). *Standards for educational and psychological testing*. Washington: American Educational Research Association.
- Atweh, B., Clarkson, P., & Nebres, B. (2003). Mathematics education in international and global contexts. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Second*

- international handbook of mathematics education* (pp. 185–229). Dordrecht: Springer Netherlands.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133–180.
- Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzales, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study*. Chestnut Hill: Boston College.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond Dichotomies. *Zeitschrift für Psychologie*, 223(1), 3–13.
- Blum, W., Drücke-Noe, C., Hartung, R., & Köller, O. (2006). *Bildungsstandards Mathematik: Konkret. Sekundarstufe 1: Aufgabenbeispiele, Unterrichts Anregungen, Fortbildungsideen*. Berlin: Cornelsen Scriptor.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.
- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1–21. doi:10.1080/08957347.2011.532417.
- Brophy, J. (2000). *Teaching*. Brüssel: International Academy of Education.
- Brophy, J. (2006). Observational research on generic aspects of classroom teaching. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 755–780). Mahwah: Erlbaum.
- Buchholtz, N., Kaiser, G., & Blömeke, S. (2014). Die Erhebung mathematikdidaktischen Wissens—Konzeptualisierung einer komplexen Domäne. *Journal für Mathematik-Didaktik*, 35(1), 101–128.
- Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, 73(5), 757–783.
- Charalambous, C. Y., & Hill, H. C. (2012). Teacher knowledge, curriculum materials, and quality of instruction: Unpacking a complex relationship. *Journal of Curriculum Studies*, 44(4), 443–466.
- Clare, L., Valdés, R., Pascal, J., & Steinberg, J. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools (CSE Technical Report No. 545)*. Los Angeles: National Center for Research on Evaluation.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability*. New York: Wiley.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior. Perspectives in social psychology*. New York: Plenum.
- Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit: Fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht*. Münster: Waxmann.
- Drollinger-Vetter, B., & Lipowsky, F. (2006). Fachdidaktische Qualität der Theoriephasen. In E. Klieme, C. Pauli, & K. Reusser (Eds.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 3: Hugener, Isabelle; Pauli, Christine & Reusser, Kurt: Videoanalysen* (pp. 189–205). Frankfurt am Main: GFPE.
- Fend, H. (1981). *Theorie der Schule* (2., durchges. Aufl.). U- & -S-Pädagogik. München [u.a.]: Urban & Schwarzenberg.
- Gates Foundation (2012). *Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains. Research paper*, http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf. Accessed 22 Jan 2016.
- Hattie, J. (2009). *Visible learning. Synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
- Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.
- Hiebert, J., Gallimore, R., Garnier, H., & Stigler, J. (2003). *Teaching mathematics in seven countries. Results from the TIMSS 1999 video study*. Washington: National Center for Education Statistics.
- Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Charlotte: Information Age.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4), 430–511.
- Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56–64. doi:10.3102/0013189X12437203.
- Hill, H. C., Kapitula, L., & Umland, K. (2010). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 48(3), 794–831. doi:10.3102/0002831210387916.
- Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, 42(2), 371–406.
- Horizon Research, Inc. (2000). *Inside the classroom observation and analytic protocol*. Chapel Hill: Horizon Research, Inc.
- Howard, G. S., Maxwell, S. E., Weiner, R. L., Boynton, K. S., & Rooney, W. M. (1980). Is a behavioral measure the best estimate of behavioral parameters? Perhaps not. *Applied Psychological Measurement*, 4, 293–311.
- Jacobs, J., Garnier, H., Gallimore, R., Hollingsworth, H., Givvin, K. B., Rust, K., Kawanaka, T., Smith, M., Wearne, D., Manaster, A., Etterbeek, W., Hiebert, J., Stigler, J. (2003). *TIMSS 1999 video study technical report: volume 1: Mathematics study*, NCES (2003-012), U.S. Department of Education. Washington, DC: National Center for Education Statistics.
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 18–64). Westport: Praeger.
- Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, 49(3), 568–589. doi:10.3102/0002831212437853.
- Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster: Waxmann.
- Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54, 222–237.
- Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York: Holt, Rinehart and Winston.
- Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9(3), 231–251. doi:10.1007/s10984-006-9015-7.
- Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction*, 17(5), 494–509. doi:10.1016/j.learninstruc.2007.09.002.
- Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal*

- of *Educational Psychology*, 105(3), 805–820. doi:[10.1037/a0032583](https://doi.org/10.1037/a0032583).
- Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14, 25–47.
- Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, 19(6), 527–537. doi:[10.1016/j.learninstruc.2008.11.001](https://doi.org/10.1016/j.learninstruc.2008.11.001).
- Lotz, M., Lipowsky, F., Faust, G. (2013). *Dokumentation der Erhebungsinstrumente des Projekts "Persönlichkeits- und Lernentwicklung von Grundschulern" (PERLE). 3. Technischer Bericht zu den PERLE-Videostudien. Materialien zur Bildungsforschung: Vol. 23.3.* Frankfurt am Main: Gesellschaft zur Förderung Pädagogischer Forschung [u.a.].
- Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, 34, 120–131. doi:[10.1016/j.cedpsych.2008.12.001](https://doi.org/10.1016/j.cedpsych.2008.12.001).
- Marder, M., & Walkington, C. (2014). Classroom observation and value-added models give complementary information about quality of mathematics teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measuring Effective Teaching project* (pp. 234–277). New York: Wiley.
- Matsumura, L. C., Garnier, H. E., Pascal, J., & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and students achievement. *Educational Assessment*, 8, 207–229.
- Matsumura, L. C., Garnier, H., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale". *Educational Assessment*, 13, 267–300.
- Oser, F., Dick, A., & Patry, J.-L. (Eds.). (1992). *Effective and responsible teaching: The new synthesis*. San Francisco: Jossey Bass.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119. doi:[10.3102/0013189X09332374](https://doi.org/10.3102/0013189X09332374).
- Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, 22, 387–400.
- Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12.
- Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology*, 104, 700–712. doi:[10.1037/a0027268](https://doi.org/10.1037/a0027268).
- Rosenshine, B. (1970). Evaluation of instruction. *Review of Educational Research*, 40, 279–300.
- Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics*, 102(6), 245–253. doi:[10.1111/j.1949-8594.2002.tb17883](https://doi.org/10.1111/j.1949-8594.2002.tb17883).
- Scheerens, J. (2004). *Review of school and instructional effectiveness. Background paper prepared for the Education for All Global Monitoring Report 2005*. Paris: UNESCO.
- Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.
- Schmidt, W. H., Tatto, M. T., Bankov, K., Blömeke, S., Cedillo, T., Cogan, L., et al. (2007). *The preparation gap: Teacher education for middle school mathematics in six countries. Mathematics teaching in the 21st century (MT21)*. East Lansing: Michigan State University, Center for Research in Mathematics and Science Education.
- Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM-The International Journal on Mathematics Education*, 45(4), 607–621.
- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499. doi:[10.3102/0034654307310317](https://doi.org/10.3102/0034654307310317).
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks: Sage.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4–31.
- Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1–22.
- Smith, E., & Gorard, S. (2007). Improving teacher quality: Lessons from America's No Child Left Behind. *Cambridge Journal of Education*, 37(2), 191–206.
- Soar, R. S., Medley, D. M., & Coker, H. (1983). Teacher evaluation: A critique of currently used methods. *The Phi Delta Kappan*, 65, 239–246.
- Thompson, C. J., & Davis, S. B. (2014). Classroom observation data and instruction in primary mathematics education: Improving design and rigour. *Mathematics Education Research Journal*, 26(2), 301–323. doi:[10.1007/s13394-013-0099-y](https://doi.org/10.1007/s13394-013-0099-y).
- Veenman, S., Kenter, B., & Post, K. (2000). Cooperative learning in Dutch primary classrooms. *Educational Studies*, 26(3), 281–302.

7 Publikation II

Die zweite Publikation dieser publikationsbasierten Dissertation mit dem Titel „*Unterrichtsqualität unter einer mathematikdidaktischen Perspektive – Konzeptualisierung, Messung und Validierung*“, wird voraussichtlich im *Journal für Mathematik-Didaktik* erscheinen. In dieser Publikation wird einerseits untersucht, ob die Qualität des Mathematikunterrichts mit mehreren theoretisch entwickelten Dimensionen reliabel erfasst werden kann. Andererseits werden die Fragestellungen untersucht, welche Zusammenhänge sich zwischen den etablierten drei Basisdimensionen von Unterrichtsqualität und den zwei neu entwickelten fachspezifischen Qualitätsdimensionen sowie zu den Kompetenzfacetten der Lehrpersonen zeigen.

Die Publikation befasst sich zunächst mit der Konzeptualisierung und Operationalisierung des Beobachtungsinstruments mit drei fachübergreifenden und zwei fachspezifischen Dimensionen und es werden erste Daten der Studie TEDS-Unterricht präsentiert. Hierbei werden die fünf Dimensionen des Beobachtungsinstruments empirisch geprüft und es werden die Zusammenhänge zwischen den drei Basisdimensionen und den zwei fachspezifischen Dimensionen genauer analysiert. Darüber hinaus wird eine erste kriteriale Validierung der Beobachtungsdaten vorgenommen, indem ebenfalls Korrelationen zwischen den Dimensionen des Beobachtungsinstruments und den erfassten domänenspezifischen Kompetenzfacetten der Lehrpersonen im Rahmen von TEDS-Unterricht genauer untersucht werden.

Die Ergebnisse der Analysen zeigen, dass die fünf Dimensionen des Beobachtungsinstruments eindimensional erfasst wurden und gute Interrater-Reliabilitäten aufweisen. Ebenfalls sind die Retest-Reliabilität sowie die interne Konsistenz der Daten zufriedenstellend. Die beiden fachspezifischen Dimensionen weisen dabei hohe Korrelationen zur kognitiven Aktivierung auf, was die Vermutung stützt, dass diese Dimension insgesamt die größte Nähe zu fachspezifischen Aspekten aufweist. Darüber hinaus lassen die korrelativen Analysen zu den professionellen Kompetenzen der Lehrpersonen erste Schlüsse zur kriterialen Validität des Beobachtungsinstruments zu. Zusammenhänge mittlerer Effektstärke zeigen sich zwischen den beiden fachspezifischen Qualitätsdimensionen und den proximalen Fähigkeiten der Lehrpersonen, Unterricht unter einer mathematikdidaktischen Perspektive wahrzunehmen und zu interpretieren. Darüber hinaus lassen sich schwache Korrelationen zwischen dem mathematikdidaktischen Wissen der Lehrpersonen und der stoffdidaktischen Qualitätsdimension feststellen.

7.1 Darlegung des eigenen Anteils an der Publikation

Die vorliegende Publikation ist wiederum in enger Zusammenarbeit mit meinem Kollegen Armin Jentsch von der Universität Hamburg entstanden. Die Entwicklung des Beobachtungsinstruments, die Datenerhebung sowie die Auswertung der Beobachtungsdaten

wurden federführend von Armin Jentsch und mir durchgeführt. Insbesondere die Organisation der Unterrichtsbeobachtungen sowohl der Pilotierung als auch der Haupterhebung oblag dabei mir. Die intensive Schulung der Rater wurde ebenfalls von Armin Jentsch und mir vorgenommen. Das Kodiermanual für die Items des Beobachtungsinstruments wurde dafür von uns verfasst.

Der erste Entwurf der vorliegenden Publikation wurde überwiegend von Armin Jentsch und mir verfasst. Dabei übernahm Armin Jentsch insbesondere die Abschnitte, die sich mit der Datenauswertung befassen. Darüber hinaus war die Projektleitung von TEDS-Unterricht (Gabriele Kaiser, Johannes König und Sigrid Blömeke) bei der methodologischen Fundierung sowie erziehungswissenschaftlichen und mathematikdidaktischen Fundierung intensiv beteiligt. Da sich jedoch Ziel und Anlage der Publikation mehrfach substanziell geändert haben, gab es im Vergleich zu den beiden anderen Publikationen mehrere intensive Überarbeitungsschleifen, in denen sich die Struktur des Artikels, die Ziele sowie die verwendete Datengrundlage mehrfach geändert haben. Dies führte dazu, dass die einzelnen Kapitel der Publikationen nicht mehr jeweils einer einzelnen Person zuzuordnen sind, sondern im Gegenteil das Ergebnis der Überarbeitung von Armin Jentsch, der Projektleitung von TEDS-Unterricht, sowie mir sind. Die Einreichung und Betreuung im Editorial Manager des *Journals für Mathematik-Didaktik* erfolgte durch mich.

7.2 Abdruck der Publikation II

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (under review). Unterrichtsqualität unter einer mathematikdidaktischen Perspektive – Konzeptualisierung, Messung und Validierung. Eingereicht als zweimalig überarbeitete Fassung im: *Journal für Mathematik-Didaktik*.

Reprint permission: ist noch einzuholen nach Annahme des Beitrags.

Journal für Mathematik-Didaktik

Unterrichtsqualität unter einer mathematikdidaktischen Perspektive - Konzeptualisierung, Messung und Validierung

--Manuscript Draft--

Manuscript Number:	JMDJ-D-18-00005
Full Title:	Unterrichtsqualität unter einer mathematikdidaktischen Perspektive - Konzeptualisierung, Messung und Validierung
Article Type:	Original Article
Corresponding Author:	Lena Schlesinger Universität Hamburg GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universität Hamburg
Corresponding Author's Secondary Institution:	
First Author:	Lena Schlesinger
First Author Secondary Information:	
Order of Authors:	Lena Schlesinger
	Armin Jentsch
	Gabriele Kaiser
	Johannes König
	Sigrid Blömeke
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Im deutschsprachigen Raum hat sich in den letzten Jahren ein Modell durchgesetzt, das Unterrichtsqualität mithilfe von drei Basisdimensionen - Klassenführung, konstruktive Unterstützung der Lernenden und ihre kognitiven Aktivierung - fächerübergreifend konzeptualisiert. Damit stellt sich aus mathematikdidaktischer Sicht die Frage, inwieweit diese mit bedeutsamen fachspezifischen Merkmalen von Unterrichtsqualität zusammenhängen. In einer theoriegeleiteten Konzeptualisierung in Anlehnung an Erkenntnisse aus der Lehrerprofessionsforschung wurden zwei Dimensionen mathematikdidaktischer Qualitätsmerkmale identifiziert, wobei die eine stoffbezogene und die andere unterrichtsbezogene Merkmale fokussiert. Auf Basis dieser Konzeptualisierung wurde ein fünfdimensionales Beobachtungsinstrument mit hoch-inferent einzuschätzenden Items entwickelt, mit dem sowohl die Basisdimensionen als auch die fachspezifischen Dimensionen des Mathematikunterrichts von 37 Lehrpersonen in je zwei Doppelstunden „in vivo“ von zwei geschulten Ratern eingeschätzt wurde. Die Ergebnisse deuten eine gute Interraterreliabilität und Eindimensionalität der fünf Skalen an. Retest-Reliabilität und interne Konsistenz fallen moderat bis zufriedenstellend aus. Als erste Hinweise auf die kriteriale Validität der beiden mathematikdidaktischen Qualitätsdimensionen können Zusammenhänge mittlerer Effektstärke zur mathematikdidaktischen Unterrichtswahrnehmung der Lehrpersonen aufgefasst werden. Darüber hinaus zeigen sich schwache Korrelationen zwischen dem mathematikdidaktischen Wissen der Lehrpersonen und der stoffdidaktischen Qualitätsdimension.
Suggested Reviewers:	

Unterrichtsqualität unter einer mathematikdidaktischen Perspektive – Konzeptualisierung, Messung und Validierung

(Instructional quality from a mathematics educational perspective –
Conceptualization, measurement and validation)

Lena Schlesinger, Universität Hamburg, lena.schlesinger@uni-hamburg.de¹

Armin Jentsch, Universität Hamburg, armin.jentsch@uni-hamburg.de¹

Gabriele Kaiser, Universität Hamburg, Australian Catholic University

Johannes König, Universität zu Köln

Sigrid Blömeke, University of Oslo, CEMO

¹Universität Hamburg, Fakultät für Erziehungswissenschaft, Arbeitsgruppe Didaktik der
Mathematik, Von-Melle-Park 8, 20146 Hamburg, Tel: 040/42838-6422

Zusammenfassung

Im deutschsprachigen Raum hat sich in den letzten Jahren ein Modell durchgesetzt, das Unterrichtsqualität mithilfe von drei Basisdimensionen – Klassenführung, konstruktive Unterstützung der Lernenden und ihre kognitiven Aktivierung – fächerübergreifend konzeptualisiert. Damit stellt sich aus mathematikdidaktischer Sicht die Frage, inwieweit diese mit bedeutsamen fachspezifischen Merkmalen von Unterrichtsqualität zusammenhängen. In einer theoriegeleiteten Konzeptualisierung in Anlehnung an Erkenntnisse aus der Lehrerprofessionsforschung wurden zwei Dimensionen mathematikdidaktischer Qualitätsmerkmale identifiziert, wobei die eine stoffbezogene und die andere unterrichtsbezogene Merkmale fokussiert. Auf Basis dieser Konzeptualisierung wurde ein fünfdimensionales Beobachtungsinstrument mit hoch-inferent einzuschätzenden Items entwickelt, mit dem sowohl die Basisdimensionen als auch die fachspezifischen Dimensionen des Mathematikunterrichts von 37 Lehrpersonen in je zwei Doppelstunden „in vivo“ von zwei geschulten Ratern eingeschätzt wurde. Die Ergebnisse deuten eine gute Interraterreliabilität und Eindimensionalität der fünf Skalen an. Retest-Reliabilität und interne Konsistenz fallen moderat bis zufriedenstellend aus. Als erste Hinweise auf die kriteriale Validität der beiden mathematikdidaktischen Qualitätsdimensionen können Zusammenhänge mittlerer Effektstärke zur mathematikdidaktischen Unterrichtswahrnehmung der Lehrpersonen aufgefasst werden. Darüber hinaus zeigen sich schwache Korrelationen zwischen dem mathematikdidaktischen Wissen der Lehrpersonen und der stoffdidaktischen Qualitätsdimension.

Schlüsselwörter:

Unterrichtsqualität, Unterrichtsbeobachtungen, Mathematikdidaktische Merkmale, Ratings, Beobachtungsinstrument, Validität

Abstract

German educational research frequently discusses a generic model with which instructional quality is described by three basic dimensions: classroom management, student support and cognitive activation. From a mathematics educator's perspective, the question arises then how these dimensions relate to domain-specific aspects of instructional quality. Building on findings from research on teacher cognition, we identify two domain-specific dimensions of instructional quality in a theory-driven approach. One of these dimensions focuses on subject-related quality while the other dimension assesses teaching-related quality in mathematics instruction. Based on this conceptualization we developed an observational protocol for high-inferent rating with which both generic and also domain-specific dimensions of instructional quality were assessed "in vivo" by two trained observers. Data contain observations from 37 teachers that were assessed in two lessons each. Findings suggest one-dimensionality for all five scales and good interrater-agreement. Measures for retest-reliability and internal consistency showed at least satisfactory results. As a first approach to validating the observational protocol, we found moderate correlations between teachers' professional noticing and domain-specific dimensions of instructional quality. In addition, we found weak correlations between pedagogical content knowledge and subject-related quality of instruction.

Keywords:

instructional quality, classroom observations, domain-specific dimensions, ratings, observational instrument, validity

1. Einleitung

Was guten Unterricht ausmacht, wird seit Jahrzehnten aus verschiedenen Perspektiven intensiv diskutiert (u.a. Klieme & Rakoczy, 2008; Helmke, 2012; Blum et al., 2015; Praetorius, 2014; Lipowsky, 2015). Theoretisch und empirisch hat sich in diesem Zusammenhang der Ansatz der empirischen Bildungsforschung bewährt, Unterrichtsqualität mittels der drei Basisdimensionen *Klassenführung*, *konstruktive Unterstützung der Lernenden* und *kognitive Aktivierung* zu beschreiben (Klieme et al., 2006). So zeigte sich mehrfach deren Wirksamkeit in Bezug auf Leistungen und Motivation von Schülerinnen und Schülern (Baumert et al., 2010; Lipowsky et al., 2009; Helmke, 2012). Nicht vollständig geklärt ist jedoch bisher, welche Zusammenhänge zwischen den drei Basisdimensionen und solchen Merkmalen bestehen, die die spezifische Qualität von *Fachunterricht* beschreiben. Zur Konzeptualisierung und Messung von mathematikdidaktischer Unterrichtsqualität wurden in den letzten Jahren neben solchen fachübergreifenden Ansätzen auch Modelle und Instrumente entwickelt, die sich speziell auf Mathematikunterricht beziehen (Hill et al., 2008; Learning Mathematics for Teaching Project, 2011).

Der vorliegende Beitrag greift diese Entwicklung auf und integriert die beiden Ansätze. Ziel ist, in Anknüpfung u.a. an Überlegungen von Brunner (2017) und Blum et al. (2015) fächerübergreifende und fachspezifische Qualitätsmerkmale von Mathematikunterricht zu beschreiben und ihre empirische Erfassbarkeit sowie ihre empirischen Zusammenhänge zu untersuchen. Dazu tragen wir Befunde aus empirischer und fachdidaktischer Unterrichtsforschung zusammen und stellen ein Beobachtungsinstrument vor, das Unterrichtsqualität anhand solcher etablierten Merkmale konzeptualisiert und erfasst. Anschließend wird das Instrument in einem korrelativen Design erprobt, hinsichtlich psychometrischer Gütekriterien überprüft und dabei insbesondere kriterial validiert.

2. Forschungsstand und theoretischer Rahmen

2.1 Konzeptualisierung von Unterrichtsqualität

Unterricht ist ein komplexes System, das von verschiedenen Bezugsdisziplinen und aus unterschiedlichen Perspektiven untersucht wird. Um die vielfältigen Prozesse im Unterricht quantitativ beschreiben zu können, wurden in den letzten Jahren von Seiten der empirischen Bildungsforschung Modelle entwickelt, die Unterrichtsqualität mithilfe unterschiedlicher Dimensionen und verschiedener Qualitätsaspekte konzeptualisieren (für einen Überblick siehe Praetorius, 2014 sowie Taut & Rakoczy, 2016).

Als entscheidender Fortschritt kann die Definition von drei Basisdimensionen der Unterrichtsqualität angesehen werden (Klieme et al., 2006). Die Trias aus Klassenführung, (dem Potential zur) kognitiven Aktivierung und, mit teils unterschiedlichen Bezeichnungen, konstruktiver Unterstützung wurde im Anschluss an die TIMSS-Videostudien (Third

International Mathematics and Science Study, Hiebert et al., 2003) entwickelt. Unter einer effizienten Klassenführung werden dabei solche Maßnahmen und Strategien verstanden, die eine optimale Nutzung der Lernzeit im Unterricht ermöglichen (Helmke, 2012). Während Kounin (1970) hierunter vor allem Methoden zur Prävention von Unterrichtsstörungen fasst, beziehen neuere Arbeiten auch den Aufbau und die Klarheit organisatorischer Regeln sowie die Strukturierung des Unterrichtsverlaufs ein (Rakoczy & Pauli, 2006). Potenzial zur kognitiven Aktivierung haben problemorientierte Lernsituationen, in denen Schülerinnen und Schüler zu komplexen Denkprozessen herausgefordert werden („high-order thinking“, zusammenfassend Lotz, 2015; Mayer, 2004). Nach einem gemäßigt konstruktivistischen Lernverständnis ist der Wissenserwerb das Ergebnis aktiver und individueller Konstruktionsprozesse, die im Unterricht in sozialen Kontexten ablaufen (Aebli, 1985; Klieme & Rakoczy, 2008; Reinmann-Rothmeier & Mandl, 2006; Shuell, 1993). Solche Unterrichtskontexte werden mit der Qualitätsdimension „Konstruktive Unterstützung“ (auch als Schülerorientierung bezeichnet) beschrieben, welche eine unterstützende Lernumgebung, die Diagnose von individuellen Lernprozessen, sowie Individualisierung und Differenzierung des Unterrichtsangebots beinhaltet. Hierdurch soll dem Bedürfnis nach Kompetenz- und Autonomieerleben von Lernenden Rechnung getragen werden (Deci & Ryan, 1985; Praetorius, 2014; Rakoczy & Pauli, 2006). Während eine effiziente Klassenführung und ein kognitiv aktivierender Unterricht vor allem kognitive Lernzuwächse fokussieren (Baumert et al., 2010; Klieme & Rakoczy, 2008), zielt die Qualitätsdimension der „Konstruktiven Unterstützung“ auf die Motivation der Lernenden ab.

Das Modell der drei Basisdimensionen ist fächerübergreifend konzeptualisiert, weshalb seine Bedeutung nicht einzelnen Fächern vorbehalten ist (Praetorius et al., 2018). Während Lotz (2015) und Maier, Kleinknecht, Metz und Bohl (2010) den Versuch unternehmen, die kognitive Aktivierung von Lernenden gänzlich unabhängig vom Fachunterricht zu beschreiben, führen Klieme und Rakoczy (2008) an, dass mit dem Potential zur kognitiven Aktivierung vor allem das fachliche Lernen gemeint und der Begründungszusammenhang für diese Qualitätsdimension aus der Fachdidaktik abgeleitet sei. Dies deutet darauf hin, dass kognitive Aktivierung im Englischunterricht möglicherweise anders zu verstehen ist als im Mathematikunterricht (im Überblick Kunter & Ewald, 2016).

Von mathematikdidaktischer Seite wird entsprechend darauf hingewiesen, dass das Modell der drei Basisdimensionen nicht alle für den Mathematikunterricht bedeutsamen Merkmale abbilde. Brunner (2017) führt an, dass die aus normativer Sicht notwendige Korrektheit der fachlichen Inhalte im Rahmen der drei Basisdimensionen unberücksichtigt bliebe. Blum (2015) weist außerdem darauf hin, dass im Mathematikunterricht der Erwerb umfangreicher Kompetenzen im Vordergrund stehe (KMK, 2003), der durch die drei Basisdimensionen von Unterrichtsqualität nicht abgebildet werde. Unter einer fachlich gehaltvollen Unterrichtsgestaltung, die er neben der kognitiven Aktivierung der Lernenden und einer effektiven und schülerorientierten Unterrichtsführung als Qualitätsdimensionen von Unterricht anführt, versteht Blum dabei reichhaltige Gelegenheiten zum Kompetenzerwerb durch geeignete Aufgaben im Unterricht (Blum et al. 2015). Die verwendeten Aufgaben und deren fachliche Qualität würden dabei im

Mathematikunterricht eine besondere Rolle spielen, da sie die zentrale Schnittstelle von Schüler- und Lehrertätigkeit darstellen sowie den Unterricht in besonderem Maße inhaltlich strukturieren und organisieren (u.a. Jordan et al., 2008). Damit stellt sich die Frage, ob sich – empirisch gesehen – tatsächlich solche fachspezifischen Merkmale von Unterrichtsqualität identifizieren lassen und ob bzw. inwiefern sie mit den drei Basisdimensionen von Unterrichtsqualität zusammenhängen.

In einem Überblicksartikel stellen Charalambous und Praetorius (2018) eine Reihe von Instrumenten zur Erfassung der Qualität von Mathematikunterricht dar, die sie entlang eines Kontinuums von generisch zu fachspezifisch ordnen. Ein Instrument wird dabei als generisch bezeichnet, wenn es Unterrichtsqualität in verschiedenen Fächern erfassen soll, während die nur für ein einziges Fach entwickelten Instrumente fachspezifisch genannt werden („subject matter“, Charalambous & Praetorius, 2018).

Der generischen Seite zugeordnet werden u.a. das *Classroom Assessment Scoring System* (CLASS, Pianta & Hamre, 2009), das *Framework for Teaching* (FfT, Danielson, 2013) sowie das im Rahmen der *Pythagoras-Studie* (Klieme, Pauli & Reusser, 2009) entwickelte Beobachtungsinstrument der drei Basisdimensionen. Als fachspezifische Instrumente werden das *Mathematical Quality of Instruction* (MQI, Hill et al., 2008; Learning Mathematics for Teaching Project, 2011), das *Instructional Quality Assessment* (Matsumura et al., 2008) und das *Mathematics Scan* eingeordnet (M-Scan, Walkowiak et al., 2014). Charalambous und Praetorius (2018) beschreiben darüber hinaus Instrumente mit „hybridem“ Charakter, die sowohl fachspezifische Merkmale von Mathematikunterricht als auch solche Merkmale erfassen, die unabhängig von einem Fach bzw. fächerübergreifend als relevant erachtet werden. Eine solche Konzeptualisierung könne sich ergeben, wenn fachspezifische mit generischen Rahmenmodellen kombiniert oder diese ineinander integriert würden (z.B. Sammons et al., 2016). Beispiele für solche Hybride sind das Projekt *Teaching for Robust Understanding* (TRU Math, Schoenfeld, 2013) oder das *UTeach Observation Protocol* (UTOP, Marder & Walkington, 2014).

2.2 Fachspezifische Merkmale von Unterrichtsqualität

Nach Charalambous und Praetorius (2018) bilden fachspezifische und hybride Instrumente (letztere in ihrem fachbezogenen Teil) Merkmale mit einer besonderen Bedeutung oder Manifestation für den Mathematikunterricht ab. Im Rahmen einer systematischen Literaturrecherche identifizieren Schlesinger und Jentsch (2016) zwei Kategorien solcher Merkmale für den Mathematikunterricht. Auf der einen Seite steht eine Kategorie stofflicher Merkmale, die die im Unterricht behandelten mathematischen Inhalte und Prozesse in den Vordergrund stellt (Beispiele: Welche prozessbezogenen Kompetenzen werden angesprochen? Sind die behandelten Inhalte korrekt dargestellt? Werden Vernetzungen aufgezeigt?). Auf der anderen Seite steht eine unterrichtsbezogene, mathematikdidaktische Kategorie, die eher die didaktisch-methodische Umsetzung dieser Inhalte und Prozesse fokussiert (Beispiele: Werden im Unterricht verschiedene Repräsentationsformen eingesetzt? Liefert die Lehrperson Angebote zur Sinnkonstruktion? Wird „intelligent“ geübt?).

Aufbauend auf der TEDS-M-Studie und Nachfolgestudien operationalisieren Buchholtz, Kaiser und Blömeke (2014) eine solche zweigeteilte Konzeptualisierung in Bezug auf das Professionswissen von Mathematiklehrkräften in Form einer *stoffdidaktischen* und einer *unterrichtsbezogenen* Komponente. Das *stoffbezogene mathematikdidaktische Wissen* umfasst dabei einerseits „stofflich geprägte Fragestellungen des Lehrens von Mathematik“, andererseits eine „fachlich geprägte Diagnostik von Schülerlösungen“ (Buchholtz, Kaiser & Blömeke, 2014, S. 111). Das *unterrichtsbezogene mathematikdidaktische Wissen* umfasst unterrichtsmethodisch ausgerichtete Merkmale. Hierzu zählen u.a. der Umgang mit Heterogenität, pädagogisch motivierte Interventionsmöglichkeiten, Lehr- und Lernformen sowie Unterrichtsarrangements im Mathematikunterricht. Diese zweigeteilte stoffbezogene und unterrichtsbezogene Konzeptualisierung mathematikdidaktischen Wissens wurde in der vorliegenden Studie der Konzeptualisierung von mathematikdidaktischer Unterrichtsqualität zugrunde gelegt.

Die stoffbezogene Dimension mathematikdidaktischer Unterrichtsqualität fokussiert die mathematische, stoffliche Qualität durch eine korrekte und präzise Darstellung mathematischer Inhalte sowie eine angemessene fachliche Tiefe des Unterrichts. Die unterrichtsbezogene Dimension mathematikdidaktischer Unterrichtsqualität nimmt didaktische Unterrichtsentscheidungen in den Blick. Hierzu zählen u.a. die Verwendung verschiedener Repräsentationsformen und angemessener Beispiele, Unterrichtsarrangements wie intelligente Übungsphasen und das Aufzeigen der Relevanz der vermittelten Inhalte, um Schülerinnen und Schüler zu motivieren.

2.3 Zum Zusammenhang fachspezifischer Merkmale von Unterrichtsqualität mit kognitiven Facetten professioneller Kompetenz von Lehrpersonen

Unterrichtsqualität kann anhand von empirischen Wirkungszusammenhängen untersucht und ein bestehendes Instrument hinsichtlich verschiedener Validitätskriterien beurteilt werden. Die Grundlage bildet meist die Konzeptualisierung professionellen Wissens von Lehrpersonen nach Shulman (1986), der professionelles Wissen u.a. in die Kategorien fachliches, fachdidaktisches und pädagogisches Wissen einteilt. Nach diesem Verständnis werden in der COACTIV-Studie (Kunter et al., 2011; Baumert et al., 2010) längsschnittliche Zusammenhänge zwischen professionellen Kompetenzen von Lehrpersonen und Unterrichtsqualität sowie ihre Wirkungen auf die Mathematikleistung von Schülerinnen und Schülern untersucht. Die Autorengruppe findet dabei in einer Mediationsanalyse sowohl auf Individual- als auch auf Klassenebene Effekte mittlerer Größe. Entsprechend nehmen wir mit Baumert et al. (2010) an, dass die mathematikdidaktische Kompetenzfacette von Lehrpersonen Einfluss auf deren unterrichtliches Handeln hat.

Zusammenhänge zwischen mathematischem bzw. mathematikdidaktischem Wissen und Unterrichtsqualität werden auch von Hill, Umland, Litke und Kapitula (2012) berichtet, die in einer amerikanischen Studie mit $n = 34$ Mathematiklehrpersonen hohe Korrelationen zwischen Unterrichtsqualität und mathematischem Fachwissen der Lehrpersonen („Mathematical Knowledge for Teaching“) feststellen. In vertiefenden qualitativen Analysen gelang es zudem,

das sich in den beobachteten Unterrichtsstunden manifestierende Fachwissen der Lehrpersonen zu identifizieren, und inhaltliche Zusammenhänge zwischen Kompetenzfacetten der Lehrpersonen und ihrer Performanz im Unterricht aufzuzeigen. In einer Studie von Kersting, Givvin, Thompson, Santagata und Stigler (2012), in der bei ähnlicher Stichprobengröße dasselbe Instrument zur Erfassung des Fachwissens, jedoch ein anderes Instrument zur Erfassung der Unterrichtsqualität eingesetzt wurde, fanden sich solche Zusammenhänge zwischen dem Fachwissen der Lehrpersonen und der Unterrichtsqualität zunächst nicht. Die Forschergruppe setzte allerdings ein weiteres Instrument zur Erfassung des Fachwissens ein, bei dem die Lehrpersonen Videoaufzeichnungen von Unterrichtssituationen anschauen und dazu Testfragen beantworten sollten. Mit diesem Instrument konnten sowohl hohe Zusammenhänge zur Unterrichtsqualität als auch mittlere Korrelationen zum mathematischen Fachwissen der Lehrpersonen mit dem von der Autorengruppe um Hill verwendeten Instrument nachgewiesen werden.

Wir begreifen professionelle Kompetenz von Lehrpersonen nach Blömeke, Gustafsson und Shavelson (2015) als ein Kontinuum, auf dem sowohl kognitions- als auch performanzähnliche Facetten angeordnet und die über proximale Kompetenzfacetten wie der professionellen Wahrnehmung von Unterricht (*teacher noticing*, Sherin, Jacobs & Philipp, 2011; Van Es & Sherin, 2002) mediiert werden (Kaiser et al., 2015). Die professionelle Unterrichtswahrnehmung wird differenziert in die Subfacetten der Wahrnehmung, Interpretation und Entscheidung zentraler unterrichtlicher Ereignisse (*Perception, Interpretation, Decision-making*, bezeichnet als PID-Modell). Für den vorliegenden Beitrag liefert ein derartiges Kontinuum eine geeignete theoretische und methodologische Rahmung: Die Struktur fachdidaktischen Wissens, verstanden als distale, kognitive Kompetenzfacette einer Lehrperson, sollte sich demgemäß auf das unterrichtliche Handeln von Lehrpersonen übertragen lassen. Damit Unterrichtsinhalte etwa korrekt und präzise dargestellt werden können, der Mathematikunterricht gemäß Brunner (2017) also eine notwendige fachliche Fundierung aufweist, benötigt die unterrichtende Lehrperson ein gewisses Maß an mathematischem und mathematikdidaktischem Wissen. Unter fachspezifischen Merkmalen von Unterrichtsqualität verstehen wir in der vorliegenden Arbeit daher solche Merkmale, deren Sicherstellung notwendigerweise ein gewisses Maß an Fachwissen oder fachdidaktischem Wissen der Lehrperson erfordert. Fachspezifische Merkmale von Unterrichtsqualität sind demnach Merkmale, die besondere Herausforderungen an die Lehrperson und ihr unterrichtliches Handeln stellen (Charalambous & Praetorius, 2018). Unter generischen bzw. fachübergreifenden Merkmalen von Unterrichtsqualität verstehen wir hingegen Aspekte, die unabhängig vom fachlichen bzw. fachdidaktischen Wissen der Lehrperson sind.

3. Fragestellungen und Hypothesen

Anknüpfend an die dargelegten Arbeiten zur Beschreibung von Unterrichtsqualität verfolgt der vorliegende Beitrag zwei Ziele. Zum einen soll ein Beobachtungsinstrument zur Erfassung von Unterrichtsqualität vorgestellt werden, das gängigen psychometrischen Gütekriterien entspricht

und das die Zusammenhänge der beiden fachspezifischen Dimensionen zu den drei Basisdimensionen aufweist. Dabei operationalisieren wir in Anlehnung an die Lehrerprofessionsforschung sowohl eine stoffdidaktische als auch eine unterrichtsdidaktische Qualitätsdimension für den Mathematikunterricht. Hieraus leitet sich die erste Fragestellung ab:

Fragestellung 1: a) Lässt sich Unterrichtsqualität im Mathematikunterricht mit fünf Skalen reliabel erfassen? b) Welche Zusammenhänge weisen die fachspezifischen Dimensionen mit den Basisdimensionen auf?

Für diese Fragestellung wird angenommen, dass die eingesetzten Skalen zur Unterrichtsqualität zwar deutliche Interkorrelationen aufweisen, empirisch aber dennoch voneinander zu trennen sind.

Darüber hinaus verfolgt der Beitrag das Ziel, die beiden neuentwickelten mathematikdidaktischen Skalen zu validieren. Mit der zweiten Fragestellung thematisieren wir daher den Zusammenhang zwischen Unterrichtsqualität und professioneller Kompetenz von Lehrpersonen:

Fragestellung 2: Sind im Mathematikunterricht fachdidaktische Unterrichtsqualität und domänenspezifische Kompetenzfacetten von Lehrpersonen miteinander korreliert?

In Anlehnung an das Kompetenzmodell von Blömeke, Gustafsson und Shavelson (2015) nehmen wir an, dass diese Korrelationen positiv ausgeprägt sind und proximale Kompetenzfacetten stärker als distale Kompetenzfacetten mit stoffdidaktischer bzw. unterrichtsdidaktischer Qualität zusammenhängen.

4. Methodisches Vorgehen

4.1 Studiendesign und Stichprobe

Das im Folgenden beschriebene Beobachtungsinstrument zur Untersuchung von mathematikspezifischer Unterrichtsqualität wurde im Rahmen der Studie TEDS-Unterricht entwickelt und eingesetzt. TEDS-Unterricht ist ein Nachfolgeprojekt der Studien TEDS-M (*Teacher Education and Development Study: Learning to Teach Mathematics*) (Blömeke, Kaiser & Lehmann, 2010) und TEDS-FU (*Follow-Up*) (z.B. Blömeke et al., 2014). Dieses Forschungsprojekt wurde 2014-2016 in Hamburg durchgeführt mit dem Ziel, die professionelle Kompetenz von Mathematiklehrkräften der unteren Sekundarstufe und ihres über die Unterrichtsqualität vermittelten Einflusses auf den Leistungszuwachs von Schülerinnen und Schülern zu untersuchen. Dabei stellte die Konzeptualisierung von Unterrichtsqualität durch fachdidaktische Merkmale des Mathematikunterrichts und die drei Basisdimensionen einen wichtigen Teilaspekt dar. Dabei wurden bei einer Teilstichprobe der befragten Mathematiklehrkräfte in Hamburg Unterrichtsbeobachtungen in-vivo durchgeführt.

Das Beobachtungsinstrument war im Rahmen einer ersten Datenerhebung pilotiert worden. Hierzu wurden im Frühjahr 2015 bei 13 Lehrpersonen Unterrichtsbeobachtungen in drei Bundesländern (Hamburg, Brandenburg und Schleswig-Holstein) durchgeführt. Die Pilotierung führte zu kleineren Modifikationen des Instruments. Die im vorliegenden Beitrag verwendeten Daten der Haupterhebung im Rahmen der Studie TEDS-Unterricht entstammen einer neuen Stichprobe von 37 Hamburger Lehrpersonen der unteren Sekundarstufe. Bei jeder Lehrperson wurden im Abstand von ca. zwei Wochen zwei Unterrichtsbeobachtungen in Doppelstunden (90 Minuten) durchgeführt. Die Teilnahme der Lehrpersonen an Online-Tests und Unterrichtsbeobachtungen erfolgte auf freiwilliger Basis. Der Anteil der weiblichen Lehrpersonen betrug 50%. Im Durchschnitt waren die beobachteten Lehrpersonen 40 Jahre alt (Spannweite: 27 bis 71 Jahre) und verfügten über rund zehn Jahre Unterrichtserfahrung (Spannweite: ein bis 33 Jahre). Das erste Staatsexamen wurde im Durchschnitt mit der Note 1,6 (Spannweite: „sehr gut“ bis „befriedigend“) abgeschlossen, während das Ergebnis des zweiten Staatsexamens im Durchschnitt 1,9 betrug (Spannweite: „sehr gut“ bis „ausreichend“).

4.2 Erhebungsinstrument zur Unterrichtsqualität

Zur Erfassung der Unterrichtsqualität von Mathematikunterricht wurde ein Beobachtungsinstrument eingesetzt, das zum einen die drei Basisdimensionen und zum anderen zwei fachspezifische Dimensionen mathematikdidaktischer Unterrichtsqualität, eine unterrichtbezogene und eine stoffbezogene, zu messen beansprucht (siehe auch Schlesinger et al., 2018). Die Bewertung erfolgte auf einer vierstufigen Skala (1=“trifft überhaupt nicht zu“ bis 4=“trifft vollständig zu“).

Die Items zu den drei Basisdimensionen stammen dabei zum Teil aus dem Beobachtungsinstrument der Pythagoras-Studie (Rakoczy & Pauli, 2006), sie wurden für die vorliegende Studie jedoch adaptiert sowie um weitere Items ergänzt. In Bezug auf die Dimension Klassenführung war der Grund hierfür, dass die Items der Pythagoras-Studie jeweils große Bereiche der Klassenführung enthielten, was eine Bewertung für die Rater erschwerte. Durch die Unterteilung in Anlehnung an Instrumente der Schulinspektion (Pietsch & Tosana, 2008) und damit die Reduzierung der Komplexität erschien eine in-vivo Bewertung im Unterricht schneller umsetzbar. Das Merkmal „Klarheit“ aus der Pythagoras-Studie, das dort unzureichende Reliabilitätswerte (relativer Generalisierbarkeitskoeffizient) aufwies (Rakoczy & Pauli 2006), wurde deshalb mit den zwei Items „Transparenz des Lernprozesses“ und „Strukturierung des Unterrichts“ erhoben. Das Merkmal „Disziplinprobleme / Unterrichtsstörungen“ wurde wiederum mit zwei Items „Störungspräventives Verhalten“ und „Arbeitsatmosphäre“ erhoben, und das Merkmal „Classroom Management“ wurde durch die zwei Items „Effektive Lernzeitnutzung“ und „Sinnvolle Routinen und Regeln“ operationalisiert.

Die zweite Dimension wurde ebenfalls in Teilen verändert. Grund hierfür war, dass eine „in vivo“ Einschätzung der Unterrichtsqualität noch größere Herausforderungen an die Rater stellt als die Bewertung von Unterrichtsvideos, da das mehrfache Betrachten derselben Situation nicht möglich ist, so dass eine schlechte Inter-Rater-Reliabilität voraussichtlich auch mit einer

umfangreichen Schulung nicht kompensiert hätte werden können. Daher wurden keine Items übernommen, die in der Pythagoras-Studie schlechte Reliabilitäten aufwiesen (beispielsweise „Einforderung von Anstrengung“, „Ängstlichkeit“ oder auch „Engagement der Lehrperson“). Darüber hinaus wurden die Merkmale „Inhaltliche Relevanz“ und „Fehlerkultur“ fachspezifisch konzeptualisiert (Steuer, 2014). Die Items „Wahlmöglichkeiten“, „Eigenverantwortung“ und „Individualisierung“ wurden in zwei Items „Differenzierung“ und „selbstgesteuertes Lernen“ unterteilt. Die Items „Rückmeldung der Lehrperson“, „Anerkennung/Wertschätzung der Lehrperson“ und „Kooperation“ wurden übernommen. Das Item „Notendruck“ war zunächst operationalisiert worden, musste jedoch nach der Pilotierung gestrichen werden, weil das Unterrichtsmerkmal nicht beobachtet werden konnte.

Die dritte Dimension kognitive Aktivierung wurde ebenfalls adaptiert. Das Item „Herausfordernde Probleme und Fragen“ wurde übernommen. Die Items „Exploration des Wissens“, „Exploration von Denkweisen“ und „Evolutionärer Umgang“ wurden zu einem Item zusammengefasst, da sich diese Bereiche in der Pilotierung nicht angemessen trennen ließen. Das Item „rezeptives Lernverständnis“ wurde nicht übernommen, da es überwiegend das negativ gepolte Item zu „Herausfordernde Probleme und Fragen“ sowie „Evolutionärer Umgang“ darstellt. Zur breiteren Erfassung des Konstrukts kognitive Aktivierung haben wir schließlich drei Items „kognitiv aktivierende Qualität der Unterrichtsmethoden“, „Förderung von Metakognition“ und „Wissenssicherung“ in Anlehnung an Helmke (2012) und Lipowsky (2015) hinzugenommen.

Die drei Basisdimensionen werden so mit insgesamt 18 hoch-inferent einzuschätzenden Items erfasst (siehe Tabelle 1). Jedes Item, das von den Ratern bewertet wird, wird durch eine Reihe von Indikatoren und typischen Ankerbeispielen beschrieben. Die Indikatoren stellen Hilfestellungen dar, um die Items einschätzen zu können. Zusätzlich zu den Indikatoren wurde ein umfangreiches Kodiermanual entwickelt, das neben der Itembeschreibung auch weitere Bewertungshinweise enthält.

Tabelle 1: Skalen, Items und Indikatoren zur Erfassung von Unterrichtsqualität

Dimension	# Items	Beispielitems	Indikatoren
Klassenführung	6	Effektive Lernzeitnutzung	• Der Unterricht beginnt und endet pünktlich • Die Übergänge zwischen den Phasen erfolgen reibungslos • Die Unterrichtszeit wird für unterrichtsbezogene Inhalte genutzt
Konstruktive Unterstützung	7	Umgang mit Heterogenität	• Es gibt zusätzliches Material für Teile der Lerngruppe • Es gibt anforderungsbezogene, differenzierende Aufgabenstellungen • Es findet ad-hoc Differenzierung statt (z. B. leichtere Fragestellung)

Kognitive Aktivierung	5	Herausfordernde Fragen und Aufgaben	• Die LK zeigt ein angemessenes Frageverhalten (nicht zu kleinschrittig) • Die LK stellt kognitiv herausfordernde Aufgaben • Die LK gibt geeignete Impulse, z.B. ein Problem oder eine offene Frage
Unterrichtsbezogene Merkmale	4	Repräsentationsformen	• Es werden verschiedene Darstellungsformen verwendet • Die LK verdeutlicht Beziehungen zwischen Darstellungsformen
Stoffbezogene Merkmale	5	Fachliche Tiefe	• Es finden Verallgemeinerungen statt • Es finden innermathematische Vernetzungen statt • Mathematische Inhalte werden strukturiert und begrifflich verdichtet

Die fachspezifischen Skalen beinhalten neun hoch-inferent zu bewertende Items (fünf stoffbezogene Items und vier unterrichtsbezogene Items). Die Förderung der prozessbezogenen Kompetenzen im Mathematikunterricht (Blum et al., 2015) wurde mit sechs niedrig-inferent einzuschätzenden Items auf einer vierstufigen Likert-Skala erhoben („nicht gefördert“ bis „umfangreich gefördert“). Diese wurden für weitere Analysen zu einem Item-Paket „Kompetenzorientierung“ aggregiert, da es sich weitgehend um gegenseitig ausschließende Prozesse handelte und so eine Häufung fehlender Werte zu vermeiden.

4.3 Analyseinheit und Training der Beobachter(innen)

Insgesamt standen für die Unterrichtsbeobachtungen sechs Rater zur Verfügung, aus denen für jede Unterrichtsbeobachtung eine Zufallsauswahl von zwei Ratern getroffen wurde. Die Unterrichtsstunden im Rahmen der Studie wurden „in vivo“ bewertet, es fanden aus rechtlichen und praktischen Gründen keine Videoaufzeichnungen statt. Alle Rater hatten mindestens einen Bachelorabschluss für das Lehramt Mathematik inne, überwiegend sogar ein abgeschlossenes Masterstudium. Die Rater wurden im Vorfeld umfangreich geschult, indem Items und Indikatoren mit Hilfe eines Kodiermanuals ausführlich besprochen und anhand von Unterrichtsvideos erprobt wurden. Insbesondere wurde auch das Rating im Unterricht mehrfach durchgeführt. Dieses Training umfasste etwa 30 Stunden in einem Zeitraum von ca. vier Wochen während der vorlesungsfreien Zeit.

Innerhalb einer Doppelstunde (90 Minuten) wurde die Unterrichtsqualität nach vier annähernd gleichen Zeitintervallen von beiden Ratern unabhängig voneinander eingeschätzt. Dies bedeutet, dass die Bewertung in der Regel nach 22.5 Minuten erfolgte. Mit dem mehrfachen Rating sollte die Entwicklung der Unterrichtsqualität innerhalb einer Doppelstunde erfasst und andererseits das Rating für die Rater erleichtert werden. Jede Lehrperson wurde in zwei Doppelstunden beobachtet, so dass jeweils acht Messzeitpunkte bewertet wurden. Neben den Ratings zur Unterrichtsqualität wurden zusätzlich Unterrichtsprotokolle zu Sozialformen, Lehrer-Schüler-

Interaktionen und fachlichen Inhalten erstellt. Nach jeder Doppelstunde fand eine ausführliche Nachbesprechung zwischen den beiden Ratern statt, bei der die Stunde insbesondere im Hinblick auf mögliche systematische Fehleinschätzungen reflektiert wurde. Im Rahmen dieser Nachbesprechung konnten einzelne Ratings korrigiert werden (siehe zu entsprechendem Vorgehen z.B. König, 2015). Ziel war es dabei nicht, eine einheitliche Bewertung zu erreichen, sondern Wahrnehmungs- und Bewertungsfehler aufgrund des Zeitdrucks im Unterricht aufzudecken. Dies war notwendig, weil durch das Fehlen von Videoaufzeichnungen keine „objektive“ Klärung der Tatsachen im Anschluss an die Unterrichtsbeobachtungen möglich war.

4.4 Datenauswertung

Die Analyse der Rohwerteverteilung führte zum Ausschluss zweier Items auf Grund schlechter Verteilungseigenschaften („Schülerfeedback“ und „Förderung von Metakognition“, Häufigkeit des Modalwerts jeweils ca. 90%). Zur Beurteilung der Interraterreliabilität wurde der Intraklassenkorrelationskoeffizient bestimmt, da Ratingskalen üblicherweise als intervallskaliert angenommen werden (Wirtz & Caspar, 2002). Die Interraterreliabilität lag vor der gemeinsamen Nachbesprechung der Rater bei $.46 < ICC \leq .96$. Die Nachbesprechung der Rater führte insgesamt zu einer Steigerung der Interraterreliabilität. Dabei lag lediglich für das Item „Routinen und Regeln“ eine unbefriedigende Interraterreliabilität vor, so dass dieses von weiteren Analysen ausgeschlossen wurde. Nachfolgend wurden auf Grund der höheren Reliabilität jeweils die Ratings nach der Besprechung verwendet (Klassenführung: $.84 < ICC \leq .88$; Konstruktive Unterstützung: $.80 < ICC \leq .96$; Kognitive Aktivierung: $.80 < ICC \leq .93$; Stoffbezogene Qualität: $.82 < ICC \leq .95$; Unterrichtsbezogene Qualität: $.80 < ICC \leq .94$). Für die weiteren Analysen mit $n = 37$ Lehrpersonen wurden zunächst die Urteile der beiden Rater zu einem einzelnen Rating zusammengefasst und anschließend das arithmetische Mittel über acht Messzeitpunkte gebildet. Dieses Vorgehen kann als zusätzliche Maßnahme zur Qualitätsverbesserung der Einschätzungen interpretiert werden (Pietsch & Tosana, 2008).

Zur Testung der Eindimensionalität der fünf Skalen wurden fünf separate konfirmatorische Faktorenanalysen mit Hilfe der Software Mplus 6.1 (Muthén & Muthén, 2010) durchgeführt. Zur Evaluation der Modellpassung wurden die χ^2 -Teststatistik und der RMSEA mit zugehörigem 90%-Konfidenzintervall herangezogen. Wir berichten außerdem die Intraklassenkorrelation, d.h. das Verhältnis der Varianz zwischen Lehrpersonen (*between*) und der Gesamtvarianz in den Daten, um ein Maß für die Retest-Reliabilität zu erhalten. Hierzu wurde eine Mehrebenenanalyse durchgeführt, in die die 37 Lehrpersonen als Cluster mit je zwei Unterrichtsstunden eingingen.

Für die Zusammenhangsanalysen wurden manifeste Korrelationen geschätzt und auf Signifikanz geprüft ($p < .10$). Für die berichteten Zusammenhänge zu den Kompetenzfacetten der Lehrpersonen wurden wegen der geringen Stichprobengröße außerdem Bias-korrigierte (BCa) Bootstrap-Konfidenzintervalle mit $n = 5000$ Wiederholungen geschätzt.

4.5 Kognitive Kompetenzfacetten

Das mathematische Fachwissen (MCK) wurde mittels eines digitalisierten Papier-und-Bleistift-Tests erfasst. Dabei handelte es sich um ein verkürztes Instrument aus der internationalen Lehrerbildungsstudie TEDS-M (Blömeke, Kaiser, Lehmann, 2010), das ursprünglich für Mathematiklehrkräfte am Ende der Lehrerausbildung entwickelt worden war. Die verkürzte Fassung wurde in TEDS-FU entwickelt und mit denselben Probanden nach vier Jahren Schulpraxis erneut eingesetzt (Blömeke et al., 2014). Der Test bestand aus 26 Items mit den Themenbereichen Arithmetik, Algebra, Geometrie und Stochastik. Das mathematikdidaktische Wissen wurde auf ähnliche Weise erhoben. Hier bestand der Test aus 29 Items aus den Bereichen Stoff- und Unterrichtsdidaktik, wobei beide Testteile einen etwa gleich großen Anteil ausmachten und der Test gegenüber dem Ursprungstest leicht ausgeweitet war. Die Tests zum MCK und MPCK besaßen mit wenigen Ausnahmen ein Multiple-Choice-Antwortformat. Neben den genannten Konstrukten wurden die Fähigkeiten der Lehrpersonen im Wahrnehmen, Interpretieren und Entscheiden in Bezug auf Mathematikunterricht (M_PID) erfasst (Kaiser et al., 2015). Dazu wurden den Probanden drei kurze Videos von Unterrichtssituationen gezeigt, zu denen insgesamt 31 Items beantwortet werden sollten (Blömeke et al., 2016). Der Test bestand ungefähr jeweils zur Hälfte aus offenen Aufgaben und solchen im Multiple-Choice-Format. Die richtigen Antworten wurden zuvor durch ein Expertenrating festgelegt, das auch zur Bestimmung und Diskussion der Inhaltsvalidität diente. Alle Testteile wurden anschließend nach dem 1-PL-Modell skaliert, wobei offene Aufgaben vorher kodiert wurden, und zeigten bei der Datenanalyse eine befriedigende Reliabilität ($.66 < WLE \leq .80$). Dabei wurden übersprungene bzw. „not reached“ Items als falsch gewertet. Die manifesten Interkorrelationen der Testteile können als hoch eingeschätzt werden ($.61 < r \leq .79$).

5. Ergebnisse

Die eindimensionalen konfirmatorischen Faktorenanalysen zeigen für drei von fünf Skalen einen sehr guten Modellfit ($\chi^2 < df$). Für die Skalen zur Klassenführung ($\chi^2 = 4.93$, $df = 4$, $p = .29$, $RMSEA = .08$) und zur konstruktiven Unterstützung fällt dieser akzeptabel aus ($\chi^2 = 9.12$, $df = 8$, $p = .33$, $RMSEA = .06$). Hierzu wurde jeweils eine Residualkorrelation zugelassen, die inhaltlich zu rechtfertigen war. Die 90%-Konfidenzintervalle für den RMSEA unterscheiden sich dabei nur in ihrer oberen Grenze und bewegen sich zwischen $[.00, .13]$ für die Dimensionen stoffdidaktische Qualität und $[.00, .27]$ für Klassenführung. Die Skalenreliabilitäten können in drei von fünf Fällen als gut beurteilt werden, wie in Tabelle 2 dargestellt wird. Für die Dimensionen konstruktive Unterstützung und unterrichtsdidaktische Qualität unterschreitet der α -Wert die in der Literatur (z.B. Bortz & Döring, 2006) empfohlene Grenze von .80 allerdings.

Der Mittelwert der Skala Klassenführung fällt hoch aus ($M = 3.16$), während die Skala zur konstruktiven Unterstützung eher eine niedrige Ausprägung aufweist ($M = 1.96$). Die Mittelwerte der drei anderen Skalen zur Unterrichtsqualität rangieren um den Skalenmittelpunkt.

Die Streuung der fünf Skalen fällt insgesamt niedrig aus ($SD \leq .38$). Die als Maß für die Retest-Reliabilität angeführte Intraklassenkorrelation zeigt deutliche intraindividuelle Unterschiede an. So wird der Anteil der Varianz zwischen zwei Unterrichtsstunden derselben Lehrperson für konstruktive Unterstützung auf 42% bis 59% geschätzt, was als moderate Retest-Reliabilität interpretiert werden kann. Die Werte für die anderen Skalen fallen ähnlich aus. Damit lässt sich Fragestellung 1a insgesamt stützen, d.h. Unterrichtsqualität im Mathematikunterricht kann mit den fünf beschriebenen Skalen trotz einzelner Schwächen psychometrisch weitgehend überzeugend erfasst werden.

Tabelle 2: Deskriptive Kennwerte für die fünf Skalen zur Unterrichtsqualität für $n = 37$ Lehrpersonen. M : arithmetisches Mittel, SD : Standardabweichung, α : Cronbachs Alpha, ICC : Intraklassenkorrelation.

Skala	M	SD	α	Retest-Reliabilität
Klassenführung	3.16	.33	.86	$.51 \leq ICC < .59$
Konstruktive Unterstützung	1.96	.23	.73	$.42 \leq ICC < .59$
Kognitive Aktivierung	2.61	.35	.83	$.50 \leq ICC < .62$
Unterrichtsdidaktische Qualität	2.46	.38	.69	$.52 \leq ICC < .62$
Stoffdidaktische Qualität	2.66	.31	.77	$.41 \leq ICC < .60$

Tabelle 3 gibt die manifesten Interkorrelationen zwischen den Skalen zur Unterrichtsqualität an. Diese weisen bis auf zwei Fälle mindestens mittlere Effektstärke auf ($r \geq .34$). Die beiden Ausnahmen sind in Bezug auf Fragestellung 1b) relevant. Auch wenn die etwas geringere Skalenreliabilität in Betracht gezogen werden muss, fällt doch auf, dass die Skala zur unterrichtsbezogenen mathematikdidaktischen Qualität nicht signifikant mit den beiden Basisdimensionen Klassenführung und konstruktive Unterstützung korreliert. Ein hoher Zusammenhang ergibt sich dagegen zwischen den beiden mathematikdidaktischen Skalen und der Skala zur kognitiven Aktivierung ($r = .66$ bzw. $.69$).

Fragestellung 2 zielt auf eine Validierung der beiden neuentwickelten Skalen und betrifft die Zusammenhänge zwischen mathematikdidaktischer Unterrichtsqualität und domänenspezifischen Kompetenzfacetten. Weder für die unterrichtsbezogene noch für die stoffdidaktische Skala lässt sich ein signifikanter Zusammenhang zum mathematischen Professionswissen nachweisen. Dagegen zeigt sich tendenziell eine Korrelation zwischen der stoffdidaktischen Skala und dem mathematikdidaktischen Wissen. Das Bootstrap-Konfidenzintervall für den Korrelationskoeffizienten beträgt $[-.19, .60]$. Die Skala zur unterrichtsdidaktischen Qualität korreliert in der Tendenz ebenfalls positiv ($.07 < p \leq .09$) mit der Kompetenzfacette M_PID. Das Bootstrap-Konfidenzintervall für den Korrelationskoeffizient

beträgt [.06, .49]. Die stoffdidaktische Qualitätsdimension weist ebenfalls eine positive Korrelation mit der Kompetenzfacette M_PID auf. Das Bootstrap-Konfidenzintervall für den Korrelationskoeffizienten beträgt [.04, .50]. Obwohl die absoluten Werte tendenziell in die erwartete Richtung weisen, unterscheiden sich diese Korrelationen aber nicht signifikant ($z = -.77$, $p = .22$), damit kann entgegen unserer Erwartung kein systematischer Unterschied zwischen den Zusammenhängen zu proximalen und distalen Kompetenzfacetten nachgewiesen werden.

Tabelle 3: Manifeste Korrelationen aller eingesetzten Skalen für $n = 37$ Lehrpersonen.

Skala	1.	2.	3.	4.	5.
1. Klassenführung	1				
2. Konstruktive Unterstützung	.34*	1			
3. Kognitive Aktivierung	.54**	.36*	1		
4. Unterrichtsdidaktische Qualität	.19	.28	.69**	1	
5. Stoffdidaktische Qualität	.40*	.57**	.66**	.44**	1
MCK	.02	-.08	.06	.15	.26
MPCK	-.02	-.01	.12	.16	.28(*)
M_PID	-.01	.01	.28(*)	.30(*)	.40*

(*) $p < .10$, * $p < .05$, ** $p < .01$. Durch Rahmung hervorgehoben sind jeweils die Korrelationen, die zur Beantwortung von Fragestellung 1b) und 2 herangezogen wurden.

6. Diskussion

Im vorliegenden Beitrag werden Ergebnisse einer Studie berichtet, in der Zusammenhänge zwischen drei Basisdimensionen und zwei fachspezifischen Dimensionen von Unterrichtsqualität analysiert wurden. Zudem wurden letztere anhand von Zusammenhängen zu domänenspezifischen Kompetenzfacetten von $n = 37$ Mathematiklehrpersonen validiert. Zur Erfassung der Unterrichtsqualität wurde ein fünfdimensionales Beobachtungsinstrument mit hoch-inferent einzuschätzenden Items vorgestellt, das die genannten fünf Dimensionen der Unterrichtsqualität von Mathematikunterricht operationalisiert. Dies geschah unter Rückgriff auf aktuelle Theorien zur Lehrerprofessionsforschung (Buchholtz et al., 2014), nach denen zwischen zwei fachdidaktischen Qualitätsdimensionen unterschieden werden kann (stoffbezogen und unterrichtsbezogen).

Die Ergebnisse weisen auf die Eindimensionalität und eine gute Interraterreliabilität der fünf eingesetzten Skalen zur Unterrichtsqualität hin. Retest-Reliabilität und interne Konsistenz fallen

moderat bis zufriedenstellend aus. Es zeigte sich außerdem, dass die erfassten Konstrukte empirisch trennbar sind. Hohe Korrelationen zeigen sich dabei insbesondere zwischen den beiden fachspezifischen Dimensionen und der kognitiven Aktivierung, was im Einklang mit bisherigen Vermutungen steht, dass kognitive Aktivierung die größte Nähe zu fachspezifischen Merkmalen von Unterrichtsqualität aufweist. Die geringe Korrelation zwischen den mathematikdidaktischen unterrichtsbezogenen Merkmalen und der konstruktiven Unterstützung ist zunächst verwunderlich, zeigt jedoch, dass die erfolgreiche Implementation solcher fachspezifischen Aspekte im Unterricht nicht notwendigerweise von einer hohen konstruktiven Unterstützung abhängen. Dies verdeutlicht, dass analog zum mathematikdidaktischen Wissen als eigenständige Domäne (Shulman, 1985, 1986) auch die fachspezifischen unterrichtsbezogenen Merkmale eine eigenständige Dimension darstellen, die nicht durch generische Qualitätsmerkmale wie Klassenführung oder konstruktive Unterstützung abgedeckt werden. Bei der Messung des unterrichtsbezogenen mathematikdidaktischen Wissens und des pädagogischen Wissens konnten Buchholtz, Kaiser und Blömeke (2014) ebenfalls nur einen sehr schwachen Zusammenhang ($.18, p < .01$) feststellen, was sie unter anderem damit erklärten, dass „sich hinter dem unterrichtsbezogenen mathematikdidaktischen Wissen auch kein in den mathematikunterrichtlichen Kontext eingekleidetes pädagogisches Wissen verbirgt, sondern spezifisch unterrichtsbezogene mathematikdidaktische Inhaltsbereiche“ (S. 121). Als erste Hinweise auf die kriteriale Validität des Beobachtungsinstruments können die Zusammenhänge mittlerer Effektstärke zwischen fachdidaktischen Qualitätsdimensionen und der Fähigkeit, Unterricht unter mathematikdidaktischer Perspektive wahrzunehmen (Skala M_PID), verstanden werden. Darüber hinaus zeigen sich schwache Korrelationen zwischen dem mathematikdidaktischen Wissen der Lehrpersonen und der stoffdidaktischen Qualitätsdimension. Tendenziell deutet sich auch die angenommene größere Nähe der unterrichtsbezogenen Wahrnehmung zur Unterrichtsqualität im Vergleich zum Professionswissen an. Die Unterschiede werden allerdings nicht signifikant.

Die geringe Stichprobengröße in der dargestellten Studie bringt insofern eine Reihe von Grenzen mit sich. Einerseits führt sie zu einer niedrigen Teststärke und damit zur einer höheren Wahrscheinlichkeit für einen Fehler zweiter Ordnung, insbesondere bei den berichteten konfirmatorischen Faktorenanalysen. Die geringe Stichprobengröße wirkt sich zudem auf die Signifikanz der Korrelationskoeffizienten aus. Schwache Zusammenhänge können hier nur schwerlich aufgezeigt werden, weshalb die vorliegenden Ergebnisse mit Daten einer größeren Stichprobe repliziert werden sollten. Diese Stichprobe sollte zudem randomisiert gezogen werden, weil die vorliegenden Ergebnisse recht hohe intraindividuelle, aber geringe interindividuelle Unterschiede in der Unterrichtsqualität zeigen. Dies deutet vor dem Hintergrund der ausgesprochen guten Noten in den Staatsexamina der Lehrpersonen auf eine positive Selektion hin, was allerdings bei Gelegenheitsstichproben als Problem bekannt ist. Auf der anderen Seite ist nach Sichtung der einschlägigen Literatur aber nicht eindeutig, inwiefern bei Unterrichtsqualität tatsächlich von intraindividuellem Stabilität über einen gewissen Zeitraum ausgegangen werden kann (Praetorius et al., 2014). Neuere Arbeiten befassen sich daher insbesondere mit dem Versuch, "wahre" intraindividuelle Variabilität zu beschreiben (Meyer,

Cash & Mashburn, 2011).

Mit der vorliegenden Studie wurde ein Vorschlag zur Konzeptualisierung und zur Messung von Unterrichtsqualität im Mathematikunterricht unterbreitet. Diese Konzeptualisierung greift dabei auf den Ansatz der drei Basisdimensionen zurück und führt diesen weiter, in dem angenommen wird, dass im Mathematikunterricht weitere Merkmale als relevant betrachtet werden sollten. Diese Annahme könnte zukünftig durch hypothesentestende Arbeiten weitergeführt werden. Auch könnten faktorenanalytische Studien einen Beitrag zur Theorienentwicklung für den Mathematikunterricht leisten. Aktuelle fachspezifische Konzeptualisierungen von Unterrichtsqualität schlagen eine Hierarchisierung der Qualitätsdimensionen von Mathematikunterricht vor, in der effiziente Klassenführung und fachliche Fundierung als Voraussetzung für wirkungsvollen Unterricht modelliert werden (Brunner, 2017). Offen bleibt zum gegenwärtigen Zeitpunkt, inwieweit die prognostische Validität des Beobachtungsinstrumentes in Bezug auf Schülerleistungen nachweisbar ist und ob Unterrichtsqualität mit der vorliegenden Konzeptualisierung wie angenommen als Mediator zwischen professioneller Kompetenz von Lehrpersonen und Schülerleistungen modelliert werden kann. Hierzu werden die weiteren Ergebnisse von TEDS-Unterricht und TEDS-Validierung einen Beitrag leisten.

Literaturverzeichnis

- Aebli, H. (1985). Das operative Prinzip. *Mathematik lehren*, 11, 4-6.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (2010). Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress. *American Educational Research Journal*, 47(1), 133–180.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond Dichotomies. *Zeitschrift für Psychologie*, 223(1), 3–13.
- Blömeke, S., König, J., Busse, A., Suhl, U., Benthien, J., Döhrmann, M., & Kaiser, G. (2014). Von der Lehrerausbildung in den Beruf: Fachbezogenes Wissen als Voraussetzung einer genauen Wahrnehmung und Analyse von Unterricht. *Zeitschrift für Erziehungswissenschaft* 17(3), 509-542. doi:10.1007/s11618-014-0564-8
- Blömeke, S., Kaiser, G., & Lehmann, R. (Eds.) (2010). *TEDS-M 2008 - Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster: Waxmann.
- Blömeke, S., Busse, A., Kaiser, G., König, J., & Suhl, U. (2016). The relation between content-specific and general teacher knowledge and skills. *Teaching and Teacher Education*, 56, 35-46.
- Blum, W., Vogel, S., Drüke-Noe, C., & Roppelt, A. (Hrsg.) (2015). *Bildungsstandards aktuell: Mathematik in der Sekundarstufe II*. Braunschweig: Schroedel.
- Blum, W. (2015). Quality teaching of mathematical modelling: what do we know, what can we do? In S. J. Cho (Hrsg.), *The Proceedings of the 12th International Congress on Mathematical Education* (S. 73-96). Cham: Springer.

- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Berlin, Heidelberg: Springer.
- Brunner, E. (2017). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *Journal für Mathematikdidaktik*.
- Buchholtz, N, Kaiser, G., & Blömeke, S. (2014). Die Erhebung mathematikdidaktischen Wissens – Konzeptualisierung einer komplexen Domäne. *Journal für Mathematik-Didaktik*, 35(1), 101-128.
- Charalambous, C., & Praetorius, A.-K. (2018). Studying Instructional Quality in Mathematics through Different Lenses: In Search of Common Ground. *ZDM Mathematics Education*.
- Danielson, C. (2013). *Enhancing professional practice: A framework for teaching (2nd ed.)*. Alexandria: Association for Supervision and Curriculum Development.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior. Perspectives in social psychology*. New York: Plenum.
- Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.
- Hiebert, J., Gallimore, R., Garnier, H., ... Stigler, J. (2003). *Teaching Mathematics in Seven Countries. Results from the TIMSS 1999 Video Study*. Washington: National Center for Education Statistics.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4), 430-511.
- Hill, H. C., Umland, K., Litke, E., & Kapitula, L. R. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education*, 118(4), 489-519.
- Jordan, A., Krauss, S., Löwen, K., Blum, W., Neubrand, M., Brunner, M. et al. (2008). Aufgaben im COACTIV-Projekt: Zeugnisse des kognitiven Aktivierungspotentials im deutschen Mathematikunterrichts. *Journal für Mathematikdidaktik*, 29(2), 83-107.
- Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring Usable Knowledge: Teachers' Analyses of Mathematics Classroom Videos Predict Teaching Quality and Student Learning. *American Educational Research Journal*, 49(3), 568-589.
- Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht: Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagoras". In M. Prenzel & L. Allolio-Näcke (Eds.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (pp. 127-146). Münster: Waxmann.
- Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54, 222–237.
- Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study. In T. Janik & T. Seidel (Eds.), *The Power of Video Studies in Investigating Teaching and Learning in the Classroom* (pp. 137–160). Münster: Waxmann.

- KMK [Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland] (2003). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss*. Beschluss vom 4.12.2003.
http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2003/2003_12_04-Bildungsstandards-Mathe-Mittleren-SA.pdf. Zuletzt aufgerufen am 09. Juni 2016.
- König, J. (2015). Measuring classroom management expertise (CME) of teachers: A video based assessment approach and statistical results. *Cogent Education*, 2(1), 991178.
- Kounin, J. S. (1970). *Disciplin and group management in classrooms*. New York: Holt, Rinehart & Winston.
- Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S. & Neubrand, M. (Hrsg.) (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.
- Kunter, M., & Ewald, S. (2016). Bedingungen und Effekte von Unterricht: Aktuelle Forschungsperspektiven aus der pädagogischen Psychologie. In N. McElvany, W. Bos, H.G. Holtappels, M. M. Gebauer & F. Schwabe (Hrsg.), *Bedingungen und Effekte guten Unterrichts* (S. 9–31). Münster: Waxmann.
- Learning Mathematics for Teaching Project (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, (14), 25–47.
- Lipowsky, F. (2015). Unterricht. In E. Wild, & J. Möller (Eds.), *Pädagogische Psychologie* (2. Auflage) (S. 69-105). Berlin, Heidelberg: Springer.
- Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, 19(6), 527–537.
- Lotz, M. (2015). *Kognitive Aktivierung im Leseunterricht der Grundschule. Eine Videostudie zur Gestaltung und Qualität von Leseübungen im ersten Schuljahr*. Wiesbaden: Springer VS.
- Marder, M., & Walkington, C. (2014). Classroom Observation and Value-Added Models Give Complementary Information about Quality of Mathematics Teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measuring Effective Teaching project* (pp. 234–277). New York: John Wiley & Sons.
- Maier, U., Kleinknecht, M., Metz, K., & Bohl, T. (2010). Ein allgemeindidaktisches Kategoriensystem zur Analyse des kognitiven Potenzials von Aufgaben. *Beiträge zur Lehrerbildung*, 28(1), 84-96.
- Matsumura, L. C., Garnier, H. E., Slater, S. C. & Boston, M. D. (2008). Toward Measuring Instructional Interactions “At-Scale”. *Educational Assessment*, 13(4), 267-300.
- Mayer, R. E. (2004). Should There Be a Three-Strikes Rule Against Pure Discovery Learning? The Case for Guided Methods of Instruction. *American Psychologist*, 59(1), 14–19.
- Meyer, J. P., Cash, A. H., Mashburn, A. (2011). Occasions and the Reliability of Classroom Observations: Alternative Conceptualizations and Methods of Analysis. *Educational Assessment*, 16, 227-243.
- Muthén, B. O. & Muthén, L. K. (2010). *Mplus user's guide* (6th edition). Los Angeles, CA: Muthén & Muthén.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119.

- Pietsch, M., & Tosana, S. (2008). Beurteilereffekte bei der Messung von Unterrichtsqualität. Das Multifacetten-Rasch-Modell und die Generalisierbarkeitstheorie als Methoden der Qualitätssicherung in der externen Evaluation von Schulen. *Zeitschrift für Erziehungswissenschaft*, 11(3), 430-452.
- Praetorius, A.-K. (2014). *Messung von Unterrichtsqualität durch Ratings*. Münster: Waxmann.
- Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12.
- Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic Dimensions of Teaching Quality: The German Framework of Three Basic Dimensions. *ZDM Mathematics Education*.
- Rakoczy, K., & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse. In E. Klieme, C. Pauli & K. Reusser (Eds.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis"* (Teil 3: Hugener, Isabelle; Pauli, Christine & Reusser, Kurt: Videoanalysen, pp. 189-205). Frankfurt am Main: GPPF.
- Reinmann-Rothmeier, G., & Mandl, H. (2006). Unterrichten und Lernumgebungen gestalten. In A. Krapp & B. Weidenmann (Eds.), *Pädagogische Psychologie. Ein Lehrbuch* (pp. 613-658). Weinheim: Beltz.
- Sammons, P., Lindorff, A. M., Ortega, L. & Kington, A. (2016). Inspiring teaching: Learning from exemplary practitioners. *Journal of Professional Capital and Community*, 1(2), 124–144.
- Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM Mathematics Education*, 48(1), 29-40.
- Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *ZDM Mathematics Education*.
- Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM Mathematics Education*, 45(4), 607-621.
- Sherin, M. G., Jacobs, V. R., & Philipp, R. A. (Hrsg.) (2011). *Mathematics teacher noticing: Seeing through teachers' eyes*. New York: Routledge.
- Shuell, T. J. (1993). Toward an integrated theory of teaching and learning. *Educational Psychologist*, 28, 291-311.
- Shulman, L. S. (1985). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching*. 3. Auflage (pp. 3-36). New York: Macmillan.
- Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15, 4-14.
- Steuer, G. (2014). *Fehlerklima in der Klasse. Zum Umgang mit Fehlern im Mathematikunterricht*. Wiesbaden: Springer VS.
- Van Es, E. A., & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education*, 10(4), 571-596.
- Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices:

Evidence of validity and score reliability. *Educational Studies in Mathematics*, 85(1), 109–128.

Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

8 Publikation III

Die dritte Publikation dieser publikationsbasierten Dissertation mit dem Titel „*Subject-Specific Characteristics of Instructional Quality in Mathematics Education*“ ist 2018 ebenfalls in der Zeitschrift *ZDM Mathematics Education* erschienen. Sie bildet einen Teil des Themenheftes „*Studying Instructional Quality in Mathematics through Different Lenses: In Search of Common Ground*“, das von Charalambos Y. Charalambous und Anna-Katharina Praetorius herausgegeben wird, mit dem Ziel, Gemeinsamkeiten und Unterschiede bisheriger Beobachtungsinstrumente anhand der Bewertung dreier identischer Unterrichtsvideos herauszuarbeiten. In der Publikation wird dabei die Fragestellung untersucht, welche Stärken und Schwächen das vorliegende Beobachtungsinstrument bei der Analyse der drei Unterrichtsvideos zeigt.

Hierfür wird zunächst detailliert auf die Konzeptualisierung und Operationalisierung unseres Beobachtungsinstruments eingegangen und es werden bisherige empirische Ergebnisse des Instruments überblicksartig dargelegt. Anschließend wird das vorliegende Beobachtungsinstrument für die Analyse dreier Unterrichtsvideos aus der NCTE video library der Harvard University verwendet und die Ratings werden jeweils detailliert beschrieben und begründet. Ziel dieser Vorgehensweise ist es einerseits, aufzuzeigen, wie das Beobachtungsinstrument in der Praxis funktioniert, und andererseits Stärken und Schwächen des Instruments herauszuarbeiten. Aus diesem Grund wurde ein Mixed-Methods-Design verwendet, das die quantitativen Ratings unseres Beobachtungsinstruments mit Methoden der qualitativen Inhaltsanalyse verbindet. Die Konzeption des ZDM-Themenheftes ermöglicht es darüber hinaus, die Ratings des Beobachtungsinstruments mit anderen etablierten Instrumenten sowohl auf fachübergreifender als auch auf fachspezifischer Ebene zu vergleichen. Auf diesen Vergleich wird in Kapitel 9.3 genauer eingegangen.

Die Ergebnisse der vorliegenden Publikation zeigen, dass das Beobachtungsinstrument zwischen den drei videographierten Unterrichtsstunden differenziert, insbesondere im Hinblick auf die zwei fachspezifischen Dimensionen. Darüber hinaus erlauben die quantitativen Ratings und Ergebnisse der qualitativen Inhaltsanalyse eine detaillierte Betrachtung der Stärken und Schwächen des Beobachtungsinstruments, die zum Teil daraus resultieren, dass das Instrument im Kontext der deutschen Bildungsstandards entwickelt wurde, die drei videographierten Unterrichtsstunden jedoch aus den USA kamen.

8.1 Darlegung des eigenen Anteils an der Publikation

Die vorliegende Publikation ist wiederum in enger Zusammenarbeit mit meinem Kollegen Armin Jentsch von der Universität Hamburg entstanden. Analog zu den zwei anderen Publikationen lag mein Schwerpunkt der vorliegenden Publikation bei der Darlegung der Konzeptualisierung und Operationalisierung des Instruments, während Armin Jentsch sich intensiver mit der Datenanalyse befasst hat. Die quantitativen Ratings sowie die qualita-

tiven Analysen der drei videographierten Unterrichtsstunden erfolgten sowohl von Armin Jentsch als auch von mir. Die Analyse und Beschreibung der Stärken und Schwächen des Instruments in Abschnitt 7 wurden von Armin Jentsch und mir zusammen verfasst. Im fortlaufenden Begutachtungsprozess wurde die vorliegende Publikation substanziell von Armin Jentsch und mir unter Mitarbeit der Projektleitung von TEDS-Unterricht (Gabriele Kaiser, Johannes König und Sigrid Blömeke) überarbeitet. Die Einreichung und Betreuung im Editorial Manager der Zeitschrift *ZDM Mathematics Education* erfolgte wiederum durch mich.

8.2 Abdruck der Publikation III

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *ZDM Mathematics Education*, 50(3). Doi: 10.1007/s11858-018-0917-5

Reprinted by permission from Springer Customer Service Centre GmbH: Springer Nature, ZDM Mathematics Education, Schlesinger et al., Subject-specific characteristics of instructional quality in mathematics education, Copyright © 2018, advance online publication, 31 January 2018 (doi: 10.1007/ s11858-018-0917-5).



Subject-specific characteristics of instructional quality in mathematics education

Lena Schlesinger¹ · Armin Jentsch¹ · Gabriele Kaiser^{1,2} · Johannes König³ · Sigrid Blömeke⁴

Accepted: 22 January 2018
© FIZ Karlsruhe 2018

Abstract

Instructional research in German-speaking countries has conceptualized teaching quality recently according to three generic dimensions, namely, classroom management, student support and cognitive activation. However, as these dimensions are mainly regarded as generic, subject-specific aspects of mathematics instruction, e.g., the mathematical depth of argumentation or the adequacy of concept introductions, are not covered in depth. Therefore, a new instrument for the analysis of instructional quality was developed, which extended this three-dimensional framework by relevant subject-specific aspects of instructional quality. In this paper, the newly developed observational protocol is applied to three videotaped mathematics lessons from the NCTE video library of Harvard University to explore strengths and weaknesses of this instrument, and to examine in more detail how the instrument works in practice. Therefore, we used a mixed-methods design to extend the quantitative observer ratings, which enable high inference, by methods from qualitative content analysis. The results suggest the conclusion that the framework differentiates well between the lessons under a subject-specific perspective.

Keywords Instructional quality · Subject-specific aspects · High-inference ratings · Standardized classroom observations · Mathematics education

1 Introduction

The question of what good instruction is, has been intensively discussed in different fields of research for decades. Various studies could provide evidence that teachers' performance in the classroom has great impact on student achievement (Hill et al. 2005; Kersting et al. 2012; Lipowsky et al. 2009; Seidel and Shavelson 2007). At the same time, a more content-related issue arose, namely, whether and how instructional quality has to be conceptualized separately for different subjects (e.g., Klieme and Rakoczy 2008; Pianta

and Hamre 2009). In recent years, several theoretical frameworks and instruments have been developed not merely for mathematics instruction to assess different aspects of instructional quality, some of which may be called *generic* and others *domain- or subject-specific*, i.e., they are distinguished by mainly generic or subject-specific aspects of instructional quality. Therefore, they differ strongly in what they focus on when assessing instructional quality. Within recent years, the claim for the benefits of a combination of generic and subject-specific assessments has been put forward (Charalambous and Praetorius 2018).

Based on a framework widely accepted especially in German-speaking countries, namely the generic framework of three basic dimensions, we present the observational instrument developed within the study TEDS-Instruct in Germany, in which we extended the generic dimensions of instructional quality by additional subject-specific dimensions. As our previous studies and analyses mainly focused on psychometric issues such as rater agreement and the dimensionality of the construct, in this paper we apply the newly developed observational protocol to three videotaped mathematics lessons from the NCTE video library of Harvard University in order to explore strengths and weaknesses of this instrument

✉ Lena Schlesinger
lena.schlesinger@uni-hamburg.de

Armin Jentsch
armin.jentsch@uni-hamburg.de

¹ Faculty of Education, Mathematics Education, Universität Hamburg, Von-Melle-Park 8, 20146 Hamburg, Germany

² Australian Catholic University, Sydney, Australia

³ Universität zu Köln, Cologne, Germany

⁴ Centre of Educational Measurement (CEMO) at University of Oslo, Oslo, Norway

and to examine in more detail how the instrument works in practice. Furthermore, this is the first attempt to apply a mixed-methods approach to data gathered with the newly developed observational protocol, thereby allowing deeper insights than those possible by applying only quantitative methods, as has been the practice up to now.

This article is part of the ZDM Mathematics Education issue on ‘Studying instructional quality in mathematics through different lenses’, in which several different frameworks and instruments to assess instructional quality are compared and reflected on.

2 Theoretical rationale of the framework

For the present observational instrument we adapted the three dimensional framework of instructional quality by Klieme and colleagues, consisting of classroom management, student support and cognitive activation (Klieme and Rakoczy 2008; Lipowsky et al. 2009). The reason for choosing this framework as the basis for our instrument was its high acceptance in German-speaking countries during the last ten years. However, this framework has recently been criticized concerning the lack of content-specific aspects of instructional quality (Drollinger-Vetter 2011; Schlesinger and Jentsch 2016; Praetorius et al. 2018). Based on a systematic literature survey, we identified subject-specific characteristics of instructional quality (Schlesinger and Jentsch 2016), which were subsumed under two additional dimensions. These two subject-specific dimensions involve domain-specific aspects that are of influence to student achievements and therefore function as elements of prognostic validation. In addition, these characteristics build on solid mathematics educational theoretical frameworks (Schlesinger et al. submitted). In the following, we elaborate in more detail the reasons for extending the already existing framework by adding two subject-specific dimensions and describe how these were conceptualized.

2.1 Generic dimensions of instructional quality

The framework with three basic dimensions of instructional quality includes aspects that are regarded as crucial for good instruction. Even though it was firstly developed within mathematics instruction, the three basic dimensions are mainly conceptualized as generic dimensions, so that they can be used in different school subjects, and grades. Therefore, the framework is based on general theories of teaching and learning coming mainly from educational science and psychology, not limited to one subject. In the following, we describe the basic dimensions in more detail and give reasons for their relevance for good instruction referring to descriptions and conceptualizations that Klieme et al.

(2001) established as a result of the TIMSS video studies, using high-inference observer ratings analyzed with factor analyses.

The first dimension, classroom management, focuses on different discipline practices that ensure a high-quality learning time with an appropriate atmosphere without disruptions and conflicts that hinder students’ learning processes (Brophy 2000; Kunter et al. 2007; Lipowsky et al. 2009; Praetorius et al. 2012; Taut and Rakoczy 2016). Clear rules and routines, successful interventions when disruptions occur, and a structured and well-organized lesson are evidence-based characteristics of effective classroom management (Kounin 1970). The educational factor underlying this dimension is to maximize the learning time available for students. Different studies could confirm the positive impact of effective classroom management on students’ achievement (Seidel and Shavelson 2007) as well as a positive effect on students’ motivation (Rakoczy 2006).

The second generic dimension, student support, focuses on help for individual students provided by the teacher to assist students in their personal learning processes with adaptive teacher interventions and differentiated learning opportunities (Klieme et al. 2009; Lipowsky et al. 2009; Taut and Rakoczy 2016). A positive and respectful learning climate, constructive feedback and a good relationship between students and teacher are important aspects of this dimension. The psychological factor underlying this dimension is the assumption that students are motivated when they feel self-determined, i.e., being competent, responsible for their learning process, and socially integrated (Deci and Ryan 1985). Studies could provide evidence that effective student support has a positive impact on students’ motivation and interests (Lipowsky et al. 2009).

Cognitive activation, finally, focuses on the cognitive level of the instruction provided by the teacher that challenges students to activate high-level learning processes building on their existing knowledge (Hiebert and Grouws 2007; Klieme et al. 2009; Lipowsky et al. 2009; Praetorius et al. 2014; Taut and Rakoczy 2016). This dimension builds on the constructivist assumption that learning processes cannot be “drummed” into students by the teacher from the outside but that the teacher can only provide high-level cognitive learning opportunities (Bruner 1974; Helmke 2012). Especially the third dimension has a positive impact on students’ achievement, as has been shown by different studies (e.g. Baumert et al. 2010; Lipowsky et al. 2009). Regarding the subject-specific depth of these mainly generic quality dimensions, it seems that cognitive activation is probably most closely related to subject-specific aspects of instructional quality (Klieme and Rakoczy 2008; Schlesinger and Jentsch 2016). Still the dimension can be regarded and is conceptualized as being mainly generic due to its more general, psychologically oriented characteristics.

2.2 Subject-specific aspects of instructional quality

According to the Learning Mathematics for Teaching Project (2011) the mathematical quality of instruction is not sufficiently assessed when only generic frameworks are used for observations in mathematics classrooms. They present examples of instruction in which the lesson is well structured, students are engaged and individually supported, but the *mathematical* quality of the lesson is lacking. Blum and colleagues (2006) equally argue for a rich ‘orchestration’ of the lesson that goes beyond generic characteristics of instructional quality. Therefore, in contrast to generic dimensions, different subject-specific frameworks were developed for mathematics education. These frameworks often built on different theories of teachers’ professional knowledge for teaching.

In more detail, such mathematics educational characteristics of instructional quality may include the usage of appropriate mathematical language and various representations, well-developed teachers’ mathematical explanations at an adequate level of rigor, appropriate examples, responses to students’ mathematical errors, mathematical sense-making activities, problem-solving, proof or modeling tasks and adequate mathematical depth (e.g., including generalizations and connections) (Hiebert et al. 2003; Klieme et al. 2009; Learning Mathematics for Teaching Project 2011; Marder and Walkington 2014; Matsumura et al. 2002; Schoenfeld 2013). This listing is not meant to be complete but contains examples of what is not yet covered within generic frameworks.

Frameworks assessing the subject-specific quality of mathematics instruction are the *Uteach Observation Protocol* (UTOP, Marder and Walkington 2014), the *Mathematics-Scan* (M-Scan, Walkowiak et al. 2014), the *Instructional Quality Assessment* (IQA, Matsumura et al. 2002), the *Elementary Mathematics Classroom Observation Form* (Thompson and Davis 2014) or the *Mathematical Quality of Instruction* (MQI, Hill et al. 2008; Learning Mathematics for Teaching Project 2011) (for an overview see Schlesinger and Jentsch 2016). Finally, there are first approaches to combining the assessment of both generic and also subject-specific characteristics in mathematics instruction (e.g. *TRU Math framework*, Schoenfeld 2013). Still, to our knowledge, such an extension has not yet been developed for the framework with three basic dimensions, which are not focusing on a specific area of mathematics (e.g. geometry, quadratic equations), as was done for example in the Pythagoras study (Klieme et al. 2009).

For such an extension, we first wanted to adopt one of the existing instruments for our classroom observations and combine them with the framework of three basic dimensions. The framework by Schoenfeld seemed to be suitable as it already combined generic and subject-specific characteristics. However, it was not easily possible to match this or other frameworks with the framework of the three basic dimensions because amongst others the TRU framework

focuses primarily on students and not on the teacher. In addition, there existed no instrument that covers process-oriented aspects of mathematics education, which play an important role in the German national standards as they are regarded as crucial for the development of deep mathematical knowledge (Blum et al. 2006). As no instrument was available analyzing in depth instructional quality from a content-related perspective covering the described aspects, a new observational protocol and extension of the framework with three basic dimensions was developed referring to already existing instruments, which were enriched and further developed.

3 Description of the instrument

3.1 Subject-specific dimensions of the instrument

Based on a systematic literature review of existing classroom observation instruments (Schlesinger and Jentsch 2016), expertise in teaching mathematics within the research group, and following the discussion about the German national standards, we developed subject-specific descriptions of instructional quality, which can be empirically divided into two subject-specific dimensions (see Table 1), one covering the ‘subject-related quality’ of instruction (focus on content matter), the other one covering the ‘teaching-related quality’ (focus on practices in mathematics instruction). This classification is also consistent with the discussion on pedagogical content knowledge (Depaepe et al. 2013) that gives rise to a conceptualization with one dimension closely oriented on subject matter and the other dimension related to instructional practices (Buchholtz et al. 2014).

The first subject-specific dimension focuses on the *subject-related* quality of instruction with regard to mathematical correctness and depth (e.g. Baumert et al. 2010; Learning Mathematics for Teaching Project 2011). This first dimension is based on the assumption that students can only learn effectively if the mathematical content is presented in a correct way and covers not only superficial mathematical aspects but meaningful and deep subject matter. Summing up, it is important that the teacher is correct in his or her mathematical language and notations (Hill et al. 2008). Furthermore, teachers’ mathematical explanations and presentations need to be mathematically precise but also more in-depth and understandable for students (Schoenfeld 2013). This holds especially when responding to students’ mathematical errors that may occur during the lesson. In addition, different mathematical competencies should be dealt with during the lesson, e.g., problem solving, modeling, or proving. In the German national standards the support of these mathematical competencies in mathematics instruction is officially required (Blum et al. 2006). These national standards were developed in Germany in 2003 for the purposes

Table 1 Operationalization of two subject-specific dimensions of instructional quality besides the common three generic dimensions

Mathematics educational quality	
Subject-related quality	Teaching-related quality
Dealing with mathematical errors of students (<i>The teacher analyzes students' errors and misconceptions and uses them as an opportunity to learn</i>)	Using multiple representations (<i>The teacher uses various representations within instruction and illustrates the relation between them to facilitate students' access to the mathematical content</i>)
Teacher's mathematical correctness (<i>The teacher is mathematical correct without making content-related mistakes and being imprecise</i>)	Deliberate practice (<i>The focus of the lesson or parts of the lesson is on practicing and the teacher provides exercises that contain opportunities for exploring, reflection and self-differentiation</i>)
Teacher's mathematical explanations (<i>Teachers explanations are well-structured, precise and focusing on fundamental mathematical aspects</i>)	Appropriate mathematical examples (<i>The chosen mathematical examples focus on fundamental mathematical ideas and are suitable for the students</i>)
Mathematical depth of the lesson (<i>The teacher provides generalizations, mathematical connections and possibilities to structure the mathematical content</i>)	Relevance of mathematics for students (<i>The teacher addresses the relevance of the mathematical content and provides connections to students' everyday life so that they may bring in their own experience and interests</i>)
Support of mathematical competencies (<i>The teacher provides the opportunity to deal with mathematical processes such as problem-solving, modeling or reasoning and proof</i>)	

of securing the quality of mathematics instruction and of making students' achievement in German mathematics classrooms comparable. Within these national standards, the following mathematical competencies required that students should acquire, are distinguished: competence in (1) mathematical reasoning and proof, (2) problem solving, (3) mathematical modeling, (4) using different mathematical representations, (5) using mathematical language and communication, (6) calculating by working with symbolic and formal aspects and mathematical tools (Blum et al. 2006).

The second dimension consists of subject-specific aspects that are more *teaching-related*. For instance, it is important that the mathematical content be presented in a way that is accessible for the students within the instruction (Marder and Walkington 2014; Matsumura et al. 2002). This second dimension is based on the assumption that students can only learn effectively and be motivated if the mathematical content within the lesson is accessible and interesting for them. Therefore, it is necessary that different perspectives and representations be used to support the students in their learning processes and that the examples and tasks used are appropriate for the mathematical content (Drollinger-Vetter 2011; Kersting et al. 2012). Furthermore, students need to make sense of what they are learning, for example by means of the teacher providing real-world phenomena that help students to see their relevance for students' lives.

Regarding the three basic dimensions and the subject-specific extension with two dimensions, there exist some overlaps, which seem to be unavoidable. As already described above, it is a matter of discussion whether the three basic dimensions really are generic by nature or if they also cover

some subject-specific aspects. For example the item "dealing with mathematical errors of students" is part of student support in other frameworks and instruments (Praetorius et al. 2018). Still, in these instruments the focus of the item is much more on classroom climate aspects and less on content specific teacher decisions that need some pedagogical content knowledge or diagnostic competencies to analyze students' misconceptions successfully. The same holds for some overlaps to cognitive activation, as this basic dimension is probably most closely related to subject-specific aspects of instructional quality (Klieme and Rakoczy 2008; Schlesinger and Jentsch 2016). For example, the mathematical depth of the lesson probably has connections to the cognitive level of questions and tasks, and the use of different representations can also promote students' cognitive activation but simultaneously supports students in their personal learning processes. Therefore, we expect that these different dimensions cannot be completely separated.

3.2 Operationalization of the instrument

In order to show how we operationalized the instrument, selected item examples are presented in Table 2 as well as a range of indicators that describe incidents that can occur in mathematics instruction, and which were evaluated in order to assess instructional quality (the whole instrument can be found in the Appendix). The observational protocol consists of 26 items that are assessed by high-inference ratings for which the presented indicators describe typical examples for each item. High-inference ratings are observer ratings that have high inference in the research results. The items

Table 2 Example items and indicators of the observation protocol

Dimension	Items	Example item	Indicators
Classroom Management	6	Effective use of lesson time	The lesson starts and ends on time Transitions between lesson phases happen smoothly Lesson time is used for content-related instruction
Student support	7	Dealing with heterogeneity	Additional materials for subgroups of students exist Tasks address different types of students The teacher offers ad-hoc differentiation during lesson (e.g. by varying the cognitive level of questions)
Cognitive Activation	5	Challenging questions and tasks	The teacher asks cognitively appropriate questions The teacher presents challenging tasks The teacher spontaneously encourages cognitive challenges
Subject-related quality	4	Mathematical depth of the lesson	Mathematical generalizations are developed The topic is related to other mathematical topics The content is embedded in a broader mathematical structure
Teaching-related quality	4	Using multiple representations	Multiple representations are used during the lesson Relations between multiple representations are shown

are formulated independently of the specific mathematical topic taught, which makes the instrument applicable for mathematics instruction at all year groups at the secondary level. Four-point Likert scales from 1 = *Does not apply at all* through 4 = *Does fully apply* were used to assess the extent to which the different characteristics were observed.

In addition, six more low-inference categories were developed that assess which mathematical process-oriented competencies are supported within the lesson (0 = not supported; 1 = supported slightly; 2 = mathematical competence in the focus of the lesson). These items are ‘usage of adequate mathematical language’, ‘promotion of mathematical modeling’, ‘promotion of problem solving’, ‘reasoning and proof’, ‘adequate usage of calculations (symbolic and formal aspects)’ and ‘adequate usage of mathematical tools’.

A coding manual was developed for the instrument that contains detailed descriptions of the items with typical examples and guidelines for the assessment (see as one example Table 3).

3.3 In-vivo vs. video-based observation of mathematics instruction

Our instrument was designed to be used in-vivo without videotaping the lessons. This leads to many advantages because videotaping a larger number of mathematical lessons has become extremely difficult in Germany within recent years due to legal restrictions. In contrast, in-vivo instruments have become popular especially for assessments by governmental quality insurance institutions or superintendents of schools (Pietsch and Tosana 2008). In addition, the observers in in-vivo observations are directly in the middle of the lesson and do not see the classroom through a kind of “window” as is the case with videotaped lessons. Aspects that are not covered or not in the focus of the camera are not visible for observers that view only the videotaped lesson (Casabianca et al. 2013).

Still, in-vivo observations also have some disadvantages compared to video-based lesson ratings. Observers cannot

Table 3 Example item and detailed descriptions from the coding manual

Teacher’s mathematical correctness			
Indicators			
• The teacher makes no content-related mistakes • The teacher makes no formal mistakes • The teacher is mathematically precise in language • The teacher is mathematically precise in notation			
Rating of the item			
1	2	3	4
None of the four indicators applies A correct learning process is not possible due to teachers’ content-related and formal mistakes	The teacher makes a few mistakes that partly affect the learning process because some necessary conditions are presented incorrectly	The teacher makes some mistakes or is mathematically imprecise but these errors and imprecisions do not affect the learning process	All four indicators apply The teacher makes no mistakes and is mathematically precise

stop the lesson process so they need to make decisions very quickly, and there is no possibility of watching the lessons again. In addition, the audio quality is another key difference, because if the teacher works with single students these conversations can be difficult to hear in the classroom for in-vivo observers who are sitting far away. In video-based studies, the teacher normally wears a microphone that captures all conversations with students. Still, the placement of the microphone might also have disadvantages for videotaped lessons, as other conversations can get lost (Casabianca et al. 2013). Finally, videotaped lessons can be watched at any time of the day or within a research process so that time effects can be reduced. In-vivo observations have to take place when the lesson takes place.

Until now, there exist only few studies on the effect of observation modes by comparing in-vivo observations in the classroom with ratings of videotaped lessons (Casabianca et al. 2013). Casabianca et al. (2013) explored the effects of the observation mode using the CLASS-S (Secondary) instrument in 82 algebra lessons. Their results show that the mode did not have a great influence on the teacher inferences in a classroom observed over a year. Still they could show that there were some differences in the reliability and inferences in a single lesson. The mean scores and the reliability of inferences were partly higher for in-vivo than video observations. Another research group conducted live video observations—that means seeing a video of the lesson as it happens—for reducing reactivity issues (Liang 2015). The effects of this combination of video and live observation, however, are even more difficult to predict.

3.4 Empirical support of the instrument

The present instrument was developed and used for the first time within the study TEDS-Instruct in which the relation between teacher competencies, instructional quality, and student achievement was analyzed. 38 teachers were assessed twice (lessons of 90 min) within their instruction. To include information about the stability and variability of instructional quality, the lessons were rated multiple times, typically after periods of about 20 min. To ensure inter-rater reliability, the observations were done in-vivo by a team of six trained raters. Two observers from the team who were randomly selected rated a lesson. After each lesson the two raters carried out an intensive debriefing about the lesson and what happened within the lesson. In this debriefing it was possible to modify a rating.

In TEDS-Instruct, the generic dimensions and the two subject-specific scales reached acceptable internal consistencies (classroom management: $\alpha=0.86$; student support: $\alpha=0.71$; cognitive activation: $\alpha=0.82$; subject-related quality: $\alpha=0.77$; teaching-related quality: $\alpha=0.69$). The inter-rater reliability of $ICC>0.60$ for all items after

the debriefing was acceptable (classroom management, $0.62 < ICC \leq 0.88$; student support, $0.80 < ICC \leq 0.96$; cognitive activation, $0.80 < ICC \leq 0.93$; subject-related quality, $0.82 < ICC \leq 0.95$; teaching-related quality, $0.80 < ICC \leq 0.94$). A two-level confirmatory factor analysis showed acceptable fit indices for five dimensions ($\chi^2/df=1.52$, CFI=0.94, RMSEA=0.04) (Schlesinger et al. submitted).

4 Research questions

Three videotaped lessons were provided from the NCTE video library of Harvard University and analyzed with the standardized observational instrument to address the following research questions:

- 1) How does the instrument work in practice apart from the study in whose context it was developed?
- 2) What are the strengths and weaknesses of the newly developed instrument?

After describing the method and analysis we used, we present ratings and results for these three lessons in detail to examine advantages as well as limitations of the instrument.

5 Method and analysis

The three videotaped lessons from the NCTE video library of Harvard University, which we analyzed, are described in more detail in the paper by Charalambous and Praetorius (2018). To utilize this database in the best way, we applied a mixed-methods approach by combining qualitative and quantitative approaches. Quantitative methods, on one hand, tend to ‘forget’ the individual, which is at the same time the consequence of their main advantage, *generalizability* (Kelle and Buchholtz 2015). On the other hand, qualitative methods are useful to find explanations that are often not given by quantitative analyses only (Hiebert and Grouws 2007). In the present paper, we decided to use Qualitative Content Analysis (Mayring 2015) next to the quantitative observer ratings for the purpose of having the results complement each other.

In the present quantitative analysis, two experienced observers simultaneously rated instructional quality with our standardized observational instrument. The observers are two German PhD students in mathematics education who had finalized their studies for becoming teachers in mathematics education. The first and third video (Mr. Smith and Ms. Jones) were divided into two parts which were rated separately, and the second video (Ms. Young) was divided into three parts, i.e., a rating was carried out approximately every 20 min. We present descriptive results at item level for

all five dimensions aggregated across one lesson (median of the ratings). Due to the small sample size, we did no inference statistics calculations (e.g. estimation of standard errors). As we analyzed only one lesson per teacher, our results cannot be generalized to all lessons of the specific teacher. In addition to these quantitative ratings, we explain the ratings in more detail. We selected the expanded support of mathematical processes to be analyzed by Qualitative Content Analysis including six deductive categories (mathematical language, mathematical modeling, problem solving, reasoning and proof, calculations, mathematical tools), because of their core relevance for the development of deep mathematical knowledge (see Sect. 2.2). For this purpose, we used the transcribed lessons and coded these transcriptions using the six mentioned categories, every time something occurred in the text that was relevant and

appropriate to these categories. Finally we decided for each category whether the mathematical process was not supported ('0') within the lesson, supported slightly ('1') or if the mathematical process was in the focus of the lesson ('2').

6 Results

6.1 Descriptive statistics

We first present an overview of the descriptive statistics of the video analysis (Table 4) and then explain the ratings in more detail. We would like to stress that we analyzed only one lesson per teacher, so our results cannot be generalized as comprehensive personal evaluation.

Table 4 Median of two observers' ratings of all measurement points (two measurement points for Lesson 1 and 3 and three measurement points for Lesson 2) (1 = Does not apply at all; 4 = Does fully apply)

Item	Lesson 1 (Mr. Smith)	Lesson 2 (Ms. Young)	Lesson 3 (Ms. Jones)
Classroom management			
Effective use of lesson time	4	3	4
Clear rules and routines	3	3	4
Preventing disturbances	3	3	3
Productive atmosphere	3.5	3	3.5
Advance organization/structuring learning processes	2	3	2
Lesson structure	2	3	2
Student support			
Students' individual support	1	2	2
Dealing with heterogeneity	1	1.5	1
Self-directed learning	1	2	2
Teacher feedback to students	2	3	2.5
Teacher appreciation of students	3	2	3
Student feedback	1	1	1
Support of collaborative learning	1	2	2
Cognitive activation			
Challenging questions and tasks	1	3	2
Supporting metacognition	1	1.5	1
Activation of prior knowledge/ co-construction	2	3	2
Cognitively challenging teaching methods	2	2	2
Facilitating remembering and recalling	2	2.5	3
Subject-related quality			
Dealing with students mathematical errors	n/o	2.5	2
Teachers mathematical correctness	3.5	4	4
Teachers mathematical explanations	2	3	2
Mathematical depth of the lesson	1.5	3	1.5
Teaching-related quality			
Using multiple representations	2	3	1.5
Deliberate practice	1	n/o	1
Appropriate mathematical examples	2	3	2
Relevance of mathematics for students	1	2	1

n/o Not observable (because amongst other aspects the focus of the lesson is not on practicing)

6.1.1 Lesson by Mr. Smith

The first lesson by Mr. Smith is about 40 min long. The lesson is teacher-centered with the whole class using the smartboard. The topic of the lesson is geometry and Mr. Smith teaches his class angles.

6.1.1.1 Classroom management The lesson by Mr. Smith is characterized by high ratings in classroom management regarding discipline practices and learning time. The atmosphere is productive and no disruptions occur. The lesson time is used for content-related instruction, routines for the organization of the lesson are apparent even though the students did not take part in organizing the lesson. Deductions are made for the advance organization and the lesson structure because Mr. Smith does not inform the students about the lesson objectives or the structuring of the learning process and the lesson is not clearly separated into sections.

6.1.1.2 Student support The student support by Mr. Smith is rated low except for his appreciation of the students. He does not ask about students' individual difficulties and he does not provide individual assistance for students. Mr. Smith does not offer any differentiation during the lesson and the students have no opportunity to work in a self-directed manner. Mr. Smith's answers and feedback to students are often short without giving constructive and forward-looking feedback.

6.1.1.3 Cognitive activation The cognitive activation of Mr. Smith's lesson is also rated low. The questions and tasks within the lesson are not cognitively challenging and only repeat students' knowledge of angles of previous lessons. Although he activates their knowledge a few times, no co-construction and further development is visible. There are few cognitively activating teaching methods used within the lesson even though several students always answer in chorus to his questions. Mr. Smith does not provide any time for metacognitive processes and students have no opportunity to reflect on their learning process.

6.1.1.4 Subject-related quality As his questions focus mainly on definitions without fostering high-level thinking in his students, and they require only single-word answers by the students, no students' errors or incorrect answers occurred that he needs to deal with. In addition, he does not explore any students' misconceptions. His ratings on 'teacher's mathematical correctness' are quite high because he makes no formal mistakes. Still, he is a few times mathematically imprecise within the first 20 min ("Can an angle be greater than 360 degrees?"). Mr. Smith does not give any detailed mathematical explanations over the whole lesson and does not slow down in order to address important aspects

that need more detailed explanations for his students. The mathematical depth of the lesson is overall quite low. Mr. Smith makes almost no mathematical connections or generalizations within the whole lesson and only structures the knowledge on angles but does not address any mathematical concepts. Overall, the ratings for the subject-related quality of the lesson are mainly low.

6.1.1.5 Teaching-related quality Mr. Smith uses two different representations of angles (symbolic and figurative), but they are not always clearly connected to each other. The focus of the lesson is on practicing existing knowledge, but Mr. Smith does not explain the importance of the exercises and there are no opportunities for exploring or reflection. The examples that Mr. Smith uses within the lesson are often at a low aspiration level without focusing on real life problems. They do not vary over the lesson and are often repeating the theme. Still, they fit the topic dealt with and address different important angles, and therefore, he received the second lowest rating for this item. The relevance of the topic is not made clear to the students at all; they are operating with angles without addressing why angles could be relevant for them. In addition, they cannot bring in personal experiences or interests. Overall, the ratings for the teaching-related quality of the lesson are low.

6.1.2 Lesson by Ms. Young

The second lesson, by Ms. Young, is 70 min long and the topic is about multiplying whole numbers and doubling and halving factors. The objective of the lesson is to investigate how doubling and halving factors affect the product.

6.1.2.1 Classroom management The lesson by Ms. Jones is characterized by a moderate rating of classroom management. The lesson time is mainly used for content-related instruction even though a few minutes are lost due to organizational problems. Rules and routines are apparent but Ms. Young needs to repeat them more often so that it seems that they are not completely accepted and implemented in the class. Ms. Young explains precisely what will happen within the lesson and describes the objective of the lesson at the beginning. The atmosphere is productive most of the time but Ms. Young is not always aware of everything that happens in the classroom.

6.1.2.2 Student support The student support by Ms. Young is mainly rated as low. Her appreciation of students is low, she is often rigorous and impatient, but her feedback to the students is most of the time constructive and quite sophisticated ("Think about it. I'll come back. You know what you're trying to say, but you don't know how to communicate it. So I'll give a few minutes to think about it.") The

whole class works on the same problem, so there is nearly no differentiation within the lesson. The students work most of the time in small groups and Ms. Young walks around in the class, providing individual assistance to students a few times. Still, the lesson is very teacher-centered, i.e., Ms. Young does not use the potential of self-directed learning and collaborative learning processes between students.

6.1.2.3 Cognitive activation The cognitive activation of the lesson by Ms. Young is rated as moderate. The questions and tasks she provides for the students are cognitively challenging and after Ms. Young activates students' prior knowledge on the task, they develop their knowledge co-constructively within the lesson. Ms. Young partly provides time for metacognitive processes and students can reflect on their learning process ("That's why we are looking for strategies. Strategies that will help us to be—to get to our answer quickly without wasting too much time, efficient ones. So some problems we offer you the opportunity to do this."). The methods used within the lesson are mainly not cognitively challenging; at the end of the lesson only a few minutes are spent on securing the acquired knowledge within the class.

6.1.2.4 Subject-related quality For 'dealing with students errors' Ms. Young receives ratings between '2' and '3' because she sometimes is very fast in correcting students' errors without using mistakes or ideas as a learning opportunity or analyzing errors in more detail. In addition, tolerance for mistakes is not always visible. Her ratings on 'teachers' mathematical correctness' are high. Ms. Young's mathematical explanations are more detailed than the explanations by Mr. Smith. She explains in detail why the strategy for multiplying whole numbers is useful and how it works, but sometimes her explanations are long and not well organized so that they are not always focused on the essential aspects appropriate for the students. The mathematical depth of the lesson is higher than that of the lesson by Mr. Smith and the mathematical content is structured, generalized and abstracted from the given examples. Furthermore, the lesson offers the opportunity for the students to engage in reasoning and proving. Overall, her ratings for the subject-related quality of the lesson are mainly moderate to high.

6.1.2.5 Teaching-related quality Within the lesson, Ms. Young and her students use different representations of multiplying whole numbers (verbal, symbolic and figurative—the students can use pictures, arrays, cubes, or story problems) that they link to each other multiple times, therefore she receives high ratings for this item. The focus of the lesson is not on practicing, therefore the rater could not rate this item. The examples that Ms. Young uses within the lesson are appropriate to the topic dealt with and they are

partly connected to real life (apples in shelves). Still these examples are not personally relevant for the students and not connected to their personal interests or experiences, therefore Ms. Young receives the second lowest rating category for the 'relevance of mathematics'. Overall her ratings for the teaching-related quality of the lesson are between '2' and '3'.

6.1.3 Lesson by Ms. Jones

The third lesson, by Ms. Jones, is 56 min long. The topic is about multiplying a whole number by a fraction. The goal of the lesson is to learn three different ways for this multiplication.

6.1.3.1 Classroom management The lesson by Ms. Jones is characterized by a mainly high rating of classroom management. Clear rules and routines are visible and Ms. Jones has a good overview of her class ("When I see that everybody in the room has their hand on top of their head, I will know we're ready to move on"). The lesson time is used for content-related instruction and Ms. Jones prevents disturbances successfully most of the time, so that a productive atmosphere exists for nearly the whole lesson. Deductions are given for the lesson structure and organization as the lesson objectives are not really clear at the beginning (it is mainly a 'cold start' without knowing what will happen), and the lesson is not clearly separated into sections, but switches between single/group work and whole class mode unexpectedly.

6.1.3.2 Student support The student support by Ms. Jones is mainly rated moderate. She walks around in the class and takes some time for individual students, but most of the time this is just for the purpose of checking the progress of the students and not for giving individual assistance. Ms. Jones does not offer any differentiation during the lesson and the students have few opportunities for self-directed work. Her appreciation of students is mainly high but her feedback to students is often short and not sophisticated ("Good job, student O. Good job, student M."). At the end of the lesson she supports some collaborative learning processes but these are not in the focus of the lesson. Like Mr. Smith and Ms. Young, Ms. Jones does not ask students for feedback.

6.1.3.3 Cognitive activation The cognitive activation of Ms. Jones lesson is rated low to moderate. The questions she asks and tasks she presents are mainly not challenging and the students only copy down what the teacher is presenting by creating posters with different strategies of multiplying a whole number with a fraction. As in the case of Mr. Smith, Ms. Jones does not provide any time for metacognitive processes, and students have no opportu-

nity to reflect on their learning processes. Ms. Jones partly activates students' prior knowledge but the knowledge of multiplying a whole number by a fraction is not developed co-constructively in class but merely presented in very small steps by Ms. Jones ("I'm going to teach you three ways to do it.").

6.1.3.4 Subject-related quality Within the lesson Ms. Jones often does not use students' errors as an opportunity for further learning processes and she does not analyze or address students' wrong thinking processes to help them to understand their mistakes. In contrast, she often just waits for the right answer by the students, even if they apparently did not understand their mistakes. Therefore, her 'dealing with student errors' is only rated as '2'. Like Mr. Smith and Ms. Young, Ms. Jones also gets high ratings for her 'mathematical correctness', because she does not make any formal mistakes. Her ratings on mathematical explanations are only at the low level, because the examples she uses within her explanations do not fit the problem (patting a baby on the back five times or handing in blocks five times). In addition, her explanations do not focus on why the strategies and procedures work. The mathematical depth of the lesson is rated low, because Ms. Jones does not develop any connections to other topics or mathematical content and does not attempt any generalizations or abstractions. The whole lesson focuses on procedures ("We can't do any math if we don't have numbers"). Overall, the ratings for the subject-related quality of the lesson are mainly low ('2').

6.1.3.5 Teaching-related quality At the beginning of the lesson, there is only one representation used for multiplying a whole number with a fraction (symbolic), but throughout the lesson Ms. Jones offers other representations (figurative and also verbal) even though they are not always well connected to each other. The lesson is rated as '1' for 'deliberate practice', because the exercises that the students conduct are not focused on students' mathematical understanding regarding the mathematical concept, and the exercises are neither self-differentiated nor reflective, nor do they give students the opportunity to discover new mathematical content. Similarly to the lesson by Mr. Smith, the relevance of the topic is not made clear to the students at all and they cannot bring in any personal experiences or interests. The real life examples that Ms. Jones tries to include even do not fit well with the topic of the lesson and do not focus on the essential aspects (multiplying with a fraction) of the mathematical topic. Therefore, the ratings on the usage of 'appropriate examples' are low as well. Overall, the ratings for the teaching-related quality of the lesson are mainly low (between '1' and '2').

6.2 Qualitative analyses for the support of mathematical competencies

Regarding the 'support of mathematical competencies', the two raters coded the lessons focusing on the following six categories: usage of adequate mathematical language, promotion of mathematical modeling, promotion of problem solving, mathematical reasoning and proof, adequate usage of calculations (symbolic and formal aspects) and of mathematical tools, which reflect process-oriented mathematical competencies.

6.2.1 Mathematical language

Promoting mathematical language was slightly supported in the first and third lesson (see the examples below). This was particularly the case when the teachers asked students for the names of mathematical objects. The lesson by Ms. Young did not focus on promoting mathematical language.

First example (first lesson):

- Mr. Smith: Okay. Point B. So, I can see both are rays and B is what starts both those rays. What is a name for where two rays or where two—we could say line segments if we drew line segments—what's the name where those two meet?
- Students: Vertex.
- Mr. Smith: Vertex. Okay. And when I have more than one, what's the plural of that word?

Second example (third lesson):

- Ms. Jones: Very good. So I take 15 and I put inside. It becomes my dividend. And 4 becomes—what is that word that we use for the number that's outside the box? Raise your hand. What is that word that we use, student R?
- Student: The divisor.
- Ms. Jones: Divisor. So 15 becomes my dividend and 4 becomes my divisor, and I divide it out.

6.2.2 Mathematical modeling and problem solving

Modeling tasks could not be observed during the three lessons, i.e., the raters could not identify any tasks which fell under the category mathematical modeling. Regarding the inclusion of mathematical problem solving, it could be slightly observed in the second lesson by Ms. Young.

- Ms. Young: ... Can somebody use the story like Miss S's apples? What is happening to explain what is happening to this problem, to show that the size of the product will double if we double

one of the factors? Yes, student D, I see your hand back there. Student J, yes?

In the first and third lesson both raters did not code any situation as solving mathematical problems, i.e., students were mostly concerned with other tasks, such as calculations.

6.2.3 Reasoning and proof

Reasoning and proof was dealt with in the second lesson by Ms. Young and was in the focus of the lesson. However, for the first and third lesson by Mr. Smith and Ms. Jones both observers could not identify that the teachers emphasized mathematical reasoning in their classrooms.

Ms. Young: Is it true?

Students: Yes.

Ms. Young: How can we justify that? That's where we at. I want you to do it first, and then we share. How can we justify? How can we justify that 15 times 8 is the same as 30 times 4?

6.2.4 Calculations and mathematical tools

The usage of calculations was coded throughout all three lessons. The observers agreed that practicing (using symbolic and formal aspects) was in the focus within the first and third lesson by Mr. Smith and Ms. Jones, especially during the third one, and less focused during the second lesson by Ms. Young.

Ms. Jones: So when you were doing multiplication, it's still just repeated addition, except this time instead of adding together 2 plus 2 plus 2, you're adding together three-fourths plus three-fourths plus three-fourths.

The usage of mathematical tools was coded minimally during the first lesson dealing with geometry.

Mr. Smith: We're gonna do some stuff. You're gonna get to work on those—measuring some angles yourself. Okay. Let's measure these. Who wants to come show us how to put the protractor on one of those angles? Student G. Don't [inaudible] the protractor at the angle. Okay. What's the measure of that angle?

To sum up, in the first and third lesson teachers' foci were on training of mathematical language, i.e., repetition of mathematical concepts, and practicing (using symbolic and formal aspects). Modeling problems or proofs were not

observable during these two lessons. However, in the second lesson, the focus was on mathematical reasoning and argumentation as well as on solving mathematical problems (see Table 5).

7 Strengths and limitations of the instrument

In this paper, three videos of American mathematics lessons from the NCTE video library of Harvard University were analyzed with a new standardized observational protocol developed within the German mathematics educational context. The objective was to explore strengths and weaknesses of the newly developed instrument and to examine in more detail how our instrument can be used for analyzing mathematics instruction. The qualitative analysis performed in addition to the quantitative analysis could give further insights into what happened in the classroom from a content-related perspective; especially it enabled the identification of which mathematical processes were supported within the lessons.

The presented observational protocol has advantages but also some limitations. The instrument was developed for assessing instructional quality in-vivo without using video, which allows the evaluation of mathematical lessons at a broader range, as it is getting—at least in Germany—more and more difficult to get the permission of students and their parents for videotaping. Therefore, the items and indicators were developed in such a way that rating could be done quickly within the lesson. Videotaped lessons such as those in the presented cases can be rated more easily. The observation in-vivo necessitating fast rating restricts the observable complexity. In addition, the high-inference of the items may also lead to some rater disagreement because the ratings always need some estimation by the raters.

Comparable to other observational instruments, our instrument focuses on some aspects of instructional

Table 5 Overview of the items on 'support of mathematical competencies' after quantification (0 = not supported; 1 = supported slightly; 2 = mathematical process in the focus of the lesson)

Item	Lesson 1 (Mr. Smith)	Lesson 2 (Ms. Young)	Lesson 3 (Ms. Jones)
Mathematical language	1	0	1
Mathematical modeling	0	0	0
Problem solving	0	1	0
Reasoning and proof	0	2	0
Calculations (using symbolic and formal aspects)	1	1	2
Mathematical tools	1	0	0

quality and does not consider other aspects. The evaluation is completely standardized, so that aspects which cannot be assessed by external raters are not covered by this instrument. This standardization does not allow the observed aspects of instructional quality to be seen as an entire “to-do-list” for good instruction, especially with reference to the depth of subject-specific aspects that are observed (Steinweg 2011). Two aspects that we first wanted to include in the instrument but finally needed to remove due to raters’ difficulties in their observation and rating were teacher’s adaptivity and the separation of learning and assessment situations in the lesson. Even though these aspects are important for good instruction, it was not possible to observe these aspects in the lesson without knowing in advance the whole lesson planning of the teacher. Another advantage—but also limitation—of the instrument is its applicability in various mathematics classrooms and different instructional settings—as done in the research described in this paper—as the items and indicators are formulated independently of specific mathematical topics. However, due to this fact that the instrument does not focus on a specific mathematical topic, it is not possible to assess the conceptual coherence of the content presented (so-called ‘elements of understanding’, see Drollinger-Vetter 2011). For this purpose, the instrument would have had to be restricted to one topic or differentiated for each topic observed. Further developments may include this differentiation.

Regarding the three analyzed US American lessons, some of the items that are developed within the German context differentiate well between the lessons whereas others seem to be less adequate. Regarding the three basic dimensions, the first and third lesson by Mr. Smith and Ms. Jones received high ratings for classroom management and low ratings for cognitive activation. The lesson by Ms. Young in contrast received higher ratings for cognitive activation. Even though none of the lessons showed high student support, the lesson by Ms. Jones was rated slightly higher than the two other lessons. Regarding the subject-specific dimensions, with the exception of mathematical correctness, especially Mr. Smith and Ms. Jones received relatively low subject-specific ratings for their lessons whereas the subject-related and teaching-related mathematics educational quality of the lesson by Ms. Young were rated between moderate and high. Overall, based on these ratings, one could assume (if this lesson would be representative for their lessons in general) that Ms. Young’s class shows higher students’ achievement and that she has the highest amount of pedagogical content knowledge. However, these external stances for validation are beyond this study and the data provided. In any case, due to sample size—namely, three teachers and one lesson per each teacher—only descriptive results could be presented, which cannot be generalized to all lessons of a teacher. Regarding the support of mathematical competencies within

these three lessons, it seems that the focus on calculations and the dominance of rules and execution of algorithms may still be a characteristic of US American mathematics teaching, at least partly, which has been described as result of the TIMSS-video study (Hiebert et al. 2003).

A further advantage of the observational protocol, apart from its easy applicability and saving of resources, is its possibility of describing the variability of instructional quality throughout the lesson as the lessons can be divided into several parts (at least two parts) and these parts can be assessed separately by two raters. This facility is not common in German research currently but is more established in the American context, and it has not yet been realized for the three basic dimensions. Based on this approach it is possible to reduce rater bias, which is caused by long observation periods, as the coding is aggregated to one rating at the end of the lesson. This is especially important since some aspects of instructional quality vary more than others during a lesson (e.g., the student support or the level of cognitive activation).

The observational protocol assesses generic aspects of instructional quality (the three basic dimensions) as well as subject-specific aspects at the same time. Thus, it is possible to analyze relations between these different aspects. First approaches combined different instruments (e.g. CLASS and MQI) to analyze these relations in more detail (Blazar et al. 2017), but including generic as well as subject-specific aspects in one instrument has the advantage that researchers do not need to switch between different instruments, which would make the research process and the observations even more difficult.

Some generic and subject-specific aspects vary almost independently of each other so that it is not possible to generalize from one dimension of instructional quality to another. Until now, we have not examined these variations in detail. Thus further analyses of our data gathered with the observational protocol, maybe even with additional samples and in different contexts, are needed in order to analyze systematically the relations between the different aspects of instructional quality. First analyses indicate that the subject-specific dimensions have the highest correlation to cognitive activation (Schlesinger et al. submitted).

In contrast to other instruments—for example the TRU framework and analysis instrument by Schoenfeld (2013)—our focus is not directly on the students and their learning processes within the lesson. The focus of our instrument is primarily on the teacher and his or her instructional approaches and behavior, even though it is often not possible to assess the quality without considering students’ interactions and reactions to the teachers’ behavior. Anyway, the lack of focus on students’ reactions and behavior could be remedied by complementing the observers’ assessment with a student questionnaire implemented after the lesson.

To summarize, our newly developed instrument offers the opportunity to present a more extensive and complete picture of instructional quality, in contrast to frameworks that analyze only generic or subject-specific aspects. As the evaluation and the assessment of mathematics instruction and its quality has become more important within recent years especially with regard to the role of teachers' professional competencies in improving teaching, our observational instrument may be helpful for researchers as well as for practitioners. The instrument has the advantage that it enables the analysis of instructional quality from a generic as well as a subject-specific perspective, in order for researchers and educators to understand better what happens in the classroom and to give feedback to teachers for their professional development.

Appendix: Items and indicators of the Observational instrument for TEDS-Instruct

Items	Indicators
D1: classroom management (six items)	
Effective use of lesson time	The lesson starts and ends on time Transitions between lesson phases run smoothly The lesson time is used for content-related instruction
Clear rules and routines	Patterns for the organization of the lesson are apparent The students take part in organizing the lesson
Preventing disturbances	The teacher successfully prevents emerging disturbances immediately The teacher is aware of everything that happens in the classroom
Advance organization/structuring learning processes	The teacher informs the students about the lesson objectives The teacher's expectations are apparent Tasks are given in precise language
Productive atmosphere	The volume level is appropriate Students react to teacher's signals The students and the teacher do not interrupt each other
Lesson structure	A common thread is apparent in the lesson The lesson is separated clearly into sections The teacher ends the lesson in an appropriate manner

Items	Indicators
D2: student support (seven items)	
Students' individual support	The teacher asks about students' individual difficulties/individual progress The teacher takes time for individuals The teacher provides individual assistance for students
Dealing with heterogeneity	Additional materials for subgroups of students exist Tasks address different types of students The teacher offers ad-hoc differentiation during lesson (e.g. by varying the cognitive level of questions)
Self-directed learning	Students check their results independently with a sample solution The teacher encourages students to work independently Students may decide whether they would like to work in groups or not
Teacher feedback to students	The teacher's feedback is sophisticated The teacher's feedback is constructive The teacher's feedback is forward-looking
Teacher appreciation of students	The teacher is patient The teacher positively enhances the students' work The teacher encourages the students
Student feedback	The teacher asks for feedback The teacher reacts to students' feedback The teacher and students talk about issues in class
Support of collaborative learning	The teacher initiates collaborative learning processes between students The teacher provides tasks that require agreement The teacher mediates interaction processes The students help each other
D3: cognitive activation (five items)	
Challenging questions and tasks	The teacher asks cognitively appropriate questions The teacher presents challenging tasks The teacher spontaneously encourages cognitive challenges
Supporting metacognition	At least one metacognitive sub-process takes place The teacher provides time for metacognitive processes Students reflect on their learning processes

Items	Indicators
Activation of prior knowledge/ building on students' ideas/ co-construction	The teacher asks for students' beliefs concerning the topic The students explain the task in their own words The teacher activates and explores students' prior knowledge Knowledge is developed co-constructively in class
Cognitively challenging teaching methods	Cognitively challenging teaching methods are used The teacher provides enough time to think about the tasks The teaching methods correspond to the content and the class
Facilitating remembering and recalling	The teacher provides enough examples and helpful reminders The teacher provides enough repetitions Relevant steps are discussed with the whole class
D4: subject-related quality (four items)	
Dealing with mathematical errors of students	The teacher uses students' errors as opportunities to learn The teacher analyzes students' errors and misconceptions The teacher is tolerant towards students' errors Students correct their errors on their own
Teacher's mathematical correctness	The teacher does not make any content-related or formal mistakes The teacher is precise concerning mathematical language and notation
Teacher's explanations	The teacher explains slowly, especially when difficulty arises The teacher focuses on the fundamental mathematical aspects Teacher's explanations are suitable for the students Teacher's explanations are well-structured and precise
Mathematical depth of the lesson	Mathematical generalizations are developed The topic is related to other mathematical topics The content is embedded in a broader mathematical structure
D5: teaching-related quality (four items)	
Using multiple representations	Multiple representations are used during the lesson Relations between multiple representations are shown
Deliberate practice	The teacher explains the importance of the exercises The exercises contain opportunities for exploring and reflection The exercises are self-differentiating

Items	Indicators
Appropriate mathematical examples	The examples contain fundamental mathematical ideas The teacher provides an appropriate number of examples The examples are taken from students' everyday life
Relevance of mathematics for students	The teacher provides connections to students' everyday life The teacher addresses the relevance of content Students may bring in their own experience and interests to class
German national standards/Support of mathematical competencies (six items)	
Mathematical language	The teacher provides time for building new concepts The teacher corrects language errors in an appropriate manner The teacher initiates the adequate use of mathematical language
Mathematical modeling	The teacher encourages transition processes between the real world and maths The teacher encourages the validation of students' results The teacher encourages connections between different subjects
Problem solving	The teacher fosters the use of heuristic strategies The teacher poses mathematical problems The teacher encourages the development of new mathematical content
Reasoning and proof	The teacher encourages mathematical reasoning The teacher demands mathematical explanations The teacher addresses mathematical proofs
Calculations (using symbolic and formal aspects)	The teacher provides time for training the use of mathematical symbols The teacher provides time for training basic skills in mathematics
Mathematical tools	The students use mathematical tools reasonably The teacher uses mathematical tools correctly

Items are rated on a four-point Likert scale from 1 = *Does not apply at all* through 4 = *Does fully apply*

References

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge,

- cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133–180.
- Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment*, 22(2), 71–94.
- Blum, W., Drüke-Noe, C., Hartung, R., & Köller, O. (2006). *Bildungsstandards Mathematik: Konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichts Anregungen, Fortbildungsideen*. Berlin: Cornelsen Scriptor.
- Brophy, J. (2000). *Teaching*. Brüssel: International Academy of Education.
- Bruner, J. (1974). *Toward a theory of instruction*. Cambridge: Harvard University Press.
- Buchholtz, N., Kaiser, G., & Blömeke, S. (2014). Die Erhebung mathematikdidaktischen Wissens—Konzeptualisierung einer komplexen Domäne. *Journal für Mathematik Didaktik*, 35(1), 101–128.
- Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, 73(5), 757–783.
- Charalambous, C., & Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: setting the ground for understanding instructional quality more comprehensively. *ZDM Mathematics Education*, 50(3). <https://doi.org/10.1007/s11858-018-0914-8>. (this issue).
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior. Perspectives in social psychology*. New York: Plenum.
- Depaepe, F., Verschaffel, L., & Kelchtermans, G. (2013). Pedagogical content knowledge: A systematic review of the way in which the concept has pervaded mathematics educational research. *Teaching and Teacher Education*, 34, 12–25.
- Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit: Fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht*. Münster: Waxmann.
- Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.
- Hiebert, J., Gallimore, R., Garnier, H., & Stigler, J. (2003). *Teaching mathematics in seven countries. Results from the TIMSS 1999 video study*. Washington: National Center for Education Statistics.
- Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Charlotte: Information Age.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4), 430–511.
- Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, 42(2), 371–406.
- Kelle, U., & Buchholtz, N. (2015). The combination of qualitative and quantitative research methods in mathematics education: A “mixed methods” study on the development of the professional knowledge of teachers. In A. Bikner-Ahsbahr, C. Knipping & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education: Examples of methodology and methods* (pp. 321–364). Dordrecht: Springer.
- Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, 49(3), 568–589.
- Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster: Waxmann.
- Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54, 222–237.
- Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: “Aufgabenkultur” und Unterrichtsgestaltung. In BMBF (Ed.), *TIMSS—Impulse für Schule und Unterricht, Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (pp. 43–58). Bonn: BMBF.
- Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York: Holt, Rinehart and Winston.
- Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction*, 17(5), 494–509.
- Learning Mathematics for Teaching Project (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14, 25–47.
- Liang, J. (2015). Live video classroom observation: An effective approach to reducing reactivity in collecting observational information for teacher professional development. *Journal of Education for Teaching: International Research and Pedagogy*, 41(3), 235–253.
- Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, 19(6), 527–537.
- Marder, M., & Walkington, C. (2014). Classroom observation and value-added models give complementary information about quality of mathematics teaching. In T. Kane, K. Kerr & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measuring Effective Teaching project* (pp. 234–277). New York: Wiley.
- Matsumura, L. C., Garnier, H. E., Pascal, J., & Valdés, R. (2002). Measuring instructional quality in accountability systems: Classroom assignments and students achievement. *Educational Assessment*, 8, 207–229.
- Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In A. Bikner-Ahsbahr, C. Knipping & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education. Examples of methodology and methods* (pp. 365–380). Dordrecht: Springer.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119.
- Pietsch, M., & Tosana, S. (2008). Beurteilereffekte bei der Messung von Unterrichtsqualität. *Zeitschrift für Erziehungswissenschaft*, 11(3), 430–452.
- Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM Mathematics Education*, 50(3). <https://doi.org/10.1007/s11858-018-0918-4>. (this issue).
- Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction*, 22, 387–400.
- Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12.
- Rakoczy, K. (2006). Motivationsunterstützung im Mathematikunterricht: Zur Bedeutung von Unterrichtsmerkmalen für die Wahrnehmung der Schülerinnen und Schüler. *Zeitschrift für Pädagogik*, 52(6), 822–843.

- Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM Mathematics Education*, 48(1), 29–40.
- Schlesinger, L., Jentsch, A., Kaiser, G., Blömeke, S., & König, J. (under review). Messung fachspezifischer Unterrichtsqualität im Mathematikunterricht. (Manuscript submitted).
- Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM Mathematics Education*, 45(4), 607–621.
- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499.
- Steinweg, A. S. (2011). Einschätzung der Qualität von Lehr-Lernsituationen im mathematischen Anfangsunterricht—ein Vorschlag. *Journal für Mathematik Didaktik*, 32(1), 1–26.
- Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction*, 46, 45–60.
- Thompson, C. J., & Davis, S. B. (2014). Classroom observation data and instruction in primary mathematics education: improving design and rigour. *Mathematics Education Research Journal*, 26(2), 301–323.
- Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics*, 85(1), 109–128.

9 Ergebnisse

In diesem Kapitel werden die Ergebnisse der vorliegenden Dissertation auf der Basis der formulierten Fragestellungen in Kapitel 2 noch einmal zusammenfassend dargestellt. Die Präsentation beschränkt sich an dieser Stelle auf die wesentlichen Ergebnisse. Detailliertere Informationen können den jeweiligen Publikationen entnommen werden.

9.1 Theoretische Ergebnisse zur Entwicklung des Beobachtungsinstruments

Die ersten zwei Forschungsfragen der vorliegenden Dissertation beziehen sich auf die theoretische Entwicklung des Beobachtungsinstruments, indem einerseits die Frage nach einer theoretischen Konzeptualisierung von Aspekten mathematikdidaktischer Unterrichtsqualität aufgeworfen wird und andererseits sowohl theoretische als auch methodologische Herausforderungen bei der Erfassung einer mathematikdidaktischen Unterrichtsqualität mit externen Beobachterinnen und Beobachtern genauer in den Blick genommen werden.

Die erste Publikation, die in der Zeitschrift *ZDM Mathematics Education* erschienen ist, untersuchte diese beiden Forschungsfragen. Darüber hinaus konnte auch die zweite Publikation, die voraussichtlich im *Journal für Mathematik-Didaktik* erscheinen wird, hierzu einen Beitrag leisten, indem die Instrumententwicklung detailliert dargelegt wurde.

Mithilfe einer systematischen Literaturrecherche und darauf aufbauender theoretischer Analysen bisheriger standardisierter, fachspezifischer Beobachtungsinstrumente konnten zwei Perspektiven auf fachspezifische Qualitätsaspekte im Mathematikunterricht herausgearbeitet werden. Einerseits konnte eine *mathematische, stoffbezogene Perspektive* auf fachspezifische Qualitätsaspekte identifiziert werden, die sich auf die fachliche Tiefe, Korrektheit und Präzision der mathematischen Inhalte und Prozesse bezieht. Andererseits konnte eine *unterrichtsbezogene, fachdidaktische Perspektive* herausgearbeitet werden, die sich auf die didaktisch-methodische Umsetzung und Darstellung der mathematischen Inhalte und Prozesse im Unterricht bezieht. Wie bereits in Kapitel 3.4 genauer dargelegt, konnte eine solche zweigeteilte Konzeptualisierung in Einklang gebracht werden mit neueren Konzeptualisierungen des fachdidaktischen Wissens von Mathematiklehrkräften (Buchholtz et al., 2014). Eine genaue Darlegung der zwei Subdimensionen fachspezifischer Unterrichtsqualität erfolgte in Kapitel 5.2.2. Diese Ausdifferenzierung bildete die Grundlage für die Entwicklung und Operationalisierung des vorliegenden Beobachtungsinstruments.

Darüber hinaus wurden methodologische Entscheidungen insbesondere zur Anzahl der Rater, zur Anzahl der Stunden, zu Bewertungsintervallen und zum Grad der Standardisierung bisheriger Forschungsprojekte in diesem Bereich genauer analysiert, um auf der Basis dieser Ergebnisse begründete Entscheidungen für das Projekt TEDS-Unterricht treffen zu können. Solche methodologischen Entscheidungen basieren dabei oftmals auf Ergebnissen von Generalisierbarkeitsanalysen und Entscheidungsstudien (Brennan, 2011; Praetorius,

2014; Praetorius et al., 2014; Shavelson & Webb, 1991), die im Anschluss an erste Datenerhebungen durchgeführt werden, um darauf aufbauend Entscheidungen für zukünftige Forschungsvorhaben treffen zu können. Auch wenn ein projektübergreifender Konsens nicht gefunden werden konnte, so ließen sich dennoch Aspekte herausarbeiten, die für uns handlungsleitend waren:

- Jede Unterrichtsstunde sollte von mindestens zwei Ratern beobachtet und bewertet werden.
- Jede Lehrperson sollte mindestens zweimal im Unterricht besucht werden (abhängig von dem beobachteten Aspekt sind jedoch zum Teil auch deutlich mehr Stunden notwendig).
- Eine Unterteilung der Bewertungseinheit innerhalb einer Stunde erscheint wünschenswert, da manche Qualitätsaspekte innerhalb einer Stunde deutlich variieren können
- Eine Standardisierung des Inhalts der Stunde, der beobachteten Klassenstufe, der Lernziele oder der Unterrichtsstrukturierung (Einführungsstunden, Übungsstunden, ...) kann die Reliabilität der Daten erhöhen.

Aus diesem Grund wurde für das Projekt TEDS-Unterricht die Entscheidung getroffen, jede Unterrichtsstunde von zwei geschulten Ratern bewerten zu lassen sowie jede Lehrperson zweimal jeweils 90 Minuten im Unterricht der Sekundarstufe I zu besuchen. Aus methodischen und empirischen Gründen wurde darüber hinaus entschieden, jede Unterrichtsstunde in vier Messzeiträume zu unterteilen und jeden Messzeitraum einzeln zu bewerten. Eine genaue Darstellung der Analyseeinheit erfolgte in Kapitel 5.4. Weiterhin wurde entschieden, die Lehrpersonen aufzufordern, in einer der beiden beobachteten Stunden inhaltlich neuen Stoff einzuführen, um damit eine Breite der zu beobachtenden Aspekte zu ermöglichen und insbesondere Aspekte der kognitiven Aktivierung und der mathematikdidaktischen Unterrichtsqualität beobachtbar zu machen.

9.2 Empirische Ergebnisse

Die dritte und vierte Forschungsfrage der vorliegenden Dissertation fokussieren erste empirische Ergebnisse, die im Rahmen von TEDS-Unterricht entstanden sind. Einerseits wurde untersucht, ob die Qualität des Mathematikunterrichts mit mehreren theoretisch entwickelten Dimensionen reliabel erfasst werden kann. Andererseits wurden Zusammenhänge zwischen den etablierten drei Basisdimensionen von Unterrichtsqualität und den zwei neu entwickelten fachspezifischen Qualitätsdimensionen untersucht.

Die zweite Publikation dieser Dissertation, die voraussichtlich im *Journal für Mathematik-Didaktik* erscheinen wird, präsentiert die empirischen Ergebnisse zur Beantwortung dieser Forschungsfragen.

Hierzu wurde zunächst eine explorative Faktorenanalyse der fachspezifischen Qualitätsaspekte durchgeführt, um die theoretische Unterteilung einer stoffdidaktischen und einer unterrichtsbezogenen Subdimension fachspezifischer Unterrichtsqualität empirisch zu untersuchen. Die empirischen Ergebnisse zeigen, dass sich in den Daten ein stoffdidaktischer Faktor (u.a. Fachliche Korrektheit, Fachliche Erklärungen) und ein unterrichtsbezogener Faktor (u.a. Verwendung von Repräsentationsformen, Geeignete Beispiele) identifizieren lassen. Bei zwei Items (Fachliche Tiefe und Intelligentes Üben) kam es zu Doppelloadungen auf beiden Faktoren, so dass hier zugunsten der theoretischen Diskussion eine Zuordnung zu einer der zwei Subdimensionen vorgenommen wurde.

Hinsichtlich der Skalenreliabilitäten konnte gezeigt werden, dass sich alle fünf Dimensionen des Beobachtungsinstruments eindimensional erfassen lassen und dass die Interrater-Reliabilität sowie interne Konsistenz für alle fünf Dimensionen moderat bis zufriedenstellend ausfallen ($ICC \geq .80$ und $\alpha \geq .69$ für alle Dimensionen). Eine Analyse der Korrelationen zwischen den Unterrichtsqualitäts-Dimensionen zeigt, dass die zwei fachspezifischen Dimensionen hohe Korrelationen zur kognitiven Aktivierung aufweisen und dass die stoffdidaktische Dimension darüber hinaus mittlere Korrelationen zur konstruktiven Unterstützung und auch zur Klassenführung aufweist. Diese Ergebnisse stehen im Einklang mit der Vermutung, dass insbesondere die kognitive Aktivierung eine fachspezifische Komponente enthält. Eine Einordnung der Basisdimensionen von Unterrichtsqualität entlang des Kontinuums von generisch bis fachspezifisch basierend auf den Ergebnissen würde die kognitive Aktivierung am weitesten in Richtung der fachspezifischen Seite anordnen, während Klassenführung auf der generischen Seite des Kontinuums liegt. Konstruktive Unterstützung würde sich dann in der Mitte zwischen Klassenführung und kognitiver Aktivierung befinden (siehe Abbildung 3).

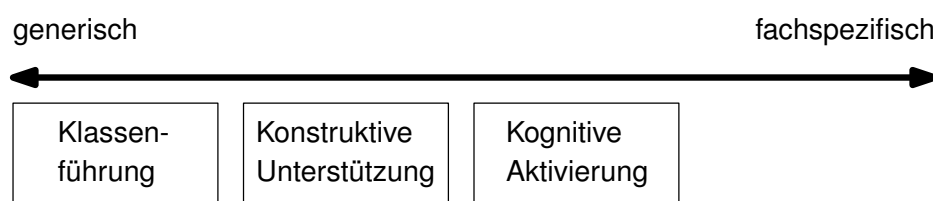


Abbildung 3: Einordnung der drei Basisdimensionen entlang des Kontinuums von generisch zu fachspezifisch anhand der empirischen Ergebnisse

9.3 Validierungsergebnisse

Die fünfte und sechste Forschungsfrage der vorliegenden Dissertation beziehen sich auf eine erste Validierung des Beobachtungsinstruments, indem einerseits Zusammenhänge zu den professionellen Kompetenzen der Lehrpersonen untersucht wurden und andererseits inhaltliche Vergleiche zu bestehenden fachspezifischen Beobachtungsinstrumenten vorgenommen wurden. Die zweite Publikation dieser Dissertation nimmt dabei die Zusammen-

hänge zu den Kompetenzen in den Fokus, während die dritte Publikation, die in *ZDM Mathematics Education* erschienen ist, einen Teil des Themenhefts bildet, in dem verschiedene Beobachtungsinstrumente miteinander verglichen werden. Auf Basis der Analysen dreier amerikanischer Unterrichtsstunden ist es somit möglich, fachspezifische Beobachtungsinstrumente inhaltlich miteinander zu vergleichen und darauf aufbauend Stärken und Schwächen des Beobachtungsinstruments zu thematisieren.

Die Ergebnisse der korrelativen Analysen zu den professionellen Kompetenzen der Lehrpersonen zeigen, dass die beiden fachspezifischen Qualitätsdimensionen Zusammenhänge mittlerer Effektstärke zu den proximalen, mathematikdidaktischen Fähigkeiten der Lehrpersonen aufweisen. Darüber hinaus zeigt die stoffdidaktische Qualitätsdimension eine schwache Korrelation zum mathematikdidaktischen Wissen der Lehrpersonen. Diese Ergebnisse können als erste kriteriale Validierung des Beobachtungsinstruments angesehen werden. Des Weiteren stützen sie die eingangs formulierte Definition fachspezifischer Qualitätsaspekte, die im Rahmen dieser Dissertation als Merkmale verstanden werden, deren Sicherstellung notwendigerweise ein gewisses Maß an Fachwissen oder fachdidaktischem Wissen der Lehrperson erfordern (siehe Kapitel 3.2).

Im Vergleich zu anderen Beobachtungsinstrumenten zeigt sich, dass das vorliegende Beobachtungsinstrument sowohl im fachübergreifenden Bereich als auch bei der Erfassung fachspezifischer Aspekte eine Reihe von inhaltlichen Überschneidungen zu anderen bestehenden Instrumenten aufweist. Eine genaue Analyse der Gemeinsamkeiten und Unterschiede aller Beobachtungsinstrumente, die Teil des ZDM-Themenheftes *„Studying Instructional Quality in Mathematics through Different Lenses: In Search of Common Ground“* bilden, findet sich in dem umfassenden Übersichtsbeitrag von Praetorius und Charalambous (2018). Für das vorliegende Beobachtungsinstrument zeigen sich bei der Erfassung der Klassenführung sowie der sozialen-emotionalen Unterstützung deutliche Überschneidungen zu den fachübergreifenden Instrumenten (z.B. CLASS, DMEE, Basisdimensionen) sowie zu den hybriden Instrumenten (z.B. TRU, UTOP). Das gleiche gilt für die Erfassung der kognitiven Aktivierung, wobei sich hier weiterhin auch deutliche inhaltliche Überschneidungen zu fachspezifischen Instrumenten (z.B. MQI, M-Scan) zeigen (bei der Beschreibung der Instrumentenentwicklung in Kapitel 5.2.1 wurden bereits detaillierte Vergleiche zu den drei Basisdimensionen dargestellt). Bei der Auswahl und Präsentation der Inhalte im Unterricht zeigen sich inhaltliche Überschneidungen zu den fachspezifischen und hybriden Instrumenten (MQI, M-Scan, TRU, UTOP). Wie bereits bei der Instrumentenentwicklung dargelegt, finden die Bereiche Leistungsüberprüfung sowie Adaptivität hingegen in dem vorliegenden Beobachtungsinstrument keine Berücksichtigung. Zusammenfassend kommen Praetorius und Charalambous (2018) beim Vergleich aller Beobachtungsinstrumente zu dem Ergebnis, dass insgesamt deutliche inhaltliche Überschneidungen bestehen:

Taken together, we can see that although many different frameworks exist,

they largely cover quite similar aspects, even across generic and content-specific frameworks. This leads to the question whether we indeed need so many frameworks or whether it would be wise to bundle our efforts and work based on a combined framework (Praetorius & Charalambous, 2018, S. 32).

Bei der Einschätzung der drei videographierten Unterrichtsstunden kommt unser Beobachtungsinstrument im Vergleich zu den anderen Beobachtungsinstrumenten in vielen Qualitätsbereichen zu ähnlichen Ergebnissen. Es zeigt sich jedoch gleichzeitig, dass zum Teil hinter gleichen Begrifflichkeiten bei einer genaueren Analyse trotzdem inhaltliche Unterschiede bestehen, die erst beim Bewerten konkreter Unterrichtsstunden deutlich werden. Ein Beispiel ist der Umgang mit Schülerfehlern, der im vorliegenden Beobachtungsinstrument unter einer fachspezifischen Perspektive bewertet wird, während andere Beobachtungsinstrumente darunter in erster Linie die Fehlertoleranz betrachten, aber keine inhaltlichen, fachspezifischen Einschätzungen vornehmen. Insgesamt zeigt sich, dass ein Vergleich anhand konkreter videographierter Unterrichtsstunden in zukünftigen Studien eine gute Möglichkeit bietet, um Gemeinsamkeiten und Unterschiede von Beobachtungsinstrumenten herauszuarbeiten und genauer zu analysieren.

10 Diskussion

Aufbauend auf den präsentierten Ergebnissen werden in dem vorliegenden Kapitel die Stärken und Schwächen des Beobachtungsinstruments genauer analysiert und es wird ein Vergleich zu bestehenden fachspezifischen Beobachtungsinstrumenten vorgenommen. Das Kapitel schließt mit einer Diskussion über mögliche Weiterentwicklungspotenziale des Beobachtungsinstruments und gibt einen Ausblick auf zukünftige Forschungsvorhaben in diesem Bereich.

10.1 Stärken und Schwächen des Beobachtungsinstruments

Das vorliegende Beobachtungsinstrument weist eine Reihe von Stärken auf, aber gleichzeitig auch einige Schwächen. Es wurde für die Verwendung bei in-vivo Unterrichtsbeobachtungen entwickelt und im Rahmen von TEDS-Unterricht hierzu auch erprobt. Da das Videographieren einerseits durch umfangreiche Genehmigungen zunehmend erschwert wird, bietet ein solches in-vivo Instrument die Möglichkeit, eine größere Anzahl an Unterrichtsstunden zu beobachten und zu bewerten. Die Items des Beobachtungsinstruments sind deshalb so entwickelt worden, dass eine mehrfache in-vivo Bewertung möglich ist. Darüber hinaus hat sich aber ebenfalls gezeigt, dass das Beobachtungsinstrument gleichermaßen auch für Videos anwendbar ist. Empirische Ergebnisse hinsichtlich des Einflusses von in-vivo Bewertungen gegenüber der Bewertung an Videos mit unserem Beobachtungsinstrument stehen jedoch noch aus. Hierzu wird TEDS-Video, eine Ergänzungsstudie zu TEDS-Validierung, die sowohl in-vivo Ratings durchführt als auch gleichzeitig Videos der beobachteten Unterrichtsstunden aufnimmt und anschließend erneut bewertet, einen Beitrag leisten können.

Das Beobachtungsinstrument erfasst einerseits die drei Basisdimensionen von Unterrichtsqualität und gleichzeitig fachspezifische Qualitätsaspekte des Mathematikunterrichts. Somit ist es möglich, Zusammenhänge zwischen fachspezifischen Aspekten und den drei Basisdimensionen genauer zu analysieren, wozu im Rahmen dieser Dissertation bereits erste Ergebnisse präsentiert wurden. Diese Zusammenhangsanalysen bieten damit eine erste Möglichkeit, die Frage nach der Fachspezifität der drei Basisdimensionen genauer zu untersuchen. Diese Kombination von generischen und fachspezifischen Aspekten innerhalb eines Beobachtungsinstruments birgt darüber hinaus den Vorteil, dass zur Bewertung und Analyse der Unterrichtsstunden nicht verschiedene Beobachtungsinstrumente kombiniert werden müssen, wie es bereits in anderen Studien durchgeführt wurde (Blazar et al., 2017). Detailliertere Analysen hinsichtlich der Zusammenhänge generischer und fachspezifischer Qualitätsaspekte im Mathematikunterricht stehen jedoch noch aus. Hier könnten Folgestudien mit anderen Stichproben weitere Ergebnisse liefern.

Die Unterteilung der Unterrichtsstunden in mehrere Bewertungseinheiten bietet die Möglichkeit, Veränderungen in der Qualität innerhalb einer Unterrichtsstunde abzubil-

den, und bietet damit ebenfalls das Potenzial für weitere Analysen. Diese Vorgehensweise ist im deutschsprachigen Raum für das Modell der drei Basisdimensionen bisher noch nicht etabliert. Darüber hinaus bietet dieses Vorgehen die Möglichkeit, Rater-Fehler zu reduzieren, die aufgrund langer Beobachtungszeiträume entstehen.

Der Fokus des vorliegenden Beobachtungsinstruments liegt im Gegensatz zu anderen Beobachtungsinstrumenten (z.B. TRU Math, Schoenfeld, 2013) in erster Linie auf der Lehrperson und ihren Handlungen im Unterricht. Dadurch werden die Schülerinnen und Schüler sowie ihre Lernprozesse nicht direkt fokussiert. Es soll jedoch hier angemerkt werden, dass eine Bewertung der Lehrperson oftmals nicht möglich ist, ohne ebenfalls die Schülerinnen und Schüler mit in den Blick zu nehmen. Eine Abgrenzung zwischen der Qualität der Handlungen der Lehrperson und den Aktivitäten der Schülerinnen und Schülern ist damit nicht immer möglich, sondern im Gegenteil bedingen sie sich gegenseitig. Eine Möglichkeit, den Fokus und die Perspektiven der Schülerinnen und Schüler detaillierter mit einzubeziehen, wäre, im Anschluss an die beobachteten Unterrichtsstunden einen Schülerfragebogen einzusetzen und die Ergebnisse dieser Befragung mit der Bewertung der Unterrichtsqualität durch die externen Beobachterinnen und Beobachter zu vergleichen und zu reflektieren.

Eine Stärke, aber gleichzeitig auch Schwäche des Beobachtungsinstruments ist die Tatsache, dass es in verschiedenen Mathematikstunden unabhängig vom Inhalt, Kontext und der Klassenstufe angewendet werden kann, da die Items unabhängig von Inhalt und Unterrichtssituation formuliert wurden. Es bleibt jedoch fraglich, inwieweit einzelne Qualitätsaspekte abhängig vom mathematischen Inhalt in der beobachteten Stunde und dem Unterrichtskontext sind. Beispielsweise ist es mit dieser Vorgehensweise nicht möglich, die inhaltliche Kohärenz der mathematischen Inhalte durch die Erfassung bestimmter Verstehenselemente (Drollinger-Vetter, 2011) zu analysieren. Hier könnten weitere Forschungsvorhaben ansetzen, die das Beobachtungsinstrument systematisch bei verschiedenen mathematischen Inhaltsbereichen einsetzen und somit analysieren, inwieweit bestimmte Qualitätsaspekte abhängig vom Inhalt und Kontext sind. Darüber hinaus könnte das vorliegende Beobachtungsinstrument für verschiedene Inhaltsbereiche weiterentwickelt und spezifiziert werden. Das gleiche gilt für verschiedene Unterrichtssituationen und Unterrichtskontexte. Auch hier wären weitere Analysen notwendig, um darauf aufbauend auch Unterrichtsqualitätsmodelle weiterzuentwickeln (Brunner, 2017; Praetorius, 2014).

Die Anwendung des Beobachtungsinstruments erfolgte in hohem Maße standardisiert. Das bedeutet, dass nur die Qualitätsaspekte erfasst und bewertet werden, die in den Items aufgeführt sind. Aspekte, die durch externe Beobachtungen nicht erfassbar sind, werden mit dem Beobachtungsinstrument nicht bewertet. Stattdessen liegt der Fokus des Beobachtungsinstruments nur auf bestimmten Aspekten von Unterrichtsqualität. In Kapitel 5.3 wurde bereits detaillierter dargestellt, welche Aspekte nicht berücksichtigt wurden. Das bedeutet gleichzeitig, dass das vorliegende Beobachtungsinstrument keine

vollständige „To-Do-Liste“ für guten Unterricht darstellt und so auch nicht verwendet werden kann (Praetorius & Charalambous, 2018; Steinweg, 2011).

Darüber hinaus bringt eine hoch-standardisierte Erfassung von Qualitätsmerkmalen gewisse Einschränkungen mit sich, die durch das quantitative Design bedingt sind. Eine Lehrperson, die im Unterricht in einem Item eine hohe Ausprägung erhält (beispielsweise im Bereich fachliche Erklärungen), zeigt bei diesem Aspekt auf Basis der Konzeption dieses Beobachtungsinstruments einen qualitativ hochwertigen Unterricht (sie erklärt nach unserem Verständnis fachlich gut). Inwieweit dieser Qualitätsaspekt einen Zusammenhang zu Schülerleistungsfortschritten aufweist, kann ebenfalls im Rahmen von TEDS-Unterricht quantitativ analysiert werden. Dies bedeutet jedoch nicht, dass diese Lehrperson für alle Schülerinnen und Schüler gleichermaßen gut sein muss. Beispielsweise kann eine fachliche Erklärung, die wir als gut bewerten, für manche Schülerinnen und Schüler nicht angemessen sein, während sie für bestimmte Schülerinnen und Schüler sehr gut passt, da die Lehrperson an bestimmten Stellen beispielsweise besonders langsam und präzise erklärt. Für solche detaillierteren Analysen ist eine Kombination mit qualitativen Methoden notwendig. Diese Schwäche der standardisierten Erhebung lässt sich auf weitere Items übertragen, deren Passung und Qualität notwendigerweise auch die beteiligten Schülerinnen und Schüler mit einbeziehen muss.

Das Beobachtungsinstrument wurde im Rahmen des ZDM-Themenheftes *„Studying Instructional Quality in Mathematics through Different Lenses: In Search of Common Ground“*, das von Charalambos Y. Charalambous und Anna-Katharina Praetorius herausgegeben wird, auf drei amerikanische Unterrichtsstunden angewendet. Es zeigte sich hierbei, dass eine Anwendung auch in anderen Länderkontexten möglich ist, es jedoch bestimmte Items gibt, die speziell im Kontext der deutschen Bildungsstandards entwickelt wurden. Hierzu zählen insbesondere die Items zu den allgemeinen mathematischen Kompetenzen, die sich nicht direkt auf andere Länderkontexte übertragen lassen, da weder die Common Core Standards noch die ursprünglichen NCTM-Standards direkt vergleichbare Kompetenzfacetten enthalten.

10.2 Weiterentwicklungspotenziale

Das Beobachtungsinstrument wurde im Rahmen dieser Dissertation detailliert beschrieben und auch hinsichtlich empirischer Gütekriterien analysiert. Darüber hinaus fand eine erste kriteriale und inhaltliche Validierung statt. Basierend auf diesen Ergebnissen bietet das Beobachtungsinstrument nun eine Reihe von Weiterentwicklungsmöglichkeiten.

Bei der theoretischen Entwicklung wurden zwei fachspezifische Qualitätsdimensionen herausgearbeitet und für das vorliegende Beobachtungsinstrument operationalisiert. Ein erstes Entwicklungspotenzial bezieht sich deshalb auf die Ausschärfung und Weiterentwicklung dieser zwei fachspezifischen Dimensionen. Eine Möglichkeit besteht darin, weitere fachspezifische Qualitätsaspekte zu identifizieren und damit eine Anreicherung der

vorliegenden Qualitätsdimensionen vorzunehmen. Insbesondere könnte hierfür genauer analysiert werden, bei welchen weiteren Lehrerhandlungen und in welchen Unterrichtssituationen sich einerseits die mathematische, stoffliche Qualität und andererseits die mathematikdidaktische Qualität im Unterricht zeigen, die bisher im vorliegenden Beobachtungsinstrument keine Berücksichtigung finden. Hierfür ist bereits die Ergänzungsstudie TEDS-Video begonnen worden, die mithilfe von Unterrichtsvideos eine genauere qualitative Analyse der fachspezifischen Unterrichtsqualität vornimmt.

Wie bereits beschrieben gab es auch gewisse Qualitätsaspekte, die sich aufgrund der externen Beobachtungssituation nicht beobachten und bewerten ließen. Hier wäre es wünschenswert, Möglichkeiten zu identifizieren, die diese Qualitätsaspekte ebenfalls im Unterricht untersuchen können. Die Ergänzungsstudie TEDS-Video könnte hier ebenfalls einen Beitrag leisten, indem zum Beispiel die videographierten Unterrichtsstunden zusammen mit der Lehrperson besprochen und reflektiert werden. Eine solche Vorgehensweise angereichert mit Methoden des nachträglichen lauten Denkens (Weidle & Wagner, 1994) bietet beispielsweise die Möglichkeit, die Adaptivität der Lehrerhandlungen genauer zu untersuchen. Diese Vorgehensweise hätte auch den Vorteil, die Kritik über die mangelnden Informationen der Beobachterinnen und Beobachter zu Klasse und Lernenden aufzunehmen und diese genauer zu untersuchen.

Die empirischen Ergebnisse des Beobachtungsinstruments im Rahmen von TEDS-Unterricht zeigen, dass manche Items Boden- und Deckeneffekte aufweisen. Die Gründe hierfür können einerseits in der Wahl der Stichprobe für TEDS-Unterricht liegen, die insgesamt als positiv selektiert angesehen werden kann, worauf die guten bis sehr guten Noten in den beiden Staatsexamina der beteiligten Lehrpersonen hinweisen. Dies erklärt beispielsweise die Deckeneffekte mancher Items in der Dimension Klassenführung. Andererseits zeigen manche Items deutliche Bodeneffekte, die sich nicht direkt mit der Stichprobenwahl erklären lassen. Hier wäre eine bessere Justierung der Items wünschenswert, damit sie besser zwischen den verschiedenen Lehrpersonen trennen können. Aus empirischen Gründen müssten solche schlecht trennenden Items aus den vorliegenden Dimensionen herausgenommen werden, da sie keine Varianz zwischen Lehrpersonen oder Unterrichtsstunden aufklären können. Aus inhaltlichen Gesichtspunkten heraus wäre es jedoch gut, diese Qualitätsaspekte ebenfalls weiter zu berücksichtigen. Beispielsweise bezieht sich diese Problematik auf die zwei Qualitätsaspekte „Schülerfeedback“ und „Metakognition“, die für guten Unterricht durchaus als sehr relevant angesehen werden können (Hattie, 2009; Helmke, 2012; Schneider & Artelt, 2010). Eine Weiterentwicklung und eine Verbesserung dieser Items erscheinen deshalb sinnvoll.

Im Rahmen dieser Dissertation wurden die Zusammenhänge zwischen den drei Basisdimensionen und den zwei fachspezifischen Qualitätsdimensionen genauer untersucht. Insbesondere zeigten sich hohe Korrelationen der zwei fachspezifischen Qualitätsdimensionen zur kognitiven Aktivierung, aber ebenfalls moderate Korrelationen zur konstruktiven

Unterstützung. Diese Zusammenhänge eröffnen weitere Entwicklungspotenziale, die sich auch auf die theoretischen Konstrukte von Unterrichtsqualität beziehen. Beispielsweise könnte eine theoretische Konzeptualisierung vorgenommen werden, bei der die fachspezifischen Qualitätsaspekte zum Teil in das Modell der drei Basisdimensionen integriert werden, so dass jeweils eine fachspezifische Anreicherung und Ausdifferenzierung stattfindet. Eine solche Weiterentwicklung böte auch die Möglichkeit, eine Hierarchisierung von Qualitätsaspekten vorzunehmen, indem innerhalb der Basisdimensionen verschiedene Ebenen fachspezifischer und fachübergreifender Art konzeptualisiert würden. Hierzu gibt es erste Überlegungen von Brunner (2017), in denen eine effiziente Klassenführung und eine fachliche Fundierung die Grundlage für wirkungsvollen Mathematikunterricht bilden. Hier wären weitere Studien notwendig, um darauf aufbauende strukturelle Analysen durchzuführen und auf Basis dieser Analysen eine Theorieentwicklung hinsichtlich der Qualität von Mathematikunterricht anzustreben.

Und schließlich wäre es aus ökonomischen Gründen heraus wünschenswert, ein Beobachtungsinstrument vorliegen zu haben, das mit möglichst wenig Items auskommt, um zu inhaltlich validen Aussagen hinsichtlich der Qualitätsdimensionen zu kommen. Eine Möglichkeit wäre, Itemparcels zu bilden, die Qualitätsaspekte miteinander vereinen und damit insgesamt die Itemanzahl zu reduzieren. Andererseits bietet diese Möglichkeit auch die Gefahr, die Inferenz der Items weiter zu erhöhen und damit die Anwendung zu erschweren und die Reliabilität des Beobachtungsinstruments zu verringern. Eine andere Möglichkeit wäre deshalb, Items inhaltlich zu streichen, die hinsichtlich empirischer Gütekriterien eine geringere Aussagekraft haben. Eine solche Entscheidung muss jedoch immer die Ziele des Forschungsprojekts mitberücksichtigen, damit die Validität des Beobachtungsinstruments nicht eingeschränkt wird.

10.3 Fazit und Ausblick

Die vorliegende Dissertation hat ein Beobachtungsinstrument zur Unterrichtsqualität im Mathematikunterricht vorgestellt und empirisch erprobt. Hierzu wurden zunächst Aspekte mathematikdidaktischer Unterrichtsqualität theoretisch konzeptualisiert und für das Beobachtungsinstrument anschließend operationalisiert. Des Weiteren wurden empirische Ergebnisse präsentiert, mithilfe derer erste Antworten hinsichtlich Reliabilität und Validität des Beobachtungsinstruments gegeben werden konnten. Hinsichtlich der Konzeptualisierung von mathematikdidaktischer Unterrichtsqualität unterbreitet die vorliegende Dissertation einen ersten Vorschlag, in dem neben den drei Basisdimensionen von Unterrichtsqualität weitere fachspezifische Merkmale identifiziert wurden, die als relevant für guten Mathematikunterricht angesehen werden.

Abschließend soll hier noch einmal angemerkt werden, dass die Stichprobe mit 37 Lehrpersonen eher klein war, so dass weitere Studien notwendig sind, um die präsentierten Ergebnisse auch in größeren Kontexten zu prüfen und zu replizieren. Darüber hinaus

war die Stichprobe aus gegebenen Gründen positiv selektiert. Auch hier wären weitere Analysen mit einer randomisiert gezogenen Stichprobe wünschenswert. Und schließlich sind weitere Analysen zur Stabilität und Generalisierbarkeit der Daten notwendig, um Aussagen treffen zu können, die über die Ergebnisse der vorliegenden Studie hinausgehen (Brennan, 2011; Hill, Charalambous & Kraft, 2012; Praetorius, 2014; Praetorius et al., 2014; Shavelson & Webb, 1991).

Das vorliegende Beobachtungsinstrument ist im Rahmen der Studie TEDS-Unterricht entwickelt worden mit dem Ziel, die professionelle Kompetenz von Mathematiklehrkräften der unteren Sekundarstufe und ihren über die Unterrichtsqualität vermittelten Einfluss auf den Leistungszuwachs von Schülerinnen und Schülern zu untersuchen. Diese Analysen stehen aktuell noch aus. Damit einher geht die Frage, inwieweit die prognostische Validität des Beobachtungsinstruments in Bezug auf Schülerleistungen nachweisbar ist. Hierzu werden die Analysen im Rahmen der Studie TEDS-Unterricht eine Antwort geben können.

Es wäre außerdem für zukünftige Studien interessant, das vorliegende Beobachtungsinstrument mit anderen (sowohl fachübergreifenden als auch fachspezifischen) Beobachtungsinstrumenten zu kombinieren, um auch quantitativ genauer zu untersuchen, welche Gemeinsamkeiten und Unterschiede bei den erfassten Qualitätsaspekten und fachübergreifenden sowie fachspezifischen Dimensionen identifizierbar sind. Das ZDM-Themenheft *„Studying Instructional Quality in Mathematics through Different Lenses: In Search of Common Ground“*, das von Charalambos Y. Charalambous und Anna-Katharina Praetorius herausgegeben wird, konnte auf qualitativer Ebene hier bereits erste Ansätze generieren.

Und schließlich wurden eine Reihe von Stärken und Schwächen des Beobachtungsinstruments herausgearbeitet und diskutiert sowie Weiterentwicklungspotenziale aufgezeigt. Es wäre wünschenswert, diese Möglichkeiten in zukünftigen Forschungsvorhaben aufzugreifen und das Beobachtungsinstrument entsprechend weiterzuentwickeln, um den aufgezeigten Schwächen zu begegnen und die Aussagekraft des Beobachtungsinstruments für zukünftige Studien zu erhöhen.

Literatur

- Aebli, H. (1985). Das operative Prinzip. *Mathematik lehren*, 11(1985), 4–6.
- Baruk, S. (1989). *Wie alt ist der Kapitän? Über den Irrtum in der Mathematik*. Basel: Birkhäuser Verlag.
- Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, 9(4), 469–520.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American educational research journal*, 47(1), 133–180.
- Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, 56(3), 205–213.
- Blazar, D., Braslow, D., Charalambous, C. Y. & Hill, H. C. (2017). Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment*, 22(2), 71–94.
- Blömeke, S. (2002). *Universität und Lehrerausbildung*. Bad Heilbrunn/Obb.: Julius Klinkhardt.
- Blömeke, S., Gustafsson, J.-E. & Shavelson, R. J. (2015). Beyond dichotomies. *Zeitschrift für Psychologie*, 223(1), 3–13.
- Blömeke, S., Kaiser, G. & Lehmann, R. (2010). *TEDS-M 2008 – Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster: Waxmann Verlag.
- Blömeke, S., König, J., Busse, A., Suhl, U., Benthien, J., Döhrmann, M. & Kaiser, G. (2014). Von der Lehrerausbildung in den Beruf: Fachbezogenes Wissen als Voraussetzung für Wahrnehmung, Interpretation und Handeln im Unterricht. *Zeitschrift für Erziehungswissenschaft*, 17(3), 509–542.
- Blum, W., Drüke-Noe, C., Hartung, R. & Köller, O. (2012). *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen*. Berlin: Cornelsen Scriptor.
- Blum, W., Vogel, S., Drüke-Noe, C. & Roppelt, A. (2015). *Bildungsstandards aktuell: Mathematik in der Sekundarstufe II*. Braunschweig: Diesterweg.
- Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1–21.
- Brophy, J. (2000). *Teaching*. Brüssel: International Academy of Education.
- Bruner, J. (1974). *Toward a theory of instruction*. Cambridge: Harvard University Press.
- Brunner, E. (2017). Qualität von Mathematikunterricht: Eine Frage der Perspektive. *Journal für Mathematik-Didaktik*. doi: 10.1007/s13138-017-0122-z
- Buchholtz, N., Kaiser, G. & Blömeke, S. (2014). Die Erhebung mathematikdidaktischen Wissens–Konzeptualisierung einer komplexen Domäne. *Journal für Mathematik-Didaktik*, 35(1), 101–128.
- Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K. & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching.

- Educational and Psychological Measurement*, 73(5), 757–783.
- Charalambous, C. Y. & Praetorius, A.-K. (2018). Studying mathematics instruction through different lenses: setting the ground for understanding instructional quality more comprehensively. *ZDM Mathematics Education*, 50(3).
- Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive?* Münster: Waxmann Verlag.
- Creemers, B. & Kyriakides, L. (2007). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. London, New York: Routledge.
- Danielson, C. (2011). *Enhancing professional practice: A framework for teaching (2nd edn)*. Alexandria: Association for Supervision and Curriculum Development.
- Deci, E. L. & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Boston: Springer.
- Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit: Fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht*. Münster: Waxmann Verlag.
- Drollinger-Vetter, B. & Lipowsky, F. (2006). Fachdidaktische Qualität der Theoriephasen. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie »Unterrichtsqualität, Lernverhalten und mathematisches Verständnis« (Teil 3: Videoanalysen)* (S. 189–205). Frankfurt am Main: GfP.
- Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures*. Dordrecht: Springer Netherlands.
- Hattie, J. (2009). *Visible learning. synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
- Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts* (Bd. 6). Seelze: Klett/Kallmeyer.
- Hiebert, J., Gallimore, R., Garnier, H. & Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study*. Washington: National Center for Education Statistics.
- Hiebert, J. & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Hrsg.), *Second handbook of research on mathematics teaching and learning* (S. 371–404). Charlotte, NC: Information Age.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L. & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4), 430–511.
- Hill, H. C., Charalambous, C. Y. & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56–64.
- Hill, H. C., Rowan, B. & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, 42(2), 371–406.

- Jordan, A., Krauss, S., Löwen, K., Blum, W., Neubrand, M., Brunner, M., ... Baumert, J. (2008). Aufgaben im COACTIV-Projekt: Zeugnisse des kognitiven Aktivierungspotentials im deutschen Mathematikunterricht. *Journal für Mathematik-Didaktik*, 29(2), 83–107.
- Kaiser, G. (1995). Realitätsbezüge im Mathematikunterricht - Ein Überblick über die aktuelle und historische Diskussion. In G. Grauman, T. Jahnke, G. Kaiser & J. Meyer (Hrsg.), *Materialien für einen realitätsbezogenen Mathematikunterricht, Bd. 2* (S. 66–84). Hildesheim: Franzbecker.
- Kaiser, G., Blömeke, S., König, J., Busse, A., Döhrmann, M. & Hoth, J. (2017). Professional competencies of (prospective) mathematics teachers – cognitive versus situated approaches. *Educational Studies in Mathematics*, 94(2), 161–182.
- Kaiser, G., Blum, W., Borromeo Ferri, R. & Greefrath, G. (2015). Anwendungen und Modellieren. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 357–383). Heidelberg: Springer Verlag.
- Kaiser, G., Busse, A., Hoth, J., König, J. & Blömeke, S. (2015). About the complexities of video-based assessments: Theoretical and methodological approaches to overcoming shortcomings of research on teachers' competence. *International Journal of Science and Mathematics Education*, 13(2), 369–387.
- Kane, T. J. & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains (research paper, met project)*. Bill & Melinda Gates Foundation.
- Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R. & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, 49(3), 568–589.
- Klafki, W. (1964). *Das pädagogische Problem des Elementaren und die Theorie der kategorialen Bildung*. Weinheim: Beltz.
- Klieme, E., Lipowsky, F., Rakoczy, K. & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (S. 127–146). Münster: Waxmann Verlag.
- Klieme, E., Pauli, C. & Reusser, K. (2009). The pythagoras study: Investigating effects of teaching and learning in swiss and german mathematics classrooms. In T. Janik & T. Seidel (Hrsg.), *The power of video studies in investigating teaching and learning in the classroom* (S. 137–160). Münster: Waxmann Verlag.
- Klieme, E. & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54(2), 222–237.
- Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur“ und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.), *TIMSS-Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 43–57). Bonn: Bundesministerium für Bildung und Forschung.

- König, J., Blömeke, S., Klein, P., Suhl, U., Busse, A. & Kaiser, G. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? a video-based assessment approach. *Teaching and Teacher Education*, 38, 76–88.
- König, J., Blömeke, S., Paine, L., Schmidt, W. H. & Hsieh, F.-J. (2014). Teacher education effectiveness: Quality and equity of future primary and future lower secondary teachers' general pedagogical knowledge. In S. Blömeke, F.-J. Hsieh, G. Kaiser & W. H. Schmidt (Hrsg.), *International perspectives on teacher knowledge, beliefs and opportunities to learn: Teds-m results* (S. 187–206). Dordrecht: Springer Netherlands.
- König, J., Doll, J., Buchholtz, N., Förster, S., Kaspar, K., Rühl, A.-M., ... Kaiser, G. (2017). Pädagogisches Wissen versus fachdidaktisches Wissen? Struktur des professionellen Wissens bei angehenden Deutsch-, Englisch- und Mathematiklehrkräften im Studium. *Zeitschrift für Erziehungswissenschaft*. doi: 10.1007/s11618-017-0765-z
- König, J. & Kramer, C. (2016). Teacher professional knowledge and classroom management: On the relation of general pedagogical knowledge (GPK) and classroom management expertise (CME). *ZDM Mathematics Education*, 48(1-2), 139–151.
- Kounin, J. S. (1970). *Discipline and group management in classrooms*. New York: Holt, Rinehart & Winston.
- Kunter, M. & Baumert, J. (2006). Who is the expert? construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9(3), 231–251.
- Kunter, M. & Ewald, S. (2016). Bedingungen und Effekte von Unterricht: Aktuelle Forschungsperspektiven aus der pädagogischen Psychologie. In N. McElvany, W. Bos, H. Holtappels, M. M. Gebauer & F. Schwabe (Hrsg.), *Bedingungen und Effekte guten Unterrichts* (S. 9–31). Münster: Waxmann Verlag.
- Kunter, M. & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. In M. Kunter, J. Baumert & W. Blum (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 85–113). Münster: Waxmann Verlag.
- Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14(1), 25–47.
- Liang, J. (2015). Live video classroom observation: an effective approach to reducing reactivity in collecting observational information for teacher professional development. *Journal of Education for Teaching*, 41(3), 235–253.
- Lipowsky, F. (2015). Unterricht. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (2. Auflage) (S. 73–101). Berlin, Heidelberg: Springer.
- Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E. & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the pythagorean theorem. *Learning and instruction*, 19(6), 527–537.
- Marder, M. & Walkington, C. (2014). Classroom observation and value added models give complementary information about quality of mathematics teaching. In T. Kane, K. Kerr & R. Pianta (Hrsg.), *Designing teacher evaluation systems: New guidance from the measuring*

- effective teaching project* (S. 234–277). New York: Wiley.
- Matsumura, L. C., Garnier, H., Slater, S. C. & Boston, M. D. (2008). Toward measuring instructional interactions 'at-scale'. *Educational Assessment*, 13(4), 267–300.
- Pianta, R. C. & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational researcher*, 38(2), 109–119.
- Pietsch, M. & Tosana, S. (2008). Beurteilereffekte bei der Messung von Unterrichtsqualität. *Zeitschrift für Erziehungswissenschaft*, 11(3), 430–452.
- Praetorius, A.-K. (2014). *Messung von Unterrichtsqualität durch Ratings*. Münster: Waxmann Verlag.
- Praetorius, A.-K. & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM Mathematics Education*, 50(3).
- Praetorius, A.-K., Klieme, E., Herbert, B. & Pinger, P. (2018). Generic dimensions of teaching quality: the german framework of three basic dimensions. *ZDM Mathematics Education*, 50(3).
- Praetorius, A.-K., Lenske, G. & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and instruction*, 22(6), 387–400.
- Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K. & Klieme, E. (2014). One lesson is all you need? stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12.
- Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht*. Münster: Waxmann Verlag.
- Rakoczy, K. & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie »Unterrichtsqualität, Lernverhalten und mathematisches Verständnis« (Teil 3: Videoanalysen)* (S. 189–205). Frankfurt am Main: GfPF.
- Reinmann, G. & Mandl, H. (2006). Unterrichten und Lernumgebungen gestalten. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie* (S. 613–658). Weinheim: Beltz.
- Reiss, K. & Hammer, C. (2013). *Grundlagen der Mathematikdidaktik. Eine Einführung für den Unterricht in der Sekundarstufe*. Basel: Birkhäuser.
- Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R. & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School science and mathematics*, 102(6), 245–253.
- Schilcher, A., Krauss, S., Rincke, K. & Hilbert, S. (2017). Ausblick–Aus FALKE wird FALKE. In S. Krauss, A. Lindl, A. Schilcher, M. Fricke, A. Göhring & B. Hofmann (Hrsg.), *Falko: Fachspezifische Lehrerkompetenzen: Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik* (S. 439–452). Münster: Waxmann Verlag.
- Schlesinger, L. & Jentsch, A. (2016). Theoretical and methodological challenges in measuring

- instructional quality in mathematics education using classroom observations. *ZDM Mathematics Education*, 48(1-2), 29–40.
- Schneider, W. & Artelt, C. (2010). Metacognition and mathematics education. *ZDM Mathematics Education*, 42(2), 149–161.
- Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM Mathematics Education*, 45(4), 607–621.
- Seidel, T. & Prenzel, M. (2010). Beobachtungsverfahren. In H. Holling & B. Schmitz (Hrsg.), *Handbuch der psychologischen Methoden und Evaluation* (S. 139–152). Göttingen: Hogrefe.
- Sekretariat der ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). (2003). *Bildungsstandards im Fach Mathematik für den mittleren Schulabschluss. Beschluss vom 4.12.2003*. München: Luchterhand.
- Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks: Sage.
- Sherin, M., Jacobs, V. & Philipp, R. (2011). *Mathematics teacher noticing: Seeing through teachers' eyes*. Routledge: Routledge.
- Shuell, T. J. (1993). Toward an integrated theory of teaching and learning. *Educational Psychologist*, 28(4), 291–311.
- Shulman, L. S. (1985). Paradigms and research programs in the study of teaching. teoksessa mc wittrock (toim.). In M. C. Wittrock (Hrsg.), *Handbook of research on teaching* (S. 3–36). New York: Macmillan.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational researcher*, 15(2), 4–14.
- Steinweg, A. S. (2011). Einschätzung der Qualität von Lehr-Lernsituationen im mathematischen Anfangsunterricht—ein Vorschlag. *Journal für Mathematik-Didaktik*, 32(1), 1–26.
- Steuer, G. (2014). *Fehlerklima in der Klasse: Zum Umgang mit Fehlern im Mathematikunterricht*. Wiesbaden: Springer-Verlag.
- Taut, S. & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction*, 46, 45–60.
- Thompson, C. J. & Davis, S. B. (2014). Classroom observation data and instruction in primary mathematics education: improving design and rigour. *Mathematics Education Research Journal*, 26(2), 301–323.
- Van de Pol, J., Volman, M. & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, 22(3), 271–296.
- Van Es, E. A. & Sherin, M. G. (2002). Learning to notice: Scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher Education*, 10(4), 571–596.
- Vinner, S. (2007). Mathematics education: Procedures, rituals and man's search for meaning. *The Journal of Mathematical Behavior*, 26(1), 1–10.
- Vom Hofe, R. (1995). *Grundvorstellungen mathematischer Inhalte*. Heidelberg: Spektrum.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*.

- Cambridge: Harvard university press.
- Wagenschein, M. (1999). *Verstehen lehren: genetisch, sokratisch, exemplarisch*. Weinheim: Beltz.
- Waldis, M., Grob, U., Pauli, C. & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile. In K. Reusser, C. Pauli & M. Waldis (Hrsg.), *Unterrichtsgestaltung und Unterrichtsqualität–Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (S. 171–208). Münster: Waxmann.
- Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E. & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics*, 85(1), 109–128.
- Weidle, R. & Wagner, A. C. (1994). Die Methode des Lauten Denkens. In G. L. Huber & H. Mandl (Hrsg.), *Verbale Daten. Eine Einführung in die Grundlagen und Methoden der Erhebung und Auswertung* (Bd. 2, S. 81–103). Weinheim, Basel: Beltz.
- Winter, H. (2015). *Entdeckendes Lernen im Mathematikunterricht. Einblicke in die Ideengeschichte und ihre Bedeutung für die Pädagogik (3. Auflage)*. Heidelberg: Springer.
- Wittmann, E. C. & Müller, G. N. (1990). *Handbuch produktiver Rechenübungen*. Stuttgart: Klett.
- Wood, D., Bruner, J. S. & Ross, G. (1976). The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2), 89–100.
- Wynands, A. (2012). Intelligentes üben. In W. Blum, C. Drüke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen* (S. 113–125). Berlin: Cornelsen Scriptor.
- Zech, F. (1998). *Grundkurs Mathematikdidaktik: theoretische und praktische Anleitungen für das Lehren und Lernen von Mathematik*. Weinheim: Beltz.

A Zusammenfassung

Die vorliegende Dissertation befasst sich mit der Entwicklung und Erprobung eines standardisierten Beobachtungsinstruments zur Erfassung mathematikdidaktischer Unterrichtsqualität. Diesbezüglich werden zunächst Aspekte in Form zweier Subdimensionen (stoffdidaktisch und unterrichtsdidaktisch) theoretisch konzeptualisiert und es werden sowohl theoretische als auch methodologische Herausforderungen bei der Erfassung mathematikdidaktischer Unterrichtsqualität mit externen Beobachterinnen und Beobachtern erörtert. Anschließend werden empirische Ergebnisse präsentiert, mit denen erste Aussagen hinsichtlich Reliabilität und Validität des Beobachtungsinstruments getroffen werden können. Es wird einerseits untersucht, ob die Qualität des Mathematikunterrichts mit mehreren theoretisch entwickelten Dimensionen reliabel erfasst werden kann und welche Zusammenhänge zwischen den etablierten drei Basisdimensionen von Unterrichtsqualität (Klassenführung, konstruktive Unterstützung und kognitive Aktivierung) und den zwei neu entwickelten fachspezifischen Qualitätsdimensionen existieren. Andererseits werden Zusammenhänge zu den professionellen Kompetenzen der Lehrpersonen untersucht und es werden inhaltliche Vergleiche zu bestehenden fachspezifischen Beobachtungsinstrumenten vorgenommen, um darauf aufbauend Stärken und Schwächen sowie Weiterentwicklungspotenziale des Beobachtungsinstruments zu identifizieren. Insgesamt werden hierzu in der vorliegenden Dissertation drei Zeitschriftenartikel präsentiert. Die Verfassung der Dissertation sowie die Entwicklung des Beobachtungsinstruments erfolgte im Rahmen des Projekts TEDS-Unterricht, einem Nachfolgeprojekt der Studien TEDS-M (*Teacher Education and Development Study: Learning to Teach Mathematics*) und TEDS-FU (*Follow-Up*). Die Erhebung der Unterrichtsqualität erfolgte hierfür in-vivo mit hoch-inferenten Einschätzungen von externen Beobachterinnen und Beobachtern.

Mithilfe einer systematischen Literaturrecherche und der darauf aufbauenden theoretischen Analyse bisheriger standardisierter, fachspezifischer Beobachtungsinstrumente konnten zwei Perspektiven auf fachspezifische Qualitätsaspekte im Mathematikunterricht herausgearbeitet werden. Einerseits konnte eine *mathematische, stoffbezogene Perspektive* identifiziert werden, die sich auf die fachliche Tiefe, Korrektheit und Präzision der mathematischen Inhalte bezieht. Andererseits konnte eine *unterrichtsbezogene, fachdidaktische Perspektive* herausgearbeitet werden, die sich auf die didaktisch-methodische Umsetzung und Darstellung der mathematischen Inhalte im Unterricht bezieht. Eine solche zweigeteilte Konzeptualisierung kann in Einklang gebracht werden mit neueren Konzeptualisierungen des fachdidaktischen Wissens von Mathematiklehrkräften.

Die empirischen Ergebnisse zeigen, dass die fünf Dimensionen des Beobachtungsinstruments im Rahmen von TEDS-Unterricht jeweils eindimensional mit zufriedenstellenden Interrater-Reliabilitäten erfasst werden können. Die zwei neu entwickelten fachspezifischen Qualitätsdimensionen zeigen dabei hohe Korrelationen zur kognitiven Aktivierung. Darüber hinaus weist die stoffdidaktische Qualitätsdimension mittlere Korrelationen zur

konstruktiven Unterstützung und auch zur Klassenführung auf. Im Vergleich zu anderen (fachspezifischen) Beobachtungsinstrumenten zeigt sich eine Reihe von inhaltlichen Gemeinsamkeiten, aber auch Unterschieden sowohl im fachübergreifenden als auch fachspezifischen Bereich, die auch bei der Analyse von Mathematikstunden aus dem amerikanischen Raum deutlich werden.

Insgesamt zeigt das entwickelte Beobachtungsinstrument eine Reihe von Stärken im Bereich der Konzeption und Methode, aber auch Schwächen, die bei der Erprobung deutlich wurden, so dass es Weiterentwicklungspotenziale gibt, um den aufgezeigten Schwächen zu begegnen und die Aussagekraft des Beobachtungsinstruments für zukünftige Studien zu erhöhen.

B English summary

The dissertation deals with the development and testing of a standardized observational instrument for measuring instructional quality in mathematics education. In this regard, characteristics in the form of two mathematics educational subdimensions (subject-related and teaching-related) are theoretically conceptualized. In addition, the dissertation discusses theoretical as well as methodological challenges in measuring mathematics educational quality of instruction with external observers. Subsequently, empirical results are presented, with which first statements regarding reliability and validity of the observational instrument can be made. On the one hand, the dissertation examines, whether the quality of mathematics teaching with several theoretically developed dimensions can be reliably measured and if correlations exist between the established three basic dimensions of teaching quality (classroom management, student support and cognitive activation) and the two newly developed subject-specific quality dimensions. On the other hand, the dissertation examines correlations to the professional competences of mathematics teachers and presents comparisons to established subject-specific observational instruments to identify strengths, weaknesses and potentials for further development of the observational instrument. In total, this dissertation presents three articles in peer-reviewed journals. The dissertation and the development of the observational instrument were drafted as part of the TEDS-Instruct project, a follow-up study to the TEDS-M (*Teacher Education and Development Study: Learning to Teach Mathematics*) and TEDS-FU (*Follow-Up*) studies. The quality of teaching is assessed in vivo with high-inference ratings by external observers.

Based on a systematic literature review and a theoretical analysis of established, standardized, subject-specific observational instruments, two perspectives on subject-specific quality aspects in mathematics education could be distinguished. On the one hand, a *mathematical, subject-related perspective* could be identified, which refers to the mathematical depth, correctness and precision of the mathematical content. On the other hand, a *teaching-related, educational perspective* could be identified, which relates to the implementation and presentation of the mathematical content in the classroom. Such a two-fold conceptualization can be reconciled with recent conceptualizations of mathematics teachers' pedagogical content knowledge.

The empirical results show that the five dimensions of the observational instrument can be measured one-dimensionally with satisfactory interrater-agreement. The two newly developed subject-specific quality dimensions show high correlations to cognitive activation. In addition, the subject-related quality dimension has medium correlations to student support and to classroom management. Compared to other (subject-specific) observational instruments, there are several similarities but also differences in both the generic and subject-specific dimensions, which became also evident in the analysis of three American mathematics lessons.

Overall, the developed observational instrument shows a few strengths, but also weaknesses that became apparent during practice, so that there is potential for further development to address the identified weaknesses and to enhance the explanatory power of the observational instrument for future studies.

C Beobachtungsinstrument TEDS-Unterricht

Schule: _____

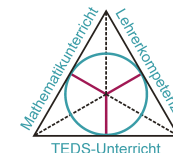
Klasse: _____

Lehrkraft: _____

Datum: _____ Zeit: _____ ☐ 1. Termin

Rater: _____ Anzahl SuS: _____ ☐ 2. Termin

Beobachtungsinstrument TEDS-Unterricht



D1: Klassenführung (6 Items)	Indikatoren	MZP 1	MZP 2	MZP 3	MZP 4
Effektive Lernzeitnutzung D11001	- Der Unterricht beginnt und endet pünktlich - Die Übergänge zwischen den Phasen erfolgen reibungslos - Die Unterrichtszeit wird für unterrichtsbezogene Inhalte genutzt				
Sinnvolle Routinen und Regeln D11002	- Eingeübte Muster für die Lernorganisation sind ersichtlich - Die SuS übernehmen Organisationsaufgaben				
Störungspräventives Verhalten D11003	- Die LK unterbindet aufkommende Störungen sofort und zielsicher - Die LK hat den Überblick über das Unterrichtsgeschehen				
Transparenz des Lernprozesses D11004	- Die LK informiert die SuS über Lernziele / Unterrichtsablauf - Die LK macht ihre Erwartungen transparent - Arbeitsaufträge sind präzise formuliert				
Arbeitsatmosphäre D11005	- Die Lautstärke ist angemessen - Die SuS reagieren auf Signale und Hinweise - Die SuS und die LK unterbrechen sich nicht gegenseitig				
Strukturierung des Unterrichts D11007	- Der Unterricht besitzt einen „roten Faden“ - Der Unterricht ist in Phasen unterteilt - Die LK findet einen angemessenen inhaltlichen Abschluss der Stunde				
D2: Konstruktive Unterstützung (7 Items)	Indikatoren	MZP 1	MZP 2	MZP 3	MZP 4
Individuelle Förderung D21001	- Die LK informiert sich über individuelle Schwierigkeiten / Fortschritte - Die LK nimmt sich Zeit für einzelne SuS - Die LK gibt den SuS individuelle Hilfestellungen				
Umgang mit Heterogenität/Differenzierung D21002	- Es gibt zusätzliches Material für Teile der Lerngruppe - Es gibt anforderungsbezogene, differenzierende Aufgabenstellungen - Es findet ad-hoc Differenzierung statt (z. B. leichtere Fragestellung)				
Selbstgesteuertes Lernen D21003	- Die SuS überprüfen ihre Ergebnisse selbstständig an einer Musterlösung - Die LK fordert eigenständiges Arbeiten in hohem Maße ein - Die LK ermöglicht den SuS die Wahl von Sozialformen				
Rückmeldungen der LK D21006	- Die Rückmeldungen der LK sind differenziert - Die Rückmeldungen der LK sind sachlich und konstruktiv - Die Rückmeldungen der LK sind zukunftsgerichtet				
Wertschätzung D21007	- Die LK hat Geduld - Die LK verstärkt positiv durch Lob - Die LK ermutigt die SuS				

Feedback der SuS D21008	- Die LK holt sich ein Feedback ein - Die LK geht auf Anregungen ein - Die LK redet mit den SuS über die Klasse				
Förderung von Kooperation D21009	- Die LK initiiert gemeinsame Lernprozesse zwischen den SuS - Die LK stellt Aufgaben, die Einigung erfordern - Die LK nimmt die Rolle eines Mediators ein - Die SuS helfen sich gegenseitig				
D3: Kognitive Aktivierung (6 Items)	Indikatoren	MZP 1	MZP 2	MZP 3	MZP 4
Herausfordernde Fragen und Aufgaben D31001	- Die LK zeigt ein angemessenes Frageverhalten (nicht zu kleinschrittig) - Die LK stellt kognitiv herausfordernde Aufgaben - Die LK gibt geeignete Impulse, z.B. ein Problem, offene Frage				
Förderung von Metakognition D31002	- Es findet mindestens ein metakognitiver Teilprozess statt - Die LK gibt Zeit für metakognitive Prozesse - Methodische Vorgehensweisen werden reflektiert				
Aktivierung des Wissens D31003	- Die LK fragt nach den Vorstellungen der SuS zum Thema - Die SuS erklären das Thema in eigenen Worten - Die LK aktiviert das bei SuS vorhandene Wissen				
Evolutionärer Umgang (Ko-Konstruktion) D31004	- Die Fachbegriffe werden aus den Vorstellungen der SuS entwickelt - Die LK nutzt Fragetechniken zur Exploration und Weiterentwicklung - Es findet eine gemeinsame Konstruktion neuen Wissens statt				
Qualität der Methoden D31005	- Es werden kognitiv aktivierende Methoden verwendet - Die LK gibt angemessene Zeit zum Nachdenken im Plenum - Die Methodenwahl ist dem Inhalt und den SuS angemessen				
Wissenssicherung D31006	- Die LK verwendet Erinnerungshilfen oder Beispiele - Es kommen Wiederholungen zur Vernetzung vor - Relevante Zwischenschritte werden besprochen				
D4: Mathematikdidaktische Unterrichtsqualität (14 Items)	Indikatoren	MZP 1	MZP 2	MZP 3	MZP 4
Konstruktiver Umgang mit Fehlern D41001	- Die LK nutzt Fehler als Lernanlass - Die LK analysiert Schülerfehler und fehlerhafte Denkweisen - Es herrscht eine Fehlertoleranz in Lernsituationen - Fehler werden von den SuS selbst korrigiert				
Fachliche Korrektheit der LK D41002	- Die LK macht keine inhaltlichen oder formalen Fehler - Die LK ist bei sprachlichen und schriftlichen Formulierungen präzise				
Darstellungsformen D41004	- Es werden verschiedene Darstellungsformen verwendet - Die LK verdeutlicht Beziehungen zwischen Darstellungsformen				

Intelligentes Üben D4I005	- Die LK erklärt, wozu die Übungen dienen - Die Übungen enthalten entdeckende und reflexive Aspekte - Die Übungen sind selbstdifferenzierend				
Geeignete Beispiele D4I006	- Die Beispiele deuten auf das Wesentliche hin - Die LK verwendet eine ausreichende Anzahl an Beispielen - Es werden Beispiele aus der Lebenswelt der SuS verwendet				
Förderung von Fachsprache D4I007	- Es wird Zeit für die Begriffsbildung genutzt - Die LK korrigiert sprachliche Fehler adaptiv - Die LK initiiert die angemessene Verwendung von Fachsprache				
Modellieren D4I008	- Die LK fördert Übersetzungsprozesse zwischen RS und MMⁱ - Die LK fördert die Validierung der Ergebnisse im Hinblick auf die RS - Es finden außermathematische Vernetzungen und fachübergreifende Zusammenhänge statt				
Problemlösen und Entdecken D4I009	- Die LK fördert die Verwendung heuristischer Strategien - Die LK stellt mathematische Probleme - Die LK fördert die mathematische Entdeckung neuer Inhalte				
Begründen und Argumentieren D4I010	- Die LK ermöglicht mathematische Argumentationen - Die LK fordert explizit mathematische Begründungen ein - Die LK thematisiert Beweise formaler oder präformaler Art				
Mathematische Fertigkeiten D4I011	- Es findet ein Training mit mathematischen Symbolen statt - Die LK ermöglicht das Training mathematischer Basistechniken				
Mathematische Werkzeuge D4I012	- Die SuS setzen mathematische Werkzeuge sinnvoll ein - Die LK verwendet mathematische Werkzeuge korrekt				
Relevanz D4I013	- Es werden Bezüge zur Lebenswelt der SuS aufgezeigt - Es wird thematisiert, wozu die Inhalte gut sind - SuS können eigene Erfahrungen und Interessen einbringen				
Lehrervortrag (Erklärungen) D4I014	- Die LK erklärt besonders bei schwierigen Stellen langsam - Die LK beschränkt sich auf das Wesentliche - Die LK erklärt adressatengerecht - Die Erklärungen der LK sind klar strukturiert und präzise				
Fachliche Tiefe D4I015	- Es finden Verallgemeinerungen statt - Es finden innermathematische Vernetzungen statt - Mathematische Inhalte werden strukturiert und begrifflich verdichtet				

ⁱ RS: Realsituation, MM: Mathematisches Modell

D Kodiermanual

Rating-Manual Unterrichtsbeobachtungen TEDS-Unterricht

Das vorliegende Rating-Manual wurde im Rahmen der Studie ***TEDS-Unterricht*** entwickelt. TEDS-Unterricht ist ein Nachfolgeprojekt der Studien TEDS-M (Teacher Education and Development Study: Learning to Teach Mathematics) (Blömeke, Kaiser & Lehmann, 2010) und TEDS-FU (Follow Up) (z. B. Blömeke et al., 2014) mit dem Ziel, die professionelle Kompetenz von Mathematiklehrkräften der Sekundarstufe I und ihres über die Unterrichtsqualität vermittelten Einflusses auf den Leistungszuwachs von Schülerinnen und Schülern empirisch zu untersuchen. Dazu wurde Anfang 2015 in Hamburg mit der Datenerhebung begonnen.

Basierend auf den drei ***Basisdimensionen von Unterrichtsqualität*** und weiteren ***fachspezifischen Qualitätsaspekten*** wurde ein Beobachtungsinstrument zur Erhebung der Unterrichtsqualität entwickelt, dass für Ratings bei Beobachtungen in vivo, d.h. ohne Videographie, umsetzbar ist. Zudem soll das Instrument unabhängig vom mathematischen Inhaltsbereich in verschiedenen Klassenstufen einsetzbar sein. Das Beobachtungsinstrument knüpft dabei an vorliegende Beobachtungsinstrumente an, u.a. an das der Pythagoras-Studie (Klieme, Pauli & Reusser, 2006) und das MQI-Beobachtungsinstrument (Learning Mathematics for Teaching Project, 2011).

Insgesamt erfordert das ***Bewerten der Items*** jeweils eine hohe Interpretationsleistung von den Beobachtern, da für die Bewertung mehrere Aspekte gleichermaßen berücksichtigt werden müssen. In die Bewertung fließt ein

- wie häufig ein bestimmtes Verhalten oder Merkmal zwischen zwei Messzeitpunkten auftritt,
- wie intensiv ein bestimmtes Verhalten oder Merkmal auftritt,
- wie viel Prozent der Schülerinnen und Schüler der Klasse an einem bestimmten Verhalten teilhaben,
- ob ein bestimmtes Merkmal allen Schülerinnen und Schülern grundsätzlich zur Verfügung steht, auch wenn es *nicht* von allen genutzt wird (Fokus auf das Angebot der Lehrperson, vgl. Helmke, 2012).

Grundsätzlich gilt für die Bewertung der einzelnen Items, dass aus diesen vier Aspekten ein Gesamturteil gebildet wird (Klieme et al., 2006). Ein ggf. davon abweichendes Rating, das auf bestimmte oder andere Aspekte fokussiert, wird bei der Item-Beschreibung und den entsprechenden Kodierhinweisen genauer erläutert.

Neben den Hinweisen zur Bewertung der einzelnen Items, die in dem vorliegenden Kodiermanual beschrieben werden, gibt es ***übergreifende Rating-Hinweise***, die für alle Items gleichermaßen gelten. Es ist grundsätzlich für jedes Item möglich, ein „nicht beobachtbar“ zu vergeben, wenn zwischen zwei Messzeitpunkten keine Bewertung möglich ist. Es muss dabei jedoch immer unterschieden werden, ob man zu diesem

Aspekt nichts beobachten konnte (so dass keine Bewertung erfolgen kann, da man es nicht weiß), oder ob man beobachten konnte, dass zu dem Aspekt nichts passiert ist (beispielsweise wurde „nicht differenziert“ oder es wurde sich „kein Feedback von den Schülerinnen und Schülern eingeholt“). Im letzteren Fall erfolgt keine Vergabe von „nicht beobachtbar“, sondern analog zur Beschreibung im Kodiermanual die Vergabe einer „1=trifft überhaupt nicht zu“.

Da das Bewerten des Unterrichts fachliches und fachdidaktisches Wissen voraussetzt, ist für das Rating ein mathematikdidaktischer Hintergrund erforderlich (beispielsweise Mathematik-Lehramtsstudierende im Masterstudium). Es findet im Vorfeld ein umfangreiches **Training** der Rater statt, in dem Items und Indikatoren mithilfe des entwickelten Kodiermanuals ausführlich besprochen und anhand von Unterrichtsvideos in ihrer Anwendung erprobt werden. Die Unterrichtsstunden werden dabei in vivo, also unmittelbar im Unterricht mit dem Beobachtungsinstrument bewertet. Innerhalb einer 90-minütigen Doppelstunde wird die Unterrichtsqualität viermal von zwei Beobachtern unabhängig voneinander eingeschätzt, die **Analyseeinheit** beträgt also etwa 22,5 Minuten. Abhängig von der Phasierung des Unterrichts sind kleine Abweichungen davon möglich. Die Beobachter einigen sich ggf. per Handzeichen, wann sie mit dem Rating beginnen. Dieses wird immer gleichzeitig von beiden Ratern vorgenommen. Die Bewertung erfolgt auf einer **vierstufigen Skala** (1=„trifft überhaupt nicht zu“; 4=„trifft vollständig zu“). Neben den Ratings soll in jeder Unterrichtsstunde ein Protokoll angefertigt werden, in dem mindestens Informationen zu den Inhalten und der Phasierung des Unterrichts, den Sozialformen, Medien und Unterrichtsmethoden enthalten sind. Zum anderen sollten die Beobachter die Lehrperson darum bitten, ihnen die verwendeten Aufgaben auszuhändigen (z.B. eine Kopie des Arbeitsblattes).

Die Beobachter verhalten sich **während der Unterrichtsstunde** möglichst unauffällig und sitzen nach Möglichkeit hinten im Klassenraum. Insbesondere gehen sie auf Fragen von Schülerinnen und Schülern nicht ein. Um auch außerhalb von Plenumsphasen einen guten Überblick über das Lehrerhandeln zu behalten, kann die **Stimme der Lehrperson** über Mobiltelefon und Kopfhörer zu den Beobachtern übertragen werden, wenn die Lehrperson dazu ihre Erlaubnis erteilt hat. So können auch Einzelgespräche der Lehrperson mit Schülerinnen und Schülern in die Bewertung der Unterrichtsqualität einbezogen werden.

In einigen Fällen gibt es **besondere Situationen** im Klassenraum, die eine Bewertung einzelner Items erschweren. Hierzu zählt beispielsweise die Anwesenheit eines Sonderpädagogen im Klassenraum, der ggf. für einzelne Schülerinnen und Schüler eine ähnliche Rolle wie die Lehrperson einnehmen kann. Hier gilt es, für die Bewertung der Items zu entscheiden, inwieweit die Verantwortung des Unterrichtsgeschehens bei der Lehrperson liegt. Trägt die Lehrperson für das Verhalten des Sonderpädagogen im Unterricht maßgeblich die Verantwortung?

Nach jeder Doppelstunde findet eine ausführliche **Nachbesprechung** zwischen den beiden Ratern statt, bei der die Stunde im Hinblick auf systematische Fehleinschätzungen reflektiert wird. Im Rahmen dieser Nachbesprechung können einzelne Ratings auf Basis eines gemeinsamen theoretischen Verständnisses der Rating-Dimensionen konsensuell korrigiert werden, falls in der Diskussion grobe Missdeutungen festgestellt werden (z.B. König, 2015; Rakoczy & Pauli, 2006). Ziel ist es dabei nicht, Einigkeit über die gewählte Rating-Kategorie zu erreichen, sondern nur einen Konsens über das beobachtete Verhalten der Lehrperson zu finden. Dies ist notwendig, weil durch fehlende Videoaufzeichnungen im Anschluss an die Unterrichtsbeobachtungen keine „objektive“ Klärung der Tatsachen möglich ist.

Der **Aufbau des Rating-Manuals** sieht wie folgt aus. Nach einer umfassenden Beschreibung der jeweiligen Qualitätsdimension wird jedes dazugehörige Item zunächst inhaltlich beschrieben und die Grundidee des Items dargestellt. Darüber hinaus wird jedes Item durch einige Indikatoren präzisiert, die im Unterricht beobachtet und bewertet werden können. Bei diesen Indikatoren handelt es sich um typische Beispiele, deren Sichtbarkeit im Unterricht insgesamt aber weder hinreichend noch notwendig ist. Dies liegt darin begründet, dass als Erhebungsmethode hoch inferente Ratings zum Einsatz kommen, die immer eine Interpretation der Beobachter erfordern (Rakoczy & Pauli, 2006). Aus diesem Grund enthält das Kodiermanual neben der Item-Beschreibung und den Indikatoren auch Ankerbeispiele und weitere Bewertungshinweise für jedes Item.

Dimension 1: Klassenführung

Eine effektive Klassenführung durch die Lehrperson stellt eine wichtige Komponente für erfolgreiches Lernen von Schülerinnen und Schülern dar. Es handelt sich hierbei um eine Dimension der Tiefenstruktur von Unterricht, in der verschiedene Merkmale zusammengefasst werden. Die Grundlage ist dabei eine möglichst störungsfreie und optimale Nutzung der zur Verfügung stehenden Lernzeit. Bereits zu Zeiten behavioristischer Unterrichtsforschung konnten Disziplinprobleme als prädiktiv für erfolglosen Unterricht nachgewiesen werden, gerade weil hierdurch ein gewisser Teil der Lernzeit ungenutzt bleibt (Helmke, 2012).

Insbesondere konnte durch die maßgebliche Arbeit von Kounin (1976) ein empirischer Zusammenhang zwischen Disziplin im Unterricht und störungspräventivem Handeln der Lehrperson festgestellt werden (z.B. *Allgegenwärtigkeit*). Es wurde somit außerdem dargestellt, dass dieses erfolgreiche Agieren weniger mit besonderen „Techniken“, spezifischen Reaktionen auf Fehlverhalten oder ausgewählten Sozialformen, d.h. mit Oberflächenmerkmalen, erklärbar ist (vgl. Klieme & Rakoczy, 2008).

Tatsächlich konnten diese Erkenntnisse in Bezug auf etwa die Schaffung von Lerngelegenheiten immer wieder bestätigt werden (*opportunities to learn*, z.B. Brophy, 2000). Darüber hinaus konnte vielfach empirisch gezeigt werden, dass ein störungsfreier und gut organisierter Unterricht positive Wirkungen sowohl auf kognitive als auch affektiv-motivationale Merkmale bei Schülerinnen und Schülern zeigt (Klieme & Rakoczy, 2008; Seidel & Shavelson, 2007). Dies resultiert zum einen in einer höheren Unterrichtsqualität (*Prozess*), zum anderen aber auch in der langfristig positiven Entwicklung von Schülerleistungen (*Produkt*). Daraus ergibt sich, dass eine gute Klassenführung als notwendige Bedingung für erfolgreiches Lernen betrachtet werden kann (Helmke, 2012).

Weitere Merkmale von Unterricht wie *advanced organizing* bzw. die Strukturiertheit des Unterrichts stehen oft in enger Verbindung mit der Störungsfreiheit des Unterrichts (vgl. Klieme & Rakoczy, 2008), weshalb sie ebenfalls hier aufgeführt werden.

Effektive Lernzeitnutzung			
Beschreibung des Items Quelle: Eigenentwicklung nach Helmke (2012) Das Item beschreibt, inwieweit eine effektive Nutzung der zur Verfügung stehenden Zeit stattfindet, d. h. inwieweit Schülerinnen und Schüler stets etwas Sinnvolles zu arbeiten haben. Dazu zählt etwa, dass der Unterricht pünktlich beginnt und endet und dass die Zeit im Unterricht vor allem für unterrichtsbezogene Inhalte genutzt wird. Negativ auf die Lernzeit können sich Wartezeiten, organisatorische Probleme und Schwierigkeiten beim Umgang mit Medien auswirken. Besonders entscheidende Phasen sind daher der Unterrichtsbeginn und die Übergänge zwischen den Phasen (Helmke, 2012).			
Indikatoren • Der Unterricht beginnt und endet pünktlich • Die Übergänge zwischen den Phasen erfolgen reibungslos • Die Unterrichtszeit wird für unterrichtsbezogene Inhalte genutzt			
Anmerkungen/Hilfen zum Rating Die unten aufgelisteten Zeitangaben beziehen sich jeweils auf die Beobachtungsphase direkt vor einem Messzeitpunkt. Für Abzüge genügt es nicht, dass z. B. einige wenige Schülerinnen und Schüler bereits ihre Aufgaben abgeschlossen haben. In diesem Item wird nicht bewertet, ob die Schülerinnen und Schüler tatsächlich sinnvoll an ihren Aufgaben arbeiten. Entscheidend ist, ob die Lehrperson den organisatorischen Rahmen für eine sinnvolle Beschäftigung der Klasse schafft, für die die zur Verfügung stehende Zeit verwenden werden kann. Sind allerdings über die Hälfte der Schülerinnen und Schüler bereits fertig mit der Aufgabe, ist es ggf. für sie nicht mehr möglich, die Zeit weiter sinnvoll inhaltlich zu nutzen. In diesem Fall gibt es für dieses Item entsprechend der Angaben unten Abzüge.			
1	2	3	4
<i>Es wird mehr als fünfzehn Minuten Zeit verschwendet.</i>	<i>Es wird weniger als fünfzehn, aber mind. zehn Minuten Zeit verschwendet.</i>	<i>Es wird weniger als zehn, aber mind. fünf Minuten Zeit verschwendet.</i>	<i>Es wird weniger als fünf Minuten Zeit verschwendet.</i>

Sinnvolle Routinen und Regeln

Beschreibung des Items

Quelle: Eigenentwicklung nach Helmke (2012)

Mit diesem Item wird erfasst, inwieweit im Unterricht ein sinnvolles Regelsystem existiert und inwieweit es Routinen gibt, die von allen akzeptiert und eingehalten werden. Solche Regeln und Routinen werden meist zu Beginn des Schuljahres in der Klasse etabliert und müssen von der Lehrperson und den Schülerinnen und Schülern über das Schuljahr hinweg gefestigt werden, so dass das Regelsystem fester Bestandteil der Unterrichtskultur ist und bleibt. Routinen, verstanden als Handlungen, die immer den gleichen Ablauf haben, symbolische Bedeutung besitzen und in ihrer Bedeutung von den Beteiligten unmittelbar verstanden werden, strukturieren den zeitlichen Ablauf und das Unterrichtsgeschehen und geben somit Orientierung und Sicherheit. Diese beziehen sich ebenso wie Regeln auf verschiedene Aspekte des Unterrichts (beispielsweise Begrüßungs- und Abschiedsrituale, Routinen zur Konfliktlösung).

Indikatoren

- Eingebaute Muster für die Lernorganisation sind ersichtlich
- Die SuS übernehmen Organisationsaufgaben

Anmerkungen/Hilfen zum Rating

Bei einem reibungslosen Verlauf des Unterrichts ohne beobachtbare Routinen und Regeln vergibt man eine „3“. Für eine „4“ müssen besondere Regeln beobachtbar sein, d. h. etwa Austeildienste, Tafeldienst, Meldekette, Ermahnungskarten, Konfliktlöserituale usw., für ein hohes Rating genügt aber nicht der plakative Aushang von Regeln im Klassenzimmer.

1	2	3	4
<i>Der Unterrichtsverlauf kommt durch fehlende Regeln und Routinen für mehr als zehn Minuten ins Stocken.</i>	<i>Der Unterrichtsverlauf kommt durch fehlende Regeln und Routinen für weniger als zehn Minuten ins Stocken.</i>	<i>Der Unterricht verläuft bzgl. organisatorischer Aspekte reibungslos.</i>	<i>Der Unterricht verläuft reibungslos und es gibt besondere Regeln oder Routinen.</i>

Störungspräventives Verhalten

Beschreibung des Items

Quelle: Eigenentwicklung nach Kounin (1970)

Dieses Item beschreibt einerseits, inwieweit die Lehrperson den Unterricht störungsfrei gestaltet bzw. andererseits, wie mit aufkommenden Störungen umgegangen wird. Zu einem störungspräventiven Verhalten im Unterricht gehört, dass die Lehrkraft jederzeit einen Überblick über das Unterrichtsgeschehen und die Schüleraktivitäten hat (*Allgegenwärtigkeit*). Gelingt es der Lehrkraft z.B., auf Probleme so frühzeitig einzugehen, bevor diese Probleme zur Störung werden? Zu einem störungspräventiven Verhalten gehört außerdem, dass die Lehrperson in Gruppenphasen das produktive Arbeiten aller Schülerinnen und Schüler sicherstellt.

Indikatoren

- Die LK unterbindet aufkommende Störungen sofort und zielsicher
- Die LK hat den Überblick über das Unterrichtsgeschehen (*Allgegenwärtigkeit*)

Anmerkungen/Hilfen zum Rating

Gibt es keine Störungen, so geht man von einer guten präventiven Arbeit aus und gibt eine „4“. Mögliches Szenario für eine „3“: In Gruppenarbeitsphasen liegt die Aufmerksamkeit der Lehrperson oft bei nur einzelnen Schülerinnen und Schülern, andere scheinen sich deshalb unbeobachtet zu fühlen und arbeiten unproduktiver.

1	2	3	4
Es gibt Störungen, welche von der Lehrperson mit einem Zeitverlust bearbeitet werden, der in Summe zehn Minuten übersteigt.	Es gibt Störungen, welche von der Lehrperson mit einem Zeitverlust bearbeitet werden, der in Summe zehn Minuten nicht übersteigt.	Es gibt Störungen, welche von der Lehrperson mit einem Zeitverlust bearbeitet werden, der in Summe fünf Minuten nicht übersteigt.	Es gibt keine Störungen.

Transparenz des Lernprozesses			
Beschreibung des Items Quelle: adaptiert von Rakoczy & Pauli (2006), Clausen et al. (2003), Trepke et al. (2003) Dieses Item soll erfassen, inwieweit der Lehrperson ein umfangreiches <i>advanced organizing</i> gelingt, d.h. dass Schülerinnen und Schüler jederzeit über die Organisation ihres Lernprozesses Bescheid wissen. Hierbei soll der Klasse die Abfolge des Unterrichtsgeschehens, die Lernziele, Anforderungen und ggf. Arbeitsaufträge transparent gemacht werden. Die Lehrperson soll somit strukturierende Hinweise geben, die über den Unterrichtsinhalt hinausgehen. Dabei steht die Klarheit dieser Ausführungen im Vordergrund.			
Indikatoren • Die LK informiert die Schülerinnen und Schüler über Lernziele / Unterrichtsablauf • Die LK macht ihre Erwartungen transparent • Die Arbeitsaufträge sind präzise formuliert			
Anmerkungen/Hilfen zum Rating Ein „Kaltstart“, bei dem kein <i>advanced organizing</i> betrieben wird, den Schülerinnen und Schülern aber überwiegend offensichtlich klar ist, worum es geht, führt zur Vergabe einer „2“. Für eine höhere Bewertung müssen weitere Aspekte des Unterrichts transparent gemacht werden, z. B. durch Anschrieb an die Tafel.			
1	2	3	4
Es trifft kein Indikator zu und es entsteht der Eindruck, dass den SuS nicht klar ist, was von ihnen erwartet wird.	Es trifft nicht notwendig ein Indikator zu, aber es entsteht der Eindruck, dass den SuS klar ist, was von ihnen erwartet wird.	Mindestens ein Indikator trifft zu.	Alle Indikatoren treffen zu.

Arbeitsatmosphäre			
Beschreibung des Items Quelle: adaptiert von Rakoczy & Pauli (2006), Clausen et al. (2003) Dieses Item soll erfassen, inwiefern eine gute Arbeitsatmosphäre im Unterricht vorherrscht. Negativ wirken sich in diesem Sinne Disziplinprobleme, eine erhöhte Lautstärke oder allgemeine Unruhe aus, die ein produktives Arbeiten innerhalb der Klasse erschweren oder sogar verhindern können. Ebenso können äußere Faktoren (z.B. Baulärm) den Unterricht stören und ggf. das Durchsetzungsvermögen der Lehrperson einschränken. Auch der Umgang der Schülerinnen und Schüler untereinander hat letztlich Auswirkungen auf die Arbeitsatmosphäre.			
Indikatoren • Die Lautstärke ist angemessen • Die SuS reagieren auf Signale und Hinweise • Die SuS und die LK unterbrechen sich nicht gegenseitig			
Anmerkungen/Hilfen zum Rating			
1	2	3	4
Überwiegend schlechte Arbeitsatmosphäre.	Wiederholte Unruhe in der Klasse.	Überwiegend gute Arbeitsatmosphäre.	Angenehme Arbeitsatmosphäre.

Strukturierung des Unterrichts			
Beschreibung des Items Quelle: Eigenentwicklung nach Helmke (2012) Dieses Item soll erfassen, inwieweit der Unterricht unter einer didaktischen Perspektive klar strukturiert und gegliedert ist, so dass ein inhaltlicher roter Faden erkennbar wird. Dazu gehört insbesondere, dass die Stunde eine klare Phasierung und einen inhaltlich sinnvollen Abschluss aufweist. Das Verhalten der Lehrperson ist dabei etwa im Hinblick auf logische oder inhaltliche Brüche in der Stundenplanung entscheidend.			
Indikatoren • Der Unterricht besitzt einen „roten Faden“ • Der Unterricht ist in Phasen unterteilt • Die LK findet einen angemessenen inhaltlichen Abschluss der Stunde			
Anmerkungen/Hilfen zum Rating Eine gute Strukturierung des Unterrichts ist die Abwesenheit von logischen oder inhaltlichen Brüchen. Es ist zu jedem Zeitpunkt klar, in welcher Phase die Schülerinnen und Schüler sich befinden und welche Unterrichtsphase sich anschließt.			
1	2	3	4
Kein Indikator trifft zu, d.h. der „rote Faden“ und die Phasierung des Unterrichts sind überwiegend nicht erkennbar.	Der „rote Faden“ und die Phasierung des Unterrichts sind stellenweise nicht erkennbar.	Der Unterricht ist überwiegend gut strukturiert, der „rote Faden“ ist größtenteils erkennbar.	Der Unterricht ist sehr gut strukturiert und klar in Phasen unterteilt. Der „rote Faden“ ist immer erkennbar.

Dimension 2: Konstruktive Unterstützung

Gemäß konstruktivistischer Lerntheorien ist eine aktive Beteiligung von Schülerinnen und Schülern im Unterricht notwendig für erfolgreiche Lernprozesse (Reinmann-Rothmeier & Mandl, 2006; Turner et al., 1998). Hierfür benötigen Lernende eine unterstützende Umgebung, welche von der Lehrperson vor allen Dingen eine aufmerksame Diagnose von Schwierigkeiten und Fortschritten im Lernprozess erfordert (Kunter & Voss, 2011). Das Konzept der konstruktiven Unterstützung beinhaltet somit Merkmale, wie sie in der Literatur zum *scaffolding* beschrieben werden (Pea, 2004; Wood, Bruner & Ross, 1976). Dabei geht es vor allem um die strukturierte Anpassung der Lerngelegenheiten für einzelne Schülerinnen und Schüler (Reiser, 2004). Dies kann durch Begleitung, Vereinfachung oder sonstige Unterstützungsmöglichkeiten geschehen, die adaptiv eingesetzt werden sollten (z.B. nach dem Prinzip der minimalen Hilfe (Zech, 1998)).

Ein weiterer Aspekt, der sich im Zusammenhang mit konstruktiver Unterstützung von Lernenden als erfolgreich erwiesen hat, ist die Qualität der Beziehung zwischen Lehrperson und Klasse (Kunter & Voss, 2011). Die Merkmale der Qualität der Beziehung zwischen Lehrkraft und Lernenden lassen sich im Rahmen der Forschung zum Unterrichtsklima einordnen (Den Brok, Brekelmans & Wubbels, 2004; Fraser, 1991). Hieran anschließend sollten Unterrichtsgespräche von beiderseitiger Wertschätzung geprägt sein und von der Art, dass alle Beteiligten als autonome und selbstständige Personen voneinander wahrgenommen werden (Cornelius-White, 2007). Insbesondere liefert die Qualität der Beziehung von Lehrenden und Klasse einen großen Beitrag zur emotionalen und motivationalen Unterstützung des Lernenden. Dies wirkt sich vor allem dann positiv aus, wenn etwa die Lehrperson besonders geduldig mit Fehlern und Verständnisschwierigkeiten umgeht und die Beiträge der Schülerinnen und Schüler zum Unterricht wertschätzt (Kunter & Voss, 2011).

Schließlich soll diese Dimension auch den Umgang der Lehrperson mit der gegebenen Heterogenität in der Klasse erfassen. Individualisierung und Differenzierung ist vor diesem Hintergrund eine besondere Herausforderung, für das Kompetenz- und Autonomieerleben des einzelnen Lernenden jedoch von großer Bedeutung (Rakoczy & Pauli, 2006).

Individuelle Förderung			
Beschreibung des Items Quelle: Eigenentwicklung Dieses Item beschreibt, inwieweit es der Lehrperson gelingt, Schülerinnen und Schüler in ihrem individuellen Lernprozess und gemäß ihrer individuellen Voraussetzungen zu unterstützen. Es kommt dabei nicht nur auf die Zeit an, die einzelnen Schülerinnen und Schülern gewidmet wird, sondern auch auf die Adaptivität in Bezug auf Hilfestellungen, Erklärungen etc. Insbesondere sollte die Lehrperson diagnosebasiert handeln. In diesem Zusammenhang ist der Begriff des „Scaffolding“ aus der Literatur einschlägig, d.h. die Lehrperson bemüht sich um individuelle Hilfestellungen für einzelne Schülerinnen und Schüler (<i>Prinzip der minimalen Hilfe</i>, Zech, 1998).			
Indikatoren • Die LK informiert sich über individuelle Schwierigkeiten / Fortschritte • Die LK nimmt sich Zeit für einzelne SuS • Die LK gibt den SuS individuelle Hilfestellungen			
Anmerkungen/Hilfen zum Rating In Plenumsphasen ist i. A. höchstens eine „2“ zu vergeben. Dazu muss dann aber explizit nach dem Verständnis der Schülerinnen und Schüler gefragt werden bzw. es muss in besonderem Maße auf Verständnisfragen von einzelnen Schülerinnen und Schüler eingegangen werden. Für eine hohe Bewertung ist es nicht zwingend erforderlich, dass die Indikatoren für den Großteil der Schülerinnen und Schüler beobachtet werden (<i>Qualität</i> der individuellen Förderung). Wichtig ist hierbei, dass diese Förderung prinzipiell jedem Schüler und jeder Schülerin in der Klasse zur Verfügung steht. Sollte in der Klasse ein Sozialpädagoge o.ä. vorhanden sein und wird er explizit mit ins Unterrichtsgeschehen einbezogen, kann eine höhere Bewertung bis „4“ gegeben werden. Voraussetzung ist dabei, dass die Schülerinnen und Schüler durch den Sozialpädagogen eine höhere individuelle Förderung erhalten.			
1	2	3	4
Für >50% der Interaktionen zwischen LK und SuS trifft kein Indikator zu.	Für >50% der Interaktionen zwischen LK und SuS trifft genau ein Indikator zu.	Für >50% der Interaktionen zwischen LK und SuS treffen genau zwei Indikatoren zu.	Für >50% der Interaktionen zwischen LK und SuS treffen alle Indikatoren zu.

Umgang mit Heterogenität/Differenzierung			
Beschreibung des Items Quelle: adaptiert von Rakoczy & Pauli (2006), Clausen et al. (2003) Schülerinnen und Schüler weisen eine große Heterogenität etwa durch unterschiedliche individuelle Voraussetzungen und Leistungsniveaus auf. Daher werden Erwartungen der Lehrperson auch als individuell unterschiedlich anspruchsvoll von ihnen aufgefasst. Damit die Förderung der gesamten Klasse dennoch gleichermaßen möglich ist, müssen sich die Arbeitsaufträge der Lehrperson an eben diesen individuellen Voraussetzungen der einzelnen Schülerinnen und Schüler orientieren. Außerdem ist es ein erklärtes Ziel erfolgreichen Lehrerhandelns, der Heterogenität der Schülerinnen und Schüler auch im Hinblick auf andere Aspekte wie Geschlecht oder Sprache gerecht zu werden.			
Indikatoren • Es gibt zusätzliches Material für Teile der Lerngruppe • Es gibt anforderungsbezogene, differenzierende Aufgabenstellungen • Es findet ad-hoc Differenzierung statt (z. B. leichtere Fragestellung)			
Anmerkungen/Hilfen zum Rating Um das Differenzierungspotenzial der Aufgaben beurteilen zu können, können die Unterlagen aus der Unterrichtsstunde hinzugezogen werden. Nicht gemeint ist, dass die LK verschiedene Aufgaben nach methodischen Gesichtspunkten verteilt (z.B. Gruppenpuzzle).			
1	2	3	4
Es findet keinerlei Differenzierung statt.	Differenzierung findet in geringem Maße oder ad-hoc statt (z.B. Zusatzaufgaben)	Es wird in hohem Maße differenziert, aber lediglich in einer Form.	Es wird in mehreren Formen und aufwändig differenziert (z.B. inhaltlich und nach Schwierigkeit).

Selbstgesteuertes Lernen			
Beschreibung des Items Quelle: adaptiert von Rakoczy & Pauli (2006), Kunter (2005) Mit diesem Item wird die Qualität und Anzahl der Freiheitsgrade erfasst, die den Schülerinnen und Schülern bezüglich ihres eigenen Lernprozesses zur Verfügung stehen. Positiv wirkt sich somit aus, wenn die Lehrperson der Klasse selbst die verantwortliche Entscheidung gewisser didaktisch-methodischer oder organisatorischer Gestaltungsmöglichkeiten überlässt. Die Lehrperson liefert also ein Unterrichtsangebot, durch das sich Schülerinnen und Schüler selbst fordern und fördern können. Um solch ein Angebot der Einflussnahme nutzen zu können, wird der Klasse ggf. Materialien zum eigenständigen Lernen und zur Ergebnissicherung angeboten. Negativ wirkt sich dagegen aus, wenn der Lernprozess vollständig durch die Instruktionen der Lehrperson determiniert ist.			
Indikatoren • Die SuS überprüfen ihre Ergebnisse selbstständig an einer Musterlösung • Die LK fordert eigenständiges Arbeiten in hohem Maße ein • Die LK ermöglicht den Schülerinnen und Schülern die Wahl von Sozialformen			
Anmerkungen/Hilfen zum Rating Wahlmöglichkeiten können auch vorkommen, ohne dass sie explizit genannt werden.			
1	2	3	4
Kein Indikator trifft zu.	Genau ein Merkmal trifft zu (z.B. eigenständiges Überprüfen an einer Musterlösung, Wahl der Sozialform).	Mehrere Merkmale treffen zu.	SuS arbeiten weitgehend ohne direkte Instruktion der LK.

Rückmeldungen der Lehrkraft			
Beschreibung des Items Quelle: Rakoczy & Pauli (2006), Kobarg & Seidel (2003) Damit Schülerinnen und Schüler Fortschritte in ihrem Lernprozess machen können, muss die Lehrperson Rückmeldungen sachlich und konstruktiv äußern. Hilfestellungen sollten differenziert formuliert werden, Kritik und Korrektur sollte stets auch eine zukunftsgerichtete Komponente beinhalten. Zudem sollten Rückmeldungen in einem angenehmen Tonfall erfolgen, um den motivierenden Charakter zu erhalten und in ausreichendem Maß vorhanden sein.			
Indikatoren • Die Rückmeldungen der LK sind differenziert • Die Rückmeldungen der LK sind sachlich und konstruktiv • Die Rückmeldungen der LK sind zukunftsgerichtet			
Anmerkungen/Hilfen zum Rating Wenn sowohl sachliche als auch unsachliche Rückmeldungen beobachtet werden, wird eine 2 vergeben. Manchmal ist fraglich, welche Äußerungen überhaupt Rückmeldungen sind, da sie „ungefragt“ erfolgen. Auch wenn dieser Umstand das Bild dominiert, wird eine 2 vergeben.			
1	2	3	4
Mehr als 50% der Rückmeldungen erfüllen keinen der Indikatoren.	Mehr als 50% der Rückmeldungen erfüllen genau einen Indikator.	Mehr als 50% der Rückmeldungen erfüllen genau zwei Indikatoren.	Mehr als 50% der Rückmeldungen erfüllen alle Indikatoren.

Wertschätzung			
Beschreibung des Items Quelle: adaptiert von Rakoczy & Pauli (2006) Wichtig für die Motivationsunterstützung von Schülerinnen und Schülern ist auch die Wertschätzung, die ihnen von der Lehrperson entgegengebracht wird. Hier soll somit erfasst werden, mit welcher grundsätzlichen Haltung diese ihrer Klasse begegnet. Insbesondere sollte die Lehrperson geduldig und respektvoll auftreten und die Schülerinnen und Schüler in ihrer Person ernst nehmen. Dies zeigt sich etwa daran, dass die Äußerungen der Lehrperson frei von Zynismus sind.			
Indikatoren • Die LK hat Geduld • Die LK verstärkt positiv durch Lob • Die LK ermutigt die Schülerinnen und Schüler			
Anmerkungen/Hilfen zum Rating Auch (Lösungs-)Präsentationen von Schülergruppen an der Tafel, am Smartboard o.Ä. können als Wertschätzungen individueller Leistungen angesehen werden.			
1	2	3	4
Es ist keine grundsätzliche Wertschätzung der SuS zu erkennen.	<i>Die Lehrperson wirkt teilweise ungeduldig.</i> <i>Die Lehrperson übergeht Äußerungen von SuS.</i> Es gibt einzelne Situationen, in denen die Lehrkraft den SuS gegenüber wenig wertschätzend ist.	Die Lehrkraft ist grundsätzlich wertschätzend, aber gibt wenig Lob oder Ermutigung.	Alle Indikatoren treffen zu.

Feedback der Schülerinnen und Schüler			
Beschreibung des Items Quelle: Eigenentwicklung nach Helmke (2012) Professionalisierung benötigt auch die Reflexion des eigenen Handelns im Unterricht. Hierzu hat die Unterrichtsforschung der letzten Jahre gezeigt, dass Feedback von Schülerinnen und Schülern prognostisch valide für erfolgreiches Lernen der Klasse ist. Insbesondere kann eine weitere Perspektive neben der der Lehrperson zur positiven Unterrichts- und Schulentwicklung beitragen. Dieses Item soll erfassen, inwieweit den genannten Aspekten durch die Lehrperson Rechnung getragen wird.			
Indikatoren • Die LK holt sich ein Feedback ein • Die LK geht auf Anregungen ein • Die LK redet mit den SuS über Unterricht und Klasse			
Anmerkungen/Hilfen zum Rating			
1	2	3	4
Es gibt kein Feedback.	Es findet ein unreflektiertes Feedback statt (z.B. Daumenfeedback ohne Kommentare).	Es gibt einzelne Feedback-Kommentare mit Reflexion oder ein differenziertes Feedback.	Es findet ein differenziertes Feedback mit Reflexion statt.

Förderung von Kooperation			
Beschreibung des Items Quelle: Rakoczy & Pauli (2006), Clausen et al. (2003) Dieses Item soll die Qualität von Aushandlungsprozessen der Schülerinnen und Schüler untereinander erfassen. Dazu gehört zunächst, dass der Klasse überhaupt kooperative Arbeiten nahegelegt werden, darüber hinaus aber auch gerade solche Aufgaben gestellt werden, die Einigung möglich und nötig machen. Dies bedeutet, dass die Lehrperson stärker als Moderator oder Mediator auftritt und weniger selbst als „Meinungsgeber“ fungiert. Sie hat dann vielmehr die Aufgabe, Gelegenheiten für das Ausbilden demokratischer Kompetenzen zu schaffen.			
Indikatoren • Die LK initiiert gemeinsame Lernprozesse zwischen den SuS • Die LK stellt Aufgaben, die Einigung erfordern • Die LK nimmt die Rolle eines Mediators ein • Die SuS helfen sich gegenseitig			
Anmerkungen/Hilfen zum Rating Eine „4“ wird nur dann vergeben, wenn tatsächlich Einigungsprozesse nötig sind, von der Lehrperson eingefordert werden und eine angemessene Einigung gefunden wird.			
1	2	3	4
Es finden keine kooperativen Arbeiten statt.	Es finden kooperative Arbeiten statt (z. B. Gruppenarbeit, Meldekette im Plenum).	Es sind Einigungsprozesse nötig, die aber nicht explizit eingefordert werden.	Es sind Einigungsprozesse nötig, die auch stattfinden.

Dimension 3: Kognitive Aktivierung

Die Basis für diese Dimension ist ein sozial-konstruktivistisches Lehr-Lernverständnis mit der Annahme, dass Wissen das Ergebnis aktiver und individueller Konstruktionsprozesse ist, die intern, selbstgesteuert und subjektiv ablaufen, aber im Unterricht durch soziale Kontexte unterstützt werden (Reinmann-Rothmeier & Mandl, 2003). Neue Erkenntnisse knüpfen somit immer an bereits vorkonstruierte Wissensstrukturen an (Aebli, 1985; Reusser, 1998; Shuell, 1993). Nach dieser Annahme genügt es für ein tiefgehendes Verständnis von Unterrichtsinhalten also nicht, dass der Klasse ausreichend Lernzeit zur Verfügung steht, so dass bereits automatisch Lernerfolge erzielt werden könnten, wenn nur ausreichend geübt wird (Rakoczy & Pauli, 2008). Auch steht nicht eine hohe allgemeine Aktivität von Schülerinnen und Schülern im Fokus dieser Dimension, die sich z.B. in einer hohen Anzahl an Freiheitsgraden niederschlägt, wie sie etwa in o.g. Item „Selbstgesteuertes Lernen“ erfasst wird (z.B. Stefanou, Perencevich, DiCintio & Turner, 2004), sondern vielmehr um die mentale Aktivität der Klasse. Die Aufgabe der Lehrperson besteht nun darin, im Unterricht Gelegenheiten für anspruchsvolle Lernprozesse und für die interaktive *Ko-Konstruktion* von Wissen anzubieten und ferner die Schülerinnen und Schüler in ihren individuellen Konstruktionsprozessen zu unterstützen (Kunter, 2005). Dies wird mit dem Begriff der kognitiven Aktivierung umschrieben (vgl. Baumert et al., 2010; Klieme, Schümer & Knoll, 2001; Lipowsky et al., 2009), also dem Potenzial des Unterrichts, herausfordernde kognitive Aktivitäten der Lernenden hervorzubringen.

Herausfordernde Fragen und Aufgaben			
Beschreibung des Items Quelle: Rakoczy & Pauli (2006), Kunter (2005) Dieses Item soll herausfordernde Fragen und Aufgabenstellungen erfassen, mit denen sich Schülerinnen und Schüler im Unterricht auseinandersetzen. Diese Lerngelegenheiten zeichnen sich etwa dadurch aus, dass die Schülerinnen und Schüler durch kognitive Konflikte herausgefordert werden oder keine standardisierten Verfahren zur Lösung vorhanden sind. Eine Rolle spielt dabei einerseits die Offenheit der Fragestellung und ihre Komplexität, aber auch ihre Angemessenheit bezogen auf Leistungsstand und Voraussetzungen der Klasse.			
Indikatoren • Die LK stellt offene Fragen • Die LK zeigt ein angemessenes Frageverhalten (nicht zu kleinschrittig) • Die LK stellt kognitiv herausfordernden Aufgaben • Die LK gibt geeignete Impulse, z.B. ein Problem oder widersprüchliche Sachverhalte			
Anmerkungen/Hilfen zum Rating Es kann auch eine anspruchsvolle Aufgabe gestellt werden, die anschließend „kleingearbeitet wird“. Entsprechend müsste das Rating dann niedriger ausfallen. Dabei wird zwischen Herausforderung und Überforderung unterschieden. Beides bezieht sich nur auf inhaltliche Aufgaben, nicht auf anderes, wie etwa die Erstellung eines Arbeitsplans. Ein „MISSING“ wird vergeben, wenn in einer Phase überwiegend keine Aufgaben zu bearbeiten sind (z. B. Gespräch über organisatorische Dinge oder Klasseninterna).			
1	2	3	4
Kein Indikator trifft zu.	Genau ein Indikator trifft zu.	Mehrere Indikatoren treffen zu.	Alle Indikatoren treffen zu.

Förderung von Metakognition			
Beschreibung des Items Quelle: Eigenentwicklung in Anlehnung an Lipowsky (2015), Hasselhorn & Labuhn (2008) Metakognitive Prozesse werden üblicherweise in die vier Bereiche Planung, Steuerung, Regulation und Bewertung unterteilt. Mit diesem Item wird erfasst, inwiefern diese Aspekte im Unterricht eine Rolle spielen. Das Wissen über den eigenen Lernprozess steht daher in enger Beziehung zur kognitiven Aktivierung von Schülerinnen und Schülern und trägt maßgeblich zum Lernerfolg bei. Die Lehrperson hat daher die Aufgabe, die Klasse insofern metakognitiv zu fördern, dass z.B. Lernstrategien und Arbeitstechniken explizit thematisiert und weiterentwickelt werden.			
Indikatoren • Es findet mindestens ein metakognitiver Teilprozess sichtbar statt • Die LK gibt Zeit für metakognitive Prozesse (z. B. Lerntagebuch) • Methodische Vorgehensweisen werden besprochen • Der Nutzen von Methoden wird reflektiert			
Anmerkungen/Hilfen zum Rating			
1	2	3	4
Kein Indikator trifft zu.	Einzelne Aspekte von Metakognition werden berücksichtigt, sie stehen aber nicht im Zentrum des Lernprozesses.	Entweder steht ein einzelner Teilprozess im Zentrum des Unterrichts oder mehrere Teilprozesse werden bearbeitet, die aber nicht zentral sind.	Mehrere metakognitive Teilprozesse werden berücksichtigt. Insgesamt steht die Reflexion über das Lernen im Zentrum des Unterrichts.

Aktivierung des Wissens			
Beschreibung des Items Quelle: Rakoczy & Pauli (2006), Kunter (2005) Dieses Item erfasst das Einbeziehen von Vorwissen der Klasse in den Unterricht. Hierunter werden solche Konzepte und Vorstellungen der Schülerinnen und Schüler verstanden, die entweder in der Vergangenheit im Unterricht thematisiert oder außerhalb der Schule erworben wurden. Dies kann passieren, indem die Lehrperson nach einem solchen Vorwissen fragt, ohne eine Bewertung der Antworten im Sinne der Kategorien „richtig“ oder „falsch“ vorzunehmen. Insbesondere geht es dann aber auch darum, dass das Wissen der Klasse im Unterricht eine Rolle spielt, darauf aufgebaut oder mindestens in irgendeiner Form damit weitergearbeitet wird.			
Indikatoren • Die LK fragt nach den Vorstellungen der SuS zum Thema • Die SuS erklären das Thema in ihren eigenen Worten • Die LK aktiviert das bei SuS vorhandene Wissen			
Anmerkungen/Hilfen zum Rating Hilfreich: Werden tatsächlich eigene Konzepte der Schülerinnen und Schüler aufgegriffen? Inwiefern gibt es ausreichend Anknüpfungspunkte an Themen oder Aufgaben aus vorigen Stunden? Nicht in diese Kategorie fällt das einfache Aufgreifen von Schülerantworten.			
1	2	3	4
Das Wissen der SuS wird nicht aktiviert.	Das Wissen der SuS wird in Ansätzen aktiviert, ist für den Unterricht aber nicht bedeutsam.	Das Wissen der SuS wird aktiviert und ist für den Unterricht bedeutsam.	Das Wissen der SuS wird in hohem Maße aktiviert und ist für den Unterricht sehr bedeutsam.

Evolutionärer Umgang (Ko-Konstruktion)			
Beschreibung des Items Quelle: Rakoczy & Pauli (2006), Kunter (2005) Dieses Item soll erfassen, inwiefern aus dem Vorwissen von Schülerinnen und Schülern „evolutionär“ eine wissenschaftliche Sichtweise entwickelt wird. Dies könnte in Form einer Erweiterung oder Vertiefung des vorhandenen Wissens geschehen, oder so, dass neues Wissen kooperativ in der Klasse konstruiert wird. Insbesondere findet dabei eine Veränderung von Konzepten statt (<i>conceptual change</i>). Mit diesem Item soll darüber hinaus festgestellt werden, inwieweit die Lernenden in die Entstehung solcher neuen Konzepte einbezogen werden, z.B. durch Interaktion mit der Lehrperson (fragend-entwickelndes Unterrichtsgespräch).			
Indikatoren • Die Fachbegriffe werden aus den Vorstellungen der SuS entwickelt • Die LK nutzt Fragetechniken zur Exploration und Weiterentwicklung • Es findet eine gemeinsame Konstruktion neuen Wissens statt			
Anmerkungen/Hilfen zum Rating Dieses Item wird nur dann bewertet, wenn neue Dinge im Unterricht behandelt werden (nicht in Übungsphasen). Insbesondere geht es um die Qualität fragend-entwickelnder Gespräche.			
1	2	3	4
Kein Indikator trifft zu.	Die SuS werden in geringem Maße in die Wissenskonstruktion einbezogen.	Die SuS werden in hohem Maße in die Wissenskonstruktion einbezogen.	Alle Indikatoren treffen zu.

Qualität der Methoden			
Beschreibung des Items Quelle: Eigenentwicklung in Anlehnung an Helmke (2012) Zwar soll durch Beobachterratings die Tiefenstruktur des Unterrichts erfasst werden, wohingegen Unterrichtsmethoden unbestreitbar eher der Sichtstruktur zuzuordnen sind. Dennoch können gewisse Methoden ein größeres Potential zur kognitiven Aktivierung als andere aufweisen. Eine Methode ist nach diesem Verständnis genau dann von hoher Qualität, wenn sie die Schülerinnen und Schüler kognitiv aktiviert, d. h. zu eigenständigem und anspruchsvollem Denken herausfordert. In diesem Item wird also die Qualität von Unterrichtsmethoden im Hinblick auf ihre kognitive Aktivierung erfasst. Dabei spielen auch Wechsel von Methoden eine Rolle.			
Indikatoren • Es werden kognitiv aktivierende Methoden verwendet • Die LK gibt angemessene Zeit zum Nachdenken im Plenum • Die Methodenwahl ist dem Inhalt und den SuS angemessen			
Anmerkungen/Hilfen zum Rating Durch das Rating wird in erster Linie das <i>Angebot</i>, weniger die <i>Nutzung</i> beurteilt. Dennoch liefert die Wahrnehmung, ob die Schülerinnen und Schüler kognitiv aktiviert wurden, einen Hinweis auf die Qualität der Unterrichtsmethoden. Geeignete Methoden können u. a. sein: Fragend-entwickelnder Unterricht, DAB, entdeckender Unterricht, Gruppenpuzzle, Lernen durch Lehren, Expertengruppen, Planspiel etc. Zu berücksichtigen ist, dass die gewählte Methode auch zur Lerngruppe passen muss. Dies richtig einzuschätzen erfordert ggf. zusätzliche Informationen.			
1	2	3	4
Die Methode aktiviert die SuS nicht.	Die Methode aktiviert die SuS wenig.	Die Methode aktiviert die SuS überwiegend.	Die Methode aktiviert die SuS durchgängig.

Wissenssicherung			
Beschreibung des Items Quelle: Eigenentwicklung in Anlehnung an Helmke (2012), Lipowsky (2015) Aus der Unterrichtsforschung ist seit längerer Zeit bekannt, dass Inhalte erst dann dauerhaft gelernt werden können, wenn eine ausreichend häufige Wiederholung des Lernstoffes stattfindet. Im besten Fall erfolgt dies auf verschiedenen Arten, etwa mündlich im Klassengespräch sowie schriftlich in Form einer Dokumentation. In diese Kategorie fallen die Vertiefung von Unterrichtsinhalten durch Übung, aber auch Konsolidierungsphasen, die z.B. durch anspruchsvolle Diskussionen angeregt werden können. Ebenso erfasst dieses Item die Qualität der aus der Unterrichtspraxis bekannten Phase der Ergebnissicherung.			
Indikatoren • Es findet eine Wissenssicherung anhand von Beispielen oder Erinnerungshilfen statt • Es kommen Wiederholungen zur Vernetzung vor • Relevante Zwischenschritte werden besprochen			
Anmerkungen/Hilfen zum Rating Findet keine Wissenssicherung statt, da sie in dieser Phase nicht sinnvoll ist, so wird ein MISSING vergeben (z. B. in Stillarbeitsphasen). Für mündliche Wiederholungen ohne Konsolidierung wird eine „2“ vergeben (<i>was haben wir letzte Stunde gemacht?</i>).			
1	2	3	4
Kein Indikator trifft zu.	Es finden weder Konsolidierung noch Dokumentation statt.	Wissenssicherung mit Konsolidierung oder mit Dokumentation findet statt.	Wissenssicherung mit Konsolidierung und Dokumentation findet statt.

Dimension 4: Mathematikdidaktische Unterrichtsqualität (MDU)

Die Items in den drei ersten Dimensionen beschreiben in erster Linie generische Aspekte von Unterrichtsqualität, die in jedem Unterrichtsfach relevant sind. Darüber hinaus gibt es für den Mathematikunterricht jedoch eine Reihe weiterer Qualitätsaspekte, die in den drei generischen Basisdimensionen nicht erfasst werden. Blum et al. (2006) sprechen diesbezüglich von einer fachlich gehaltvollen Gestaltung von Mathematikunterricht neben den drei Basisdimensionen.

Aus diesem Grund werden in der vierten Dimension verschiedene fachspezifische Merkmale von Unterrichtsqualität berücksichtigt, die nicht bereits in den anderen drei Dimensionen enthalten sind. Diese fachspezifischen Aspekte bilden einen breiten Konsens zu den Aspekten, die in anderen fachspezifischen Beobachtungsinstrumenten für den Mathematikunterricht erhoben wurden, und bei denen vermutet wird, dass sie sich positiv auf den Leistungszuwachs der Schülerinnen und Schüler auswirken. Sie stehen im Einklang mit einem konstruktivistischen Lernverständnis, indem das Lernen als aktiver, eigenständiger Prozess angesehen wird, der nicht direkt von außen gesteuert werden kann. Ziel ist es dabei, eine Entwicklung insbesondere fachlicher Kompetenzen zu ermöglichen. Zu diesen Aspekten zählen:

- Repräsentationsformen und Darstellungswechsel
- Mathematische Fachsprache
- Mathematische Kompetenzen (z.B. Argumentieren, Problemlösen, Modellieren)
- Fachliche Tiefe (Vernetzungen und Verallgemeinerungen)
- Mathematische Korrektheit der Lehrperson
- Umgang mit Fehlern
- Geeignete Beispiele
- Erklärungen der Lehrperson
- Intelligentes Üben
- Relevanz
- (Materialien)

Konstruktiver Umgang mit Fehlern

Beschreibung des Items

Quelle: adaptiert von Rakoczy & Pauli (2006), Clausen, Reusser & Klieme (2003)

Fehler sind Ereignisse im Unterricht, die immer wieder auftreten, sowohl von Seiten der Schülerinnen und Schüler als auch in seltenen Fällen von Seiten der Lehrperson. Sie gehören zum Lernprozess der Schülerinnen und Schüler als produktive Bestandteile mit dazu. Oftmals werden diese Situationen jedoch als unerwünscht angesehen. Dabei wird übersehen, dass Fehler im Unterricht auch eine positive Funktion haben können. Fehler im Unterricht können beispielsweise dazu dienen, fehlerhafte Vorstellungen aufzudecken oder Verständnisprobleme zu klären. Das Item erfasst, inwieweit mit Fehlern im Unterricht konstruktiv umgegangen wird, und inwieweit auftretende Fehler als Lerngelegenheiten erkannt und genutzt werden anstatt diese vorschnell zu korrigieren, zu übergehen oder sogar negativ zu bewerten. Es ist dabei wichtig, dass in Lernsituationen eine Fehlertoleranz herrscht, so dass die betroffene Schülerin bzw. der betroffene Schüler sich nicht bloßgestellt fühlt oder dass insgesamt Angst im Unterricht herrscht, Fehler zu machen.

Indikatoren

- Die LK nutzt Fehler als Lernanlass
- Die LK analysiert Schülerfehler und fehlerhafte Denkweisen
- Es herrscht eine Fehlertoleranz in Lernsituationen
- Fehler werden von den Schülerinnen und Schülern selbst korrigiert

Anmerkungen/Hilfen zum Rating

Für eine 4 reicht es nicht, dass eine Fehlertoleranz in Lernsituationen herrscht, sondern es ist auch notwendig, dass mit den auftretenden Fehlern konstruktiv im Sinne der Indikatoren umgegangen wird.

1	2	3	4
Es herrscht keine Toleranz für Fehler. <i>Fehler der SuS werden lediglich als falsch dargestellt. Es entsteht das Gefühl, dass es nicht „erlaubt“ sei, Fehler zu machen.</i>	<i>Es ist zwar „erlaubt“, Fehler zu machen, allerdings werden diese nicht als Lernanlass verwendet, sondern ebenfalls lediglich als falsch dargestellt.</i> <i>Fehler werden überwiegend einfach übergangen.</i>	<i>Fehler der SuS werden nicht einfach übergangen sondern von der LK aufgegriffen. Sie werden allerdings meistens schnell korrigiert und weniger als direkter Lernanlass genommen.</i> Es herrscht überwiegend eine hohe Fehlertoleranz im Kurs.	<i>Die LK greift Fehler und fehlerhafte Denkweisen auf und nutzt sie als Chance für die Weiterentwicklung. SuS korrigieren ihre Fehler gegenseitig.</i> <i>Fehler sind im Unterricht „willkommen“ und hilfreich und es wird mit ihnen konstruktiv umgegangen.</i>

Fachliche Korrektheit der Lehrkraft			
Beschreibung des Items Quelle: adaptiert von Hill, Charalambous & Kraft (2012) Mit diesem Item soll mathematisch fehlerhaftes oder unpräzises Vorgehen der Lehrkraft erfasst werden (z. B. Fehler beim Problemlösen, bei Definitionen oder beim Lösen von Gleichungen). Außerdem sollen hier auch Schülerfehler geratet werden, die von der Lehrkraft offensichtlich nicht als falsch erkannt wurden oder umgekehrt, d. h. die Lehrkraft bezeichnet eine korrekte Schülerlösung als falsch. Darüber hinaus werden aus mathematischer Sicht unpräzise oder falsche Äußerungen der Lehrkraft bewertet (etwa unpräzise Sprache oder mathematisch ungenaue Anschriebe an der Tafel).			
Indikatoren • Die LK macht keine inhaltliche Fehler • Die LK macht keine formalen Fehler • Die LK ist bei sprachlichen Formulierungen präzise • Die LK ist bei schriftlichen Formulierungen präzise			
Anmerkungen/Hilfen zum Rating Fehler, die innerhalb der Analyseeinheit korrigiert werden, werden nicht bewertet. Es ist zu prüfen, ob die Lehrkraft ggf. vorgibt, einen „typischen Fehler“ zu machen oder falsche Schülerantworten zunächst als korrekt hervorhebt, um etwas zu erklären (siehe auch konstruktiver Umgang mit Fehlern).			
1	2	3	4
Es trifft kein Indikator zu. <i>Ein sinnvoller Lernprozess ist nicht möglich, da die LK grobe inhaltliche Fehler macht.</i>	<i>Mehrere Fehler der Lehrperson, wodurch der Lernprozess beeinträchtigt wird.</i>	<i>Die Lehrperson macht einzelne Fehler, die aber den Verstehensprozess nicht behindern, oder sie ist mehrfach unpräzise.</i>	Es treffen alle Indikatoren zu. <i>Die LK ist sprachlich und schriftlich präzise und macht keine formalen oder inhaltlichen Fehler.</i>

Darstellungsformen			
Beschreibung des Items Quelle: adaptiert von Drollinger-Vetter & Lipowsky (2006) Im Mathematikunterricht sollten die mathematischen Inhalte mithilfe verschiedener Repräsentationsformen dargestellt werden. Das Ziel der Nutzung und Verknüpfung verschiedener Darstellungsformen ist dabei, ein vertieftes inhaltliches Verständnis auf Seiten der Schülerinnen und Schüler zu fördern (Bruner, 1966). Es wird angenommen, dass ein Unterricht insbesondere dann das Verständnis der Schülerinnen und Schüler fördert, wenn diese die Möglichkeit erhalten, den mathematischen Unterrichtsgegenstand aus verschiedenen Perspektiven und auf verschiedenen Repräsentationsebenen zu begreifen. Man unterscheidet drei verschiedene Repräsentationsebenen: die symbolische, die ikonische bzw. bildliche und die enaktive Repräsentationsform. Darüber hinaus unterteilt man die symbolische Ebene noch einmal in eine sprachliche und eine formale Ebene (Zech, 1998). Das Item erfasst, inwieweit Repräsentationsformen im Unterricht vorkommen und miteinander verknüpft werden.			
Indikatoren • Es werden verschiedene Darstellungsformen verwendet • Die Lehrkraft verdeutlicht Beziehungen zwischen Darstellungsformen			
Anmerkungen/Hilfen zum Rating Entscheidend für die Vergabe einer positiven Wertung ist, dass Beziehungen zwischen einzelnen Darstellungsformen thematisiert werden.			
1	2	3	4
Es wird nur eine Darstellungsform verwendet.	Es werden in Ansätzen mehrere Darstellungsformen verwendet, deren Beziehung nicht thematisiert wird.	Es werden verschiedene Darstellungsformen verwendet und in Beziehung gesetzt.	Es werden verschiedene Darstellungsformen verwendet, die in hohem Maße aufeinander bezogen werden (z.B. <i>Animationen in Geogebra</i>).

Intelligentes Üben			
Beschreibung des Items Quellen: Eigenentwicklung in Anlehnung an Leuders (2009), Wittmann (1992) In intelligenten Übungsaufgaben werden Vorstellungen gefestigt und mathematische Begriffe und Verfahren reflektiert. Schülerinnen und Schüler untersuchen mathematische Strukturen, lösen einfache mathematische Probleme und üben dabei ganz selbstverständlich Grundfertigkeiten. Gute Übungsaufgaben zeichnen sich dadurch aus, dass sie sinnstiftend, entdeckungsoffen und selbstdifferenzierend sind sowie reflexive Aspekte enthalten. Das bedeutet, dass den Schülerinnen und Schülern zum einen transparent gemacht wird, wozu die Übungen dienen. Des Weiteren zielt das Üben nicht auf das lineare Abarbeiten von vorgezeichneten Tätigkeiten, sondern ermöglicht es den Schülerinnen und Schülern, auf ihrem individuellen Leistungsstand eigene Lösungswege zu gehen und für sich einen Nutzen aus den Aufgaben zu ziehen. Schließlich ermöglichen intelligente Übungsphasen die Reflexion über den Lerngegenstand und die eigene Tätigkeit. Das Item erfasst, inwieweit in Übungsphasen intelligent geübt wird und inwieweit die Aufgaben in diesen Phasen die genannten Aspekte enthalten.			
Indikatoren • Die Übungen enthalten entdeckende Aspekte • Die Übungen enthalten reflexive Aspekte • Die Übungen sind selbstdifferenzierend			
Anmerkungen/Hilfen zum Rating Das Item wird nur bewertet, wenn das Üben im Vordergrund der Unterrichtsstunde bzw. der Unterrichtsphase liegt. Es ist dabei ggf. notwendig, die gestellten Aufgaben genauer anzuschauen.			
1	2	3	4
Kein Indikator trifft zu.	Ein Indikator trifft zu.	Zwei Indikatoren treffen zu.	Alle Indikatoren treffen zu.

Geeignete Beispiele			
Beschreibung des Items Quelle: Eigenentwicklung Mathematische Gegenstände sind ohne den Einsatz geeigneter Beispiele im Unterricht nicht behandelbar. Dies kommt insbesondere dann zum Tragen, wenn neue Begriffe oder Verfahren eingeführt werden. Ein Beispiel für einen solchen mathematischen Gegenstand sollte dabei im besten Fall nicht nur einen Bezug zum Alltag der Schülerinnen und Schüler aufweisen, sondern auch auf andere mathematische Gebiete hin „ausstrahlen“ (Wagenschein, 1999). Dies meint, dass durch ein geeignetes Beispiel fundamentale Einsichten über die zugrundeliegende Theorie gewonnen werden können (Klafki, 1963).			
Indikatoren • Die Beispiele deuten auf das Wesentliche hin • Die LK verwendet eine ausreichende Anzahl an Beispielen • Es werden Beispiele aus der Lebenswelt der SuS verwendet			
Anmerkungen/Hilfen zum Rating Für eine „4“ ist es nicht zwingend erforderlich, dass die Beispiele aus der direkten Lebenswelt der Schülerinnen und Schüler stammen.			
1	2	3	4
Kein Indikator trifft zu.	Die Beispiele deuten nicht auf das Wesentliche hin, allerdings trifft einer der anderen Indikatoren zu.	Die Beispiele deuten auf das Wesentliche hin.	Die Beispiele deuten auf das Wesentliche hin und sind in ausreichender Anzahl vorhanden oder der Lebenswelt der SuS entnommen.

Förderung von Fachsprache			
Beschreibung des Items Quelle: adaptiert von Hill et al. (2008), Freie und Hansestadt Hamburg (2011) Im Mathematikunterricht gibt es verschiedene Sprachregister. Zum einen nutzen die Schülerinnen und Schüler ihre Alltagssprache, um eigene Interessen und Erfahrungen zu äußern, Fragen zu stellen und neue mathematische Inhalte zu erkunden. Zum anderen entwickelt sich eine gemeinschaftlich genutzte Fachsprache, mit der mathematische Objekte präzise beschrieben werden können. Für den Lernprozess ist es außerdem bedeutsam, Alltagssprachlich über mathematische Sachverhalte zu diskutieren, da dies zu einem besseren Verständnis der zugrundeliegenden Gegenstände führt. Darüber hinaus ist es jedoch für eine gemeinsame erfolgreiche Verständigung auch über den Mathematikunterricht hinaus notwendig, die mathematische Fachsprache zu erlernen und zu verwenden. Dazu zählt beispielsweise auch, relevante Informationen aus fachsprachlichen Texten zu verstehen und zu entnehmen. Dieses Item erfasst, inwieweit die Lehrkraft die korrekte Verwendung von Fachsprache anregt und fördert.			
Indikatoren • Die LK verwendet Fachsprache adressatengerecht • Die LK korrigiert sprachliche Fehler adaptiv • Die LK initiiert die angemessene Verwendung von Fachsprache			
Anmerkungen/Hilfen zum Rating Für dieses Item wird ein MISSING vergeben, wenn es im Unterricht nicht explizit um die Entwicklung oder Sicherung von Begriffen geht, d.h. wenn die Fachsprache nicht aktiv gefördert wird.			
1	2	3	4
<i>Im Unterricht spielt die Förderung von Fachsprache keine Rolle.</i>	<i>Im Unterricht spielt die Förderung von Fachsprache eine geringe Rolle.</i>	<i>Im Unterricht spielt die Förderung von Fachsprache eine große Rolle.</i>	<i>Im Unterricht steht die Förderung von Fachsprache im Vordergrund.</i>

Modellieren			
Beschreibung des Items Quelle: Eigenentwicklung in Anlehnung an KMK (2012) Die Relevanz von Anwendungen und Modellierungen im Mathematikunterricht, insbesondere zur Förderung von mathematischen Modellierungskompetenzen, ist mittlerweile weithin akzeptiert. Dabei werden unter Modellieren alle Aspekte von Beziehungen zwischen der Mathematik und der Realität verstanden (Kaiser et al., 2015). In den deutschen Bildungsstandards ist das „Mathematische Modellieren“ eine der verpflichtenden allgemeinen mathematischen Kompetenzen. Es wird dort wie folgt beschrieben: <i>„Hier geht es um den Wechsel zwischen Realsituationen und mathematischen Begriffen, Resultaten oder Methoden. Hierzu gehört sowohl das Konstruieren passender mathematischer Modelle als auch das Verstehen oder Bewerten vorgegebener Modelle. Typische Teilschritte des Modellierens sind das Strukturieren und Vereinfachen gegebener Realsituationen, das Übersetzen realer Gegebenheiten in mathematische Modelle, das Interpretieren mathematischer Ergebnisse in Bezug auf Realsituationen und das Überprüfen von Ergebnissen im Hinblick auf Stimmigkeit und Angemessenheit bezogen auf die Realsituation. Das Spektrum reicht von Standardmodellen (z. B. bei linearen Zusammenhängen) bis zu komplexen Modellierungen.“ (KMK 2012, S. 17)</i> Das Item erfasst, inwieweit im Unterricht mit den Schülerinnen und Schülern mathematisch modelliert wird.			
Indikatoren • Die LK fördert Übersetzungsprozesse zwischen Realsituation und mathematischem Modell • Die LK fördert die Validierung der Ergebnisse im Hinblick auf die Realsituation • Es finden außermathematische Vernetzungen und fachübergreifende Zusammenhänge statt			
Anmerkungen/Hilfen zum Rating Wird im Unterricht nicht modelliert, so wird ein MISSING vergeben.			
1	2	3	4
Im Unterricht wird nicht mathematisch modelliert.	Im Unterricht wird in geringem Maße mathematisch modelliert.	Im Unterricht wird in hohem Maße mathematisch modelliert.	Im Unterricht steht die mathematische Modellierung im Vordergrund.

Problemlösen und Entdecken

Beschreibung des Items

Quelle: Eigenentwicklung in Anlehnung an KMK (2012)

Das mathematische Problemlösen ist eine der allgemeinen mathematischen Kompetenzen in den deutschen Bildungsstandards. Die Relevanz des Problemlösens für den Mathematikunterricht lässt sich bereits in den Winterschen Grunderfahrungen für den allgemeinbildenden Mathematikunterricht finden (Winter, 1996). Unter einem Problem versteht man eine Aufgabe oder Anforderung, bei deren Lösen eine Schwierigkeit bzw. Barriere überwunden werden muss. Darüber hinaus werden beim Problemlösen ebenfalls die Kreativität und das logische Denken der Lernenden gefördert und Erkenntnisse, die im Rahmen des Problemlösens erworben werden, gelten als besonders nachhaltig. Zu den Problemlöseaufgaben zählen auch Aufgaben, bei denen die Schülerinnen und Schüler neue mathematische Inhaltsbereiche entdecken. Unter einer methodischen Perspektive stellt das Problemlösen damit eine Lernmethode dar, mithilfe derer die Schülerinnen und Schüler weitere Lernziele erreichen können (Heinrich et al., 2015). Das Item erfasst, inwieweit im Mathematikunterricht mathematische Probleme gestellt werden, so dass beispielsweise neue mathematische Inhalte entdeckt werden können oder auch verschiedene heuristische Strategien ihre Anwendung finden und gefördert werden (vgl. KMK, 2012).

Indikatoren

- Die LK fördert die Verwendung heuristischer Strategien
- Die LK stellt mathematische Probleme
- Die LK fördert die Entdeckung neuer mathematischer Inhalte

Anmerkungen/Hilfen zum Rating

Werden im Unterricht keine mathematischen Probleme gelöst, so wird ein MISSING vergeben.

1	2	3	4
<i>Im Unterricht werden keine mathematischen Probleme gestellt.</i>	<i>Im Unterricht werden kleine mathematische Probleme gestellt. Insbesondere ist das Ausmaß an heuristischen Strategien gering, die zur Lösung der Aufgabe angewendet werden.</i>	<i>Im Unterricht werden mathematische Probleme gestellt. Insbesondere ist das Ausmaß an heuristischen Strategien hoch, die zur Lösung der Aufgabe angewendet werden.</i>	<i>Im Unterricht liegt der Fokus auf dem Lösen mathematischer Probleme bzw. der Entdeckung neuer mathematischer Inhalte.</i>

Begründen und Argumentieren			
Beschreibung des Items Quelle: Eigenentwicklung in Anlehnung an KMK (2012) Das mathematische Argumentieren ist eine der allgemeinen mathematischen Kompetenzen der deutschen Bildungsstandards. Es wird dort wie folgt beschrieben: <i>Zu dieser Kompetenz gehören sowohl das Entwickeln eigenständiger, situationsangemessener mathematischer Argumentationen und Vermutungen als auch das Verstehen und Bewerten gegebener mathematischer Aussagen. Das Spektrum reicht dabei von einfachen Plausibilitätsargumenten über inhaltlich-anschauliche Begründungen bis zu formalen Beweisen. Typische Formulierungen, die auf die Kompetenz des Argumentierens hinweisen, sind beispielsweise „Begründen Sie!“, „Widerlegen Sie!“, „Gibt es?“ oder „Gilt das immer?“. (KMK 2012, S. 14)</i> Das Item erfasst, inwieweit die Schülerinnen und Schüler im Unterricht die Möglichkeit erhalten, mathematisch zu argumentieren oder Sachverhalte zu beweisen (präformal oder formal). Hierzu zählt auch, dass die Schülerinnen und Schüler mathematische Begründungen für Lösungswege und Ergebnisse geben, die von der Lehrperson explizit eingefordert werden.			
Indikatoren • Die LK ermöglicht mathematische Argumentationen • Die LK fordert explizit mathematische Begründungen ein • Die LK thematisiert Beweise formaler oder präformaler Art			
Anmerkungen/Hilfen zum Rating Wird im Unterricht nicht argumentiert oder begründet, so wird ein MISSING vergeben.			
1	2	3	4
Im Unterricht spielen Begründungen keine Rolle.	Im Unterricht spielen Begründungen eine geringe Rolle.	Im Unterricht spielen Begründungen eine große Rolle (z.B. beim Beweisen).	Im Unterricht stehen Argumentationsprozesse oder Beweise im Vordergrund.

Mathematische Fertigkeiten			
Beschreibung des Items Quelle: Eigenentwicklung in Anlehnung an KMK (2012) Zum erfolgreichen Mathematiktreiben gehört auch ein Repertoire an verschiedenen mathematischen Fertigkeiten. Dies bedeutet, dass mathematische Operationen ausgeführt werden können (etwa der Umgang mit Variablen, Termen und Gleichungen) und dass verschiedene mathematische Verfahren beherrscht werden. Das hierbei zugrundeliegende Regelsystem der Mathematik muss hierfür im Mathematikunterricht erworben und gefestigt werden, so dass die Schülerinnen und Schüler in der Lage sind, dieses beim Bearbeiten von Aufgaben flexibel anwenden zu können. Dieser Aspekt findet sich ebenfalls in den deutschen Bildungsstandards unter der Kompetenz „mit symbolischen, formalen und technischen Elementen der Mathematik umgehen“ wieder, wobei in diesem Item die mathematischen Hilfsmittel und Werkzeuge ausgeschlossen werden, da hierfür ein weiteres eigenes Item existiert (KMK, 2012).			
Indikatoren • Es findet ein Training mit mathematischen Symbolen statt • Die LK ermöglicht das Training mathematischer Basistechniken			
Anmerkungen/Hilfen zum Rating Werden im Unterricht keine mathematischen Fertigkeiten trainiert, so wird ein MISSING vergeben.			
1	2	3	4
<i>Im Unterricht werden keine mathematischen Fertigkeiten trainiert.</i>	<i>Im Unterricht werden in geringem Maße mathematische Fertigkeiten trainiert.</i>	<i>Im Unterricht werden in hohem Maße mathematische Fertigkeiten trainiert.</i>	<i>Im Unterricht steht das Training mathematischer Fertigkeiten im Vordergrund.</i>

Mathematische Werkzeuge			
Beschreibung des Items Quelle: Eigenentwicklung in Anlehnung an KMK (2012) Um verschiedene Aufgaben im Mathematikunterricht erfolgreich und effizient bearbeiten zu können, ist es teilweise hilfreich oder sogar notwendig, mathematische Hilfsmittel und Werkzeuge korrekt zu verwenden. Hierzu zählt das Arbeiten mit z.B. Zirkel, Geodreieck und Lineal oder auch die Verwendung digitaler Mathematikwerkzeuge wie beispielsweise Taschenrechner, CAS oder auch dynamische Geometriesoftware (DGS). Neben der Nutzung dieser Hilfsmittel und Werkzeuge ist es im Rahmen eines kompetenzorientierten Mathematikunterrichts auch notwendig, die Möglichkeiten und Grenzen derselben zu thematisieren. Das Item erfasst, inwieweit mathematische Hilfsmittel und Werkzeuge im Mathematikunterricht korrekt und sinnvoll eingesetzt werden.			
Indikatoren • Die SuS setzen mathematische Werkzeuge sinnvoll ein • Die LK verwendet mathematische Werkzeuge korrekt			
Anmerkungen/Hilfen zum Rating Werden im Unterricht keine mathematischen Werkzeuge genutzt, so wird ein MISSING vergeben.			
1	2	3	4
<i>Im Unterricht werden keine mathematischen Werkzeuge genutzt.</i>	<i>Im Unterricht werden in geringem Maße mathematische Werkzeuge genutzt.</i>	<i>Im Unterricht werden in hohem Maße mathematische Werkzeuge genutzt.</i>	<i>Im Unterricht steht die Verwendung mathematischer Werkzeuge im Vordergrund.</i>

Relevanz			
Beschreibung des Items Quelle: adaptiert von Rakoczy & Pauli (2006), Kunter (2005), Kobarg & Seidel (2003) Dieses Item erfasst, inwieweit die Lehrperson versucht, den Schülerinnen und Schülern die persönliche Relevanz der im Unterricht behandelten Inhalte aufzuzeigen. Die Schülerinnen und Schüler sollen somit nicht das Gefühl bekommen, dass Mathematik ein abstraktes Unterrichtsfach ohne Relevanz und Lebensweltbezug sei. Hierzu zählt, dass die Lehrperson den Schülerinnen und Schülern den Alltagsbezug von Mathematik durch Beispiele, Materialien und vertraute Phänomene verdeutlicht, so dass die mathematischen Inhalte für die Schülerinnen und Schüler verständlicher und anschaulicher gemacht werden. Die Schülerinnen und Schüler können dabei eigene Erfahrungen und Interessen mit einbringen und diese diskutieren. Darüber hinaus kann die Lehrperson die persönliche Relevanz der Inhalte auch dadurch verdeutlichen, indem sie den Schülerinnen und Schülern aufzeigt, dass Mathematik vielfältige Anwendungsbezüge im Alltag hat und dort auch als Werkzeug verwendet werden kann (Klieme et al., 2006).			
Indikatoren • Es werden Bezüge zur Lebenswelt der Schülerinnen und Schüler aufgezeigt • Es wird thematisiert, wozu die Inhalte gut sind • Schülerinnen und Schüler können eigene Erfahrungen und Interessen einbringen			
Anmerkungen/Hilfen zum Rating Für die höchste Bewertung des Items ist es nicht notwendig, dass mehrere Aufgaben oder Materialien mit Relevanz und Alltagsbezug verwendet werden, sondern es geht hier um die Qualität. Eine „2“ wird vergeben, wenn die Lehrperson die inhaltliche Relevanz des Themas an den Abschlussprüfungen festmacht oder anderweitig <i>extern motiviert</i> .			
1	2	3	4
Eine Sinnkonstruktion findet nicht statt.	Genau ein Indikator trifft zu (eine Sinnkonstruktion findet in geringem Maße statt).	Genau zwei Indikatoren treffen zu (eine Sinnkonstruktion findet in hohem Maße statt).	Alle Indikatoren treffen zu.

Lehrervortrag (Erklärungen der Lehrkraft)			
Beschreibung des Items Quelle: Eigenentwicklung Dieses Item beschreibt, inwieweit es der Lehrkraft gelingt, verständlich und mathematisch korrekt zu erklären. Dies können kurze Erklärungen der Lehrkraft auf Nachfragen der Schülerinnen und Schüler sein, aber auch längere Lehrervorträge, die die Lehrkraft ggf. schon im Vorfeld vorbereitet hat. Die Erklärungen der Lehrkraft müssen dabei an die Lerngruppe angepasst sein (<i>adressatengerecht</i>). Ob die Erklärung der Lehrkraft auf die Lerngruppe passt, kann z.B. an der Reaktion der Lerngruppe erkannt werden oder ob mehrere Schülerinnen und Schüler mehrfach noch einmal nachfragen müssen. In diesem Item wird auch bewertet, inwieweit die Erklärungen der Lehrkraft gut strukturiert sind und ob beispielsweise an schwierigeren, für das Verständnis substanziellen Stellen langsamer und besonders präzise erklärt wird. Die Erklärungen der Lehrkraft sollten sich dabei nicht nur auf das „Wie“ beziehen sondern auch auf das „Warum“. Beispielsweise ist es nicht nur wichtig zu wissen, wie ein Verfahren funktioniert sondern auch insbesondere, warum es funktioniert.			
Indikatoren • Die LK erklärt besonders bei schwierigen Stellen langsam • Die LK beschränkt sich auf das Wesentliche • Die LK erklärt adressatengerecht • Die Erklärungen der LK sind klar strukturiert und präzise			
Anmerkungen/Hilfen zum Rating Geratet wird ggf. auch in Arbeitsphasen (nicht Plenum).			
1	2	3	4
Kein Indikator trifft zu.	Mehrere Indikatoren treffen nicht zu.	Mehrere aber nicht alle Indikatoren treffen zu.	Alle Indikatoren treffen zu.

Fachliche Tiefe			
Beschreibung des Items Quelle: Eigenentwicklung in Anlehnung an Hill et al. (2012), Blum et al. (2006) Dieses Item beschreibt die fachliche Tiefe des Mathematikunterrichts, der den Schülerinnen und Schülern angeboten wird. Zu dieser fachlichen Tiefe zählen innermathematische Vernetzungen zwischen verschiedenen Inhaltsbereichen und Themengebieten (im Sinne einer fachlichen Breite), mathematische Verallgemeinerungen (im Sinne einer mathematischen Höhe), aber auch die Verdichtung der mathematischen Inhalte durch beispielsweise neue Begrifflichkeiten und inhaltliche Erweiterungen (mathematische Tiefe). Hierzu müssen die mathematischen Inhalte im Unterricht gut strukturiert und geordnet werden und alle Elemente, die für einen erfolgreichen Strukturaufbau und für das Verständnis erforderlich sind, müssen thematisiert werden (bezeichnet als „Verstehenselemente“ nach Drollinger-Vetter, 2011).			
Indikatoren • Es finden Verallgemeinerungen statt • Es finden innermathematische Vernetzungen statt • Es wird strukturiert und geordnet • Mathematische Inhalte werden begrifflich verdichtet			
Anmerkungen/Hilfen zum Rating Es geht eher um das fachliche „Volumen“, also Breite (Vernetzung) <i>und</i> Höhe (Verallgemeinerung) <i>und</i> Tiefe (Verdichtung). „Nur“ Strukturieren ist eine „2“.			
1	2	3	4
<i>Es finden keine Strukturierungen, Verallgemeinerungen, Vernetzungen oder Verdichtungen statt.</i>	<i>Es wird lediglich strukturiert.</i>	<i>Zusätzlich zur Strukturierung wird vernetzt oder verdichtet oder verallgemeinert.</i>	<i>Es findet eine fachliche Tiefe in allen Bereichen (Verallgemeinern, Vernetzen, Verdichten) statt.</i>

Literaturverzeichnis

- Aebli, H. (1985). Das operative Prinzip. *Mathematik lehren*, 11(1985), 4-6.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., . . . Tsai, Y.-M. (2010). Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress. *American Educational Research Journal*, 47(1), 133-180.
- Blömeke, S., König, J., Busse, A., Suhl, U., Benthien, J., Döhrmann, M. & Kaiser, G. (2014). Von der Lehrerausbildung in den Beruf: Fachbezogenes Wissen als Voraussetzung einer genauen Wahrnehmung und Analyse von Unterricht. *Zeitschrift für Erziehungswissenschaft* 17(3), 509-542. doi:10.1007/s11618-014-0564-8
- Blömeke, S., Kaiser, G., & Lehmann, R. (Eds.) (2010). *TEDS-M 2008 - Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster: Waxmann.
- Blum, W., Drücke-Noe, C., Hartung, R., & Köller, O. (2006). *Bildungsstandards Mathematik: Konkret. Sekundarstufe 1: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen*. Berlin: Cornelsen Scriptor.
- Bruner, J. (1966). *Toward a theory of instruction*. Cambridge MA: Belknap Press.
- Clausen, M., Reusser, K. & Klieme, E. (2003). Unterrichtsqualität auf der Basis hoch inferenter Unterrichtsbeurteilungen. Ein Vergleich zwischen Deutschland und der deutschsprachigen Schweiz. *Unterrichtswissenschaft*, 31 (2), 122-141.
- Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research*, 77(1), 113-143.
- Den Brok, P., Brekelmans, M., & Wubbels, T. (2004). Interpersonal teacher behaviour and student outcomes. *School Effectiveness and School Improvement*, 15(3), 407-442.
- Drollinger-Vetter, B. (2011). *Verstehenselemente und strukturelle Klarheit: Fachdidaktische Qualität der Anleitung von mathematischen Verstehensprozessen im Unterricht*. Münster: Waxmann.
- Drollinger-Vetter, B., & Lipowsky, F. (2006). Fachdidaktische Qualität der Theoriephasen. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 3 Videoanalysen)* (S. 206-233). Frankfurt am Main: GFPP/DIPF.
- Fraser, B. J. (1991). Two decades of classroom environment research. In H. J. Walberg (Ed.), *Educational environments: Evaluation, antecedents and consequences* (pp. 3-27). Elmsford, NY: Pergamon.
- Freie und Hansestadt Hamburg, Behörde für Schule und Berufsbildung (2011). *Bildungsplan Gymnasium Sekundarstufe I. Mathematik*.

- Hasselhorn, M., & Labuhn, A. S. (2008). Metakognition und selbstreguliertes Lernen. In W. Schneider, & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (S. 28–37). Göttingen: Hogrefe.
- Heinrich, F., Bruder, R., & Bauer, C. (2015). Problemlösen lernen. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme, & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 357-384). Berlin Heidelberg: Springer.
- Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze: Klett-Kallmeyer.
- Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Loewenberg Ball, D. (2008). Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study, *Cognition and Instruction*, 26(4), 430-511.
- Hill, H., Charalambous, C. Y., & Kraft, M. A. (2012). When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, 41(2), 56-64.
- Kaiser, G., Blum, W., Borromeo Ferri, R., & Greefrath, G. (2015). Anwendungen und Modellieren. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme, & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 357-384). Berlin Heidelberg: Springer.
- Klafki, W. (1963). Das pädagogische Problem des Elementaren und die Theorie der kategorialen Bildung. Weinheim.
- Klieme, E., Pauli, C. & Reusser, K. (Hrsg.) (2006). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 3 Videoanalysen)*. Frankfurt am Main: GFPP/DIPF.
- Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54, 222–237.
- Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur“ und Unterrichtsgestaltung. In E. Klieme & J. Baumert (Hrsg.), *TIMSS - Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente*. Bonn: Bundesministerium für Bildung und Forschung.
- KMK (Kultusministerkonferenz) (2012). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife* (Beschluss der Kultusministerkonferenz vom 18.10.2012).
- Kobarg, M., & Seidel, T. (2003). Prozessorientierte Lernbegleitung im Physikunterricht. In T. Seidel, M. Prenzel, R. Duit & M. Lehrke (Hrsg.), *Technischer Bericht zur Videostudie „Lehr-Lern-Prozesse im Physikunterricht“* (S. 151-200). Kiel: IPN.

- Kounin, J. S. (1970). *Disciplin and group management in classrooms*. New York: Holt, Rinehart & Winston.
- Kounin, J. S. (1976). *Techniken der Klassenführung*. Bern, Stuttgart: Hubert/Klett.
- Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht*. Münster: Waxmann.
- Kunter, M. & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. In M. Kunter, J. Baumert & W. Blum (Hrsg.), *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV* (S. 85-113). Münster: Waxmann.
- Learning Mathematics for Teaching Project (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14, 25–47.
- Leuders, T. (2009). Intelligent üben und Mathematik erleben. In T. Leuders, L. Hefendehl-Hebeker, & H.-G. Weigand (Hrsg.), *Mathemagische Momente* (S. 130–143). Berlin: Cornelsen.
- Lipowsky, F. (2015). Unterricht. In E. Wild, & J. Möller (Eds.), *Pädagogische Psychologie* (2. Auflage) (S. 69-105). Berlin, Heidelberg: Springer.
- Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction*, 19(6), 527–537. doi:10.1016/j.learninstruc.2008.11.001
- Pea, R. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences*, 13, 423–451.
- Rakoczy, K., & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis" (Teil 3 Videoanalysen)* (S. 206-233). Frankfurt am Main: GFPF/DIPF.
- Reinmann-Rothmeier, G. & Mandl, H. (2006). Unterrichten und Lernumgebungen gestalten. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie. Ein Lehrbuch* (S. 613-658). Weinheim: Beltz.
- Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences*, 13(3), 273–304.
- Reusser, K. (1998). Denkstrukturen und Wissenserwerb in der Ontogenese. In F. Klix & H. Spada (Hrsg.), *Enzyklopädie der Psychologie* (Themenbereich C; Serie II, Bd. 6) (S. 115-166). Göttingen: Hogrefe.

- Seidel, T. (2003). Überblick über Beobachtungs- und Codierungsverfahren. In T. Seidel, M. Prenzel, R. Duit & M. Lehrke (Hrsg.), *Technischer Bericht zur Videostudie „Lehr-Lern-Prozesse im Physikunterricht“*. Kiel: IPN.
- Seidel, T., & Shavelson, R. J. (2007). Teaching Effectiveness Research in the Past Decade: The Role of Theory and Research Design in Disentangling Meta-Analysis Results. *Review of Educational Research*, 77(4), 454–499.
doi:10.3102/0034654307310317
- Shuell, T. J. (1993). Toward an integrated theory of teaching and learning. *Educational Psychologist*, 28, 291–311.
- Stefanou, C. R., Perencevich, K. C., DiCintio, M., & Turner, J. C. (2004). Supporting autonomy in the classroom: Ways teachers encourage student decision making and ownership. *Educational Psychologist*, 39(2), 97–110.
- Trepke, C., Seidel, T. & Dalehefte, I.M. (2003). Zielorientierung im Physikunterricht. In T. Seidel, M. Prenzel, R. Duit & M. Lehrke (Hrsg.), *Technischer Bericht zur Videostudie „Lehr-Lern-Prozesse im Physikunterricht“* (S. 201–228). Kiel: IPN.
- Turner, J. C., Meyer, D. K., Cox, K. E., Logan, C., DiCintio, M., & Thomas, C. T. (1998). Creating contexts for involvement in mathematics. *Journal of Educational Psychology*, 90(4), 730–745.
- Wagenschein, M. (1999). *Verstehen lehren: Genetisch - Sokratisch – Exemplarisch*. Weinheim, Basel: Beltz.
- Winter, H. (1996). Mathematik und Allgemeinbildung. *Mitteilungen der Gesellschaft für Mathematik*, 61, 37–46.
- Wittmann, E. C. (1992). Wider die Flut der bunten Hunde und der grauen Päckchen: Die Konzeption des aktiv entdeckenden Lernens und produktiven Übens. In G. N. Müller & E. C. Wittmann (Hrsg.), *Handbuch produktiver Rechenübungen* (S. 152–166). Stuttgart: Klett.
- Wood, D., Bruner, J. S. & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology & Psychiatry & Allied Disciplines*, 17(2), 89–100.
- Zech, F. (1998). *Grundkurs Mathematikdidaktik. Theoretische und praktische Anleitungen für das Lehren und Lernen von Mathematik* (9. Aufl.). Weinheim: Beltz.

E Auflistung der Einzelarbeiten der Dissertation

Die vorliegende publikationsbasierte Dissertation besteht aus drei Zeitschriften-Artikeln.

Artikel 1:

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations, *ZDM Mathematics Education*, 48(1), 29-40.

Artikel 2:

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (under review). Unterrichtsqualität unter einer mathematikdidaktischen Perspektive – Konzeptualisierung, Messung und Validierung. *Journal für Mathematik-Didaktik*.

Artikel 3:

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *ZDM Mathematics Education*, 50(3). Doi: 10.1007/s11858-018-0917-5

F Liste der Publikationen von Lena Schlesinger

- Schlesinger, L. & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM Mathematics Education*, 48(1), 29-40.
- Jentsch, A. & Schlesinger, L. (2016). Mathematikdidaktische Unterrichtsqualität – Herausforderungen bei Konzeption und Messung eines theoretischen Konstrukts. In Institut für Mathematik und Informatik Heidelberg (Hrsg.), *Beiträge zum Mathematikunterricht 2016* (S. 493-496). Münster: WTM-Verlag.
- Schlesinger, L., Benecke, K. & Jentsch, A. (2016). Unterrichtsbeobachtungen zur Messung der Unterrichtsqualität im Rahmen der Studie TEDS-Unterricht. In Institut für Mathematik und Informatik Heidelberg (Hrsg.), *Beiträge zum Mathematikunterricht 2016* (S. 1519-1520). Münster: WTM-Verlag.
- Jentsch, A. & Schlesinger, L. (2017). Fachspezifische Unterrichtsqualität im Mathematikunterricht. In U. Kortenkamp & A. Kuzle (Hrsg.), *Beiträge zum Mathematikunterricht 2017* (S. 501-504). Münster: WTM-Verlag.
- Jentsch, A., & Schlesinger, L. (2017). Measuring instructional quality in mathematics education. In T. Dooley & G. Gueudet (Eds.), *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education* (CERME10, February 1 – 5, 2017). Dublin, Ireland: DCU Institute of Education & ERME.
- Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *ZDM Mathematics Education*, 50(3). Doi: 10.1007/s11858-018-0917-5
- Schlesinger, L., & Jentsch, A. (angenommen). Zur faktoriellen Struktur fachspezifischer Unterrichtsqualität im Mathematikunterricht. In Institut für Mathematik der Universität Paderborn (Hrsg.), *Beiträge zum Mathematikunterricht 2018*. Münster: WTM-Verlag.
- Jentsch, A., Schlesinger, L., Kaiser, G., Suhl, U., König, J., & Blömeke, S. (angenommen). Zusammenhänge zwischen professionellen Kompetenzen von Mathematiklehrkräften und fachspezifischer Unterrichtsqualität. In Institut für Mathematik der Universität Paderborn (Hrsg.), *Beiträge zum Mathematikunterricht 2018*. Münster: WTM-Verlag.
- Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (under review). Unterrichtsqualität unter einer mathematikdidaktischen Perspektive – Konzeptualisierung, Messung und Validierung. *Journal für Mathematik-Didaktik*.
- König, J., Kaiser, G., Blömeke, S., Jentsch, A., Schlesinger, L., Nehls, C., & Suhl, U. (under review). Teaching and learning in the lower secondary mathematics classroom:

Analyzing the links between pedagogical competence, instructional quality, and student achievement. *Learning and Instruction*.

Erklärung über die Eigenständigkeit der Dissertation

Eidesstattliche Versicherung

Hiermit versichere ich, dass die Dissertation von mir selbst verfasst wurde, und dass keine anderen als die angegebenen Hilfsmittel von mir genutzt wurden.

14. Mai 2018

(Signatur entfällt aus datenschutzrechtlichen Gründen.)

Datum

Unterschrift

Erklärung

Ich erkläre hiermit, dass ich keine kommerzielle Promotionsberatung in Anspruch genommen habe, und dass ich mich bisher keiner weiteren Doktorprüfung unterzogen habe. Insbesondere habe ich die Dissertation in der gegenwärtigen oder einer anderen Fassung an keiner anderen Fakultät eingereicht.

14. Mai 2018

(Signatur entfällt aus datenschutzrechtlichen Gründen.)

Datum

Unterschrift