Using TIMSS 2015 data to compare educational effectiveness-enhancing factors in the countries of the Gulf Cooperation Council (GCC)

Dissertation

zur Erlangung des akademischen Grades eines

Doktors der Philosophie (Dr. phil.)

an der Fakultät für Erziehungswissenschaft der Universität Hamburg

vorgelegt von

Oliver Neuschmidt aus Hamburg

Hamburg, 17.9.2018

Erstgutachter: Prof. Dr. Knut Schwippert
(Universität Hamburg)

Zweitgutachter: Prof. Dr. Tobias C. Stubbe
(Universität Göttingen)

Mündlicher Prüfer: Prof. Dr. Jan Retelsdorf
(Universität Hamburg)

Disputation: Universität Hamburg, 17. April 2019

Abstract

The countries of the Gulf Cooperation Council (GCC) are currently experiencing extensive and rapid economic and social changes while transforming from traditionally-oriented oil monarchies into knowledge societies. While during the last decades quantitative dimensions of schooling were vastly improved, qualitative dimensions of education are still lagging behind, as all GCC countries are located in the lowest quartile of the mathematics and science scales in the international comparative assessment TIMSS 2015. Additionally, the region still shows large disparities in terms of gender in favor of girls, but also between the national and the – largely higher-achieving – foreign populations. The current research project is based on secondary analyses of the TIMSS 2015 data, with the objective of identifying factors explaining achievement similarities and differences in the region in terms of mathematics and science outcomes. For this purpose, a research framework was built, which concurrently aims to take into account the special conditions in the Gulf area and the restrictions inherent in using cross-sectional large-scale assessment data for educational effectiveness research. Two main questions were formulated to attain the research objectives. 1.) To what extent does TIMSS 2015 reflect essential factors in terms of educational effectiveness research? To answer this question, data from the TIMSS 2015 background questionnaires were matched to the model factors of the research framework. Principal component, reliability, and correlation analyses with mathematics and science outcomes were used to specify a regional model of important factors in a parsimonious way. While the strength of the correlations between model variables and outcomes varied by country and subject, results indicated that TIMSS 2015 can be used to obtain a sufficient coverage of the research framework in the region. 2.) According to the framework specified, which educational factors are most effective from the perspective of EER with regard to learning outcomes on primary level in the GCC countries? To answer this question, this study used multilevel modeling techniques to deconstruct the total achievement variance into within- and between-course/school level parts. Student background factors emerged as the strongest predictors of achievement in all six countries, with the background model explaining most of the between-group variance. On the course- and school-level, clear and structured instruction, and the amount of teaching time, emerged as the most consistent factors across the region but a regional pattern in terms of common factors could not be discerned. The final models explain between 27% of the level-2 variance in Oman and 46% in Qatar for mathematics, and between 24% in Oman and 51% in Qatar for science achievement.

Keywords: GCC countries; educational effectiveness; TIMSS 2015; mathematics; science; multilevel modeling

Acknowledgements

I would like to especially thank my supervisor, Prof. Dr. Knut Schwippert for supporting me over the long period needed to do a doctorate next to a full-time job.

I also would like to give special thanks to my colleagues from the IEA research and analysis unit, Dr. Sabine Meinck and Dr. Agnes Stanzel-Piatak, who are very knowledgeable discussion partners who supported me with valuable advice in difficult phases of the project.

I also want to express my sincere gratitude to all those who crossed my path and in one or the other way and helped me during my research:

David and Leslie Rutkowski, who taught me how to write academically and supported me a lot in laying the statistical foundations needed for doing research in educational effectiveness. Still remembering their ‘brown bags’ where important research papers were discussed over a lunch break.

Clara Wilsher Beyer, who thoroughly edited my work and transferred it to proper English.

My son Hannes, who helped me a lot with a final review of all the tables and graphics and corresponding links.

My colleagues in the International Studies Unit who took over many of my duties while I was absent to work on my research project.

My wife, who had to endure spending quite a bit of family time without me and interestedly read my work and commented on it.

Table of Contents

Abstract V

Acknowledgements VI

Table of Contents VII

List of Figures XI

List of Tables XII

List of Abbreviations XV

1 INTRODUCTION 1

1.1 Introducing the Study 1

1.2 Structure of the Dissertation 4

2 EDUCATIONAL CONTEXT IN THE GCC COUNTRIES 5

2.1 Introduction 5

2.2 The Schooling System in the GCC Countries 11

2.3 The Curricula in the GCC Countries 13

2.4 Achievement of GCC Countries in International Large-Scale Assessments 15

3 RESEARCH ON FACTORS INFLUENCING STUDENT PERFORMANCE 17

3.1 Educational Effectiveness Research 17

3.1.1 Strands of educational effectiveness research 17

3.1.2 International comparative studies and educational effectiveness 19

3.2 Educational Effectiveness Research (EER) – Definitions 22

3.3 General Effectiveness Factors 30

3.3.1 Introduction 30

3.3.2 Time on task 33

3.3.3 Opportunity to learn 37

3.3.4 Student-level factors 40

3.3.5 Class-level factors 51

3.3.6 School-level factors 63

3.3.7 Context-level factors 69

3.4 School Effects 71

3.4.1 Existence of school effects 71

3.4.2 Magnitude of school effects 73

3.4.3 Consistency of school effects 77

3.4.4 Stability of school effects over time 78

3.4.5 Differential effects 79

3.4.6 Composition effects 80

4 MODELS OF EDUCATIONAL EFFECTIVENESS 83
4.1 Theoretical Foundations 83
4.2 Integrated Models of Educational Effectiveness 84
4.2.1 Scheerens' Model 85
4.2.2 Creemers' Model 87
4.2.3 The dynamic model of educational effectiveness 92

5 PROBLEM SETTING AND RESEARCH QUESTIONS 99
5.1 Problem Setting 99
5.2 Rationale for the Study 102
5.3 Aims of the Study and Research Questions 103

6 CONCEPTUAL FRAMEWORK OF THE STUDY 106
6.1 Introduction 106
6.2 Developing the Framework 106

7 IEA LARGE-SCALE ASSESSMENT TIMSS 2015 114
7.1 Introduction 114
7.2 Design and Framework of the TIMSS assessment 115
7.3 Background Instruments 116
7.4 The TIMSS Sample 117
7.5 The TIMSS Achievement Scores 119
7.6 TIMSS Data Quality Considerations 121
7.6.1 Objectivity 121
7.6.2 Reliability 121
7.6.3 Validity 122

8 RESEARCH DESIGN AND METHODS 124
8.1 Introduction 124
8.2 Using TIMSS for Educational Effectiveness Research 124
8.2.1 Secondary analysis of data 124
8.2.2 Using large-scale assessment data for educational effectiveness research 125
8.3 Data Analysis 128
8.3.1 Preliminary analyses related to disparities in terms of gender and nationality status 128
8.3.2 Identifying variables related to the proposed framework 128
8.3.3 Matching school, teacher, and student data 130
8.3.4 Preparing and exploring the data sets 131
8.3.5 Missing data 132
8.3.6 Data reduction procedures 134
8.4 Creating an Index of Economic Social and Cultural Status (ESCS) 137
8.5 Multilevel Analysis 143
8.5.1 General characteristics 144

8.5.2 Building the models 145

9 RESULTS OF VARIABLE SELECTION AND FACTOR/ RELIABILITY ANALYSES 153
9.1 Results of the Preliminary Analyses Related to Gender and Nationality 153
9.2 Variable Selection and Categorization 160
9.3 Results from the Principal Component and Reliability Analyses 161
9.3.1 Student level 162
9.3.2 Course level 165
9.3.3 School level 169
9.4 Results of the Correlation Analyses 172
9.4.1 Student level 173
9.4.2 Course and school level 175
9.4.3 Final components kept for multilevel analyses 180
9.5 Summary 182

10 RESULTS OF THE MULTI LEVEL ANALYSES 184
10.1 The Null Model 185
10.1.1 Mathematics 185
10.1.2 Science 186
10.2 The Level-1 Student Background Models 187
10.2.1 Mathematics 187
10.2.2 Science 189
10.3 The Student Background Models Including Student Composition Variables on Level 2 190
10.3.1 Mathematics 190
10.3.2 Science 191
10.4 School- and Course-Level Effectiveness Variables Without Controlling for the Student Background 192
10.4.1 Mathematics 193
10.4.2 Science 194
10.5 School Effectiveness Variables after Controlling for the Student Background 195
10.5.1 Mathematics 196
10.5.2 Science 198
10.6 Summary 199

11 DISCUSSION AND CONCLUSIONS 202
11.1 Introduction 202
11.2 Summary 202
11.3 Answering the Research Questions 204
11.4 Discussion 225
11.5 Contribution to Scientific and Practical Knowledge 235
11.6 Recommendations 236

11.6.1 Policy recommendations for the region 236
11.6.2 Recommendations regarding the TIMSS assessment in regard to educational effectiveness research 238
11.6.3 Recommendations concerning further research on educational effectiveness in the region 240
11.7 Limitations 241
11.8 Further Research 242
11.9 Conclusion 243
12 SUMMARIES 245
12.1 English Summary 245
12.2 German Summary 249
REFERENCES 255
APPENDIX A: THE TIMSS 2015 QUESTIONNAIRES 280
APPENDIX B: INDICATORS AND VARIABLE RECODING 284
APPENDIX C: ADDITIONAL ANALYSES 288
APPENDIX D: VARIANCE COMPONENTS FOR GROUP-CENTERED APPROACH 289

List of Figures

Figure 3-1: Basic system model on the functioning of education (from Scheerens, 2016, p. 6) 23

Figure 4-1: Integrated model of school effectiveness from Scheerens (1992, p. 14).....86

Figure 4-2: Creemers' comprehensive model of educational effectiveness – overview (taken Creemers, 1994, p. 27).....88

Figure 4-3: Main Characteristics of the Dynamic Model (from Creemers & Kyriakides, 2008, p. 150).....94

Figure 4-4: Factors of the dynamic model operating at student level and their assumed interrelation (from Creemers & Kyriakides, 2008, p. 94).....97

Figure 6-1: Proposed model of educational effectiveness – Summary 107

List of Tables

Table 2-1: Overview on demographics and selected indicators relevant to primary education 6
Table 2-2: Overview on primary and secondary education in the GCC countries 12
Table 2-3: Percentage of private enrollment in primary and percentage of non-nationals 12
Table 2-4: Number of TIMSS topics intended to be taught by end of grade and Test Curriculum Matching Analysis 14
Table 2-5: Participation of GCC countries in international large-scale assessments 15
Table 2-6: Overall mathematics and science scores of the GCC countries in 2011 and 2015 16
Table 3-1: Range of stability estimates (correlation coefficients) for school effects taken from Luyten (1994) 78
Table 4-1: Factors of Creemers' comprehensive model of educational effectiveness – detailed version (from Creemers, 1994, p. 119) 91
Table 4-2: Main elements of the dynamic model on the teaching level (summary taken from (Chapman et al., 2015, p. 116) 96
Table 6-1: Details of the factors for the proposed model of educational effectiveness 113
Table 7-1: TIMSS 2015 sample sizes in the GCC countries (own calculations) 118
Table 8-1: Percentages of students linked to more than one teacher 130
Table 8-2: Match between TIMSS occupation categories (Variables ASBH23A/B) and ISEI scores following the procedure of Caro and Cortés (2012) 140
Table 8-3: Match between ISCED, TIMSS educational categories, and years of schooling 141
Table 8-4: Correlation between SES variables and mathematics achievement 143
Table 8-5: Correlation between SES variables and science achievement 143
Table 9-1: TIMSS 2015 average mathematics achievement by gender 154
Table 9-2: TIMSS 2015 average science achievement by gender 154
Table 9-3: TIMSS 2015 average mathematics achievement by nationality status 155
Table 9-4: TIMSS 2015 average science achievement by nationality status 155
Table 9-5: TIMSS 2015 average mathematics achievement by nationality status and gender 157
Table 9-6: TIMSS 2015 average science achievement by nationality status and gender 157
Table 9-7: TIMSS 2015 average mathematics achievement by type of school (mixed versus segregated in terms of immigrant status) 158
Table 9-8: TIMSS 2015 average science achievement by type of school (mixed versus segregated in terms of immigrant status) 158
Table 9-9: TIMSS 2015 average mathematics achievement by teachers' gender and school type (mixed versus single-sex) 159
Table 9-10: TIMSS 2015 average science achievement by teachers' gender and school type (mixed versus single-sex) 159
Table 9-11: Questions and options selected from the TIMSS 2015 questionnaires 161
Table 9-12: Economic and social cultural status 163
Table 9-13: Early numeracy skills 163
Table 9-14: Subject motivation mathematics 164
Table 9-15: Subject motivation science 164

Table 9-16: Confidence in teaching – mathematics 165
Table 9-17: Confidence in teaching – science 166
Table 9-18: Emphasis on academic success 166
Table 9-19: Clear and structured instruction – mathematics 167
Table 9-20: Clear and structured instruction – science 167
Table 9-21: Cognitive activation 168
Table 9-22: Class environment 168
Table 9-23: Shortage in mathematics resources 169
Table 9-24: Shortage in science resources 170
Table 9-25: Emphasis on academic success (school level) 170
Table 9-26: School discipline and safety 171
Table 9-27: Absenteeism 171
Table 9-28: Correlation of student level factors with mathematics achievement 174
Table 9-29: Correlation of student level factors with science achievement 174
Table 9-30: Correlation of course and school level factors with mathematics achievement 175
Table 9-31: Correlation of course and school level factors with science achievement 176
Table 9-32: Final mathematics indicators selected for multilevel analyses and their means and standard deviations 181
Table 9-33: Final science indicators selected for multilevel analyses and their means and standard deviations 181
Table 10-1: Null models for mathematics 186
Table 10-2: Null models for science 187
Table 10-3: Student background models – mathematics 189
Table 10-4: Student background models – science 190
Table 10-5: Student background models including composition – mathematics 191
Table 10-6: Student background models including composition – science 192
Table 10-7: Course/ school level model without controlling – mathematics 194
Table 10-8: Course/ school level model without controlling – science 195
Table 10-9: Course/ school-level model with controlling – mathematics 197
Table 10-10: Course/ school level model with controlling – science 198
Table 11-1: Factors identified from questionnaires according to the specified framework 206
Table 11-2 : Significant indicators using mathematics and science as outcome variables 210
Table 11-3: Mathematics variance components and variance explained on group level 225
Table 11-4: Science variance components and variance explained on group level 225
Table A-1: Content of the TIMSS 2015 grade 4 student questionnaire 280
Table A-2: Content of the TIMSS 2015 grade 4 parent questionnaire 281
Table A-3: Content of the TIMSS 2015 grade 4 teacher questionnaire 282
Table A-4: Content of the TIMSS 2015 grade 4 school questionnaire 283
Table B-1: Results from reliability analyses for the created indices 286
Table B-2: Further recodings 287

Table C-1: Mathematics results by parental education level 288
Table C-2: Science results by parental education level 288
Table D-1: Variance components for the mathematics models 289
Table D-2: Variance components for the science models 289

List of Abbreviations

ARE	United Arab Emirates
BHR	Bahrain
CTT	Classical test theory
EER	Educational effectiveness research
ESCS	Economic, Social, and Cultural Status (index)
GCC	Gulf Cooperation Council
KWT	Kuwait
IEA	International Association for the Evaluation of Educational Achievement
IRT	Item response theory
ISEI	International Socio-Economic Index of Occupational Status
ISCED	International Standard Classification of Education
ISCO	International Classification of Occupations
KMO	Kaiser-Meyer-Olkin
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
NRC	National Research Coordinator
OECD	Organisation for Economic Co-operation and Development
OMN	Oman
PCA	Principal component analysis
PIRLS	Progress in International Reading Literacy Study

PISA	Programme for International Student Assessment
QAT	Qatar
SAU	Saudi Arabia
SAS	Statistical Analysis System (analysis software)
S.E.	Standard error
SER	School effectiveness research
SES	Socio-economic status
SLE	School learning environment
SPSS	Statistical Packages for the Social Sciences (software for data analysis)
STEM	Science, technology, engineering, and mathematics
TCMA	Test-curriculum matching analysis
TER	Teacher effectiveness research
TIMSS	Trends in International Mathematics and Science Study
UN	United Nations
UNDP	United Nations Development Programme
UNESCO	United Nations Educational, Scientific, and Cultural Organization
UIS	UNESCO Institute for Statistics

1 INTRODUCTION

1.1 Introducing the Study

Competencies in mathematics and science are regarded as an important precondition for economic development around the world; corresponding research indicates a strong relationship between cognitive skills and economic growth (Baker, Goesling, & LeTendre, 2002; Hanushek & Woessmann, 2008; Schofer, Ramirez, & Meyer, 2000). Hanushek and Woessmann (2008, p. 607), in their analysis of international comparative assessments of mathematics, science, and reading, concluded "that there is strong evidence that the cognitive skills of the population–rather than mere school attainment–are powerfully related to individual earnings, to the distribution of income, and to economic growth." Moreover, important international organizations such as the United Nations focus explicitly not only on the quality, but also on the equity, of education – as stated in their Sustainable Development Goals (United Nations, 2015). Along these lines, the OECD (2012, p. 3) also stated: "The highest performing education systems are those that combine equity with quality."

The Gulf Cooperation Council (GCC) countries show many similarities in terms of their social and cultural values, religion, and language. Due to the wealth accumulated from the export of natural resources, and the resulting rapid economic development, they have experienced tremendous transformations in almost all aspects of socio-economic life (Bahgat, 1999; Mansour & Al-Shamrani, 2015). These developments have likewise impacted the education sector, in which fast developments in terms of quantitative dimensions of schooling were achieved in a short period of time. However, these rapid developments led to an imbalance between fast economic growth and social development – which can only change at a slower pace. Due to a lack of skilled labor force in the GCC countries, the region heavily depends on a foreign workforce; in most of the GCC countries, foreigners represent more than half of the population. These developments also led to a number of other societal distortions such as a mismatch between traditional and modern schooling, an imbalance between national and foreign workers, and a rising gender gap (Bahgat, 1999, p. 129). Dwindling revenues from oil and gas now force countries in the region to diversify their economies and to follow the "knowledge economy road map laid out by international development agencies" (Weber, 2011, p. 2592) in order to become more competitive on the global market. With the quantitative dimension of schooling, such as enrollment and staffing, addressed during the last decades, the next wave of modernization programs targets the quality of education. In this context, a rising interest in monitoring educational outcomes and policy reforms has emerged in the region.

Here, international large-scale assessments play a major role, as they allow for the assessment of several different subjects and the investigation of associated contextual factors of school learning. One of the foremost international assessments in educational research is the Trends in International Mathematics and Science Study (TIMSS), conducted by the International Association for the Evaluation of Educational Achievement (IEA). The IEA is a non-profit organization aiming to help their member countries "understand effective practices in education and develop evidence-based policies to improve education" ("About Us | IEA," 2018). The TIMSS assessment is administered every four years, and the most recent administration in 2015 was administered in 57 countries and 7 benchmarking entities. All six Gulf Cooperation Council (GCC) countries participated in both target grades (grade four and grade eight) of the assessment. Despite major improvements made in some countries in the region during the last years, GCC countries still appear in the lowest quartile of the TIMSS achievement scales for both mathematics and science. Nonetheless, the achievement gap between the highest- and the lowest-achievement GCC country amounts to more than one standard deviation for science (see Table 2-6 for an overview on the mathematics and science achievement of the GCC countries). Additionally, the region exhibits large disparities in terms of gender and nationality status (Neuschmidt, 2016; Neuschmidt & Tölle, 2017).

A special motivation for this project originates from a seminar series conducted in different Arab countries between 2006 and 2007. The purpose of the seminar series was "to provide the participants with the training and skills necessary to permit them to conduct secondary analysis of their national [TIMSS 2007] datasets" (Lietz, Wagemaker, Neuschmidt, & Hencke, 2008). The common interest of nearly all the seminar participants from ministries of education in the region was to identify "malleable factors"; this resulted, in part, in research projects to identify common characteristics of effective schools. It became apparent that the analyses that could be conducted during the seminar series would not be sufficiently comprehensive to explain the achievement differences and large disparities found, which triggered the interest of the author to investigate further, basing the reasearch on a solid theoretical framework.

Literature review revealed that comprehensive investigations on educational factors affecting student outcomes, which were increasingly guided by theoretical underpinnings under the paradigm of educational effectiveness research (EER), had been undertaken in the Western hemisphere and later on in Asia; corresponding analyses in the Gulf States, however, were still missing. The Gulf region appeared to the researcher to be an especifically interesting target for further educational effectiveness research, as educational conditions in terms of historical development, culture, and political conditions are very different from educational conditions in the

West, where most of the existing educational effectiveness research had been undertaken. Moreover, the region is characterized by a certain homogeneity in terms of history, culture, language and the fact that they all accumulated large wealth through their oil and gas exports in a relatively short period of time; this allowed for the investigation of the extent to which identified educational factors work in a similar manner across the region.

The field of EER, which also should guide the current research project, started to develop around five decades ago predicated on the findings of Coleman et al. (1966, p. 325) that "schools bring little influence to bear on a child's achievement that is independent of his background and general social context." Nowadays, there is a widespread consensus among researchers that schools do indeed affect student achievement, both directly and indirectly (Chapman, Muijs, Reynolds, Sammons, & Teddlie, 2015; Mortimore, Sammons, Stoll, Lewis, & Ecob, 1988; Teddlie & Reynolds, 2000). EER has led to the development of theories and models that help explain differences among schools and other educational levels, and as such give indications for the effectiveness of schools or educational systems. While definitions of educational effectiveness have changed considerably over the past decades, the majority of studies on effectiveness research still utilize standardized achievement test results in core subjects – such as reading, mathematics, or science – as their outcome variables. More recent studies, which are based on cross-sectional data, try to disentangle organizational and instructional school practices from the effects of the student's home environment, in order to analyze school and classroom specific value-added effects, which will also be the approach in the current research project.

With the participation of all GCC countries in TIMSS, a more regional approach in the analysis of educational effectiveness factors in the Arab Region, from the perspective of EER, becomes available. TIMSS data is not specifically designed for the detection of educational effectiveness factors, but is rather framed to address multiple purposes – such as to obtain in-depth knowledge regarding different systems of educations' implemented policies and practices, and to provide robust and high-quality data for trend analyses (Martin & Mullis, 2013; Teddlie & Reynolds, 2000). Using TIMSS data for effectiveness research consequently also raises criticism.

Taking this criticism into account, the researcher will argue that given the absence of suitable, internationally comparable, longitudinal data on a school- and student level the use of large-scale assessment data in exploring educational effectiveness concepts is justified to a certain extent. Moreover, when applied in other areas of the world, these studies may expand the knowledge related to the international dimension of effectiveness research and add empirical

evidence for the generalizability and validity of models and constructs, even if findings must be interpreted with caution due to limitations in the availability of suited indicators, the cross-sectional structure, and so forth.

Thus, the current research will focus on the effectiveness of mathematics and science instruction in the GCC countries, based on a framework rooted in EER, which also endeavors to take into account the special conditions in the region under consideration as well as certain limitations occurring due to the use of the comparative large-scale assessment data at hand.

1.2 Structure of the Dissertation

The following three chapters will summarize findings from the relevant literature review for this study. Chapter 2 will present the educational context of the region, while chapter 3 will summarize the findings on educational effectiveness research. Chapter 3 will provide definitions of effectiveness, give an overview on important effectiveness factors on different educational levels, and also describe the important concepts of *time on task* and *opportunity to learn*. Finally, different properties of school effects will be discussed. Chapter 4 subsequently will present an overview on important models and constructs of effectiveness that were used as a base for the theoretical framework developed for the current research project. Chapter 5 will lay out the problem setting, the research objectives, and describe the research questions posed for this study. Based on the outcomes of the literature review, the conceptual framework will be developed in chapter 6. Chapter 7 will then introduce the TIMSS 2015 assessment and discuss issues of objectivity, validity, and reliability. Chapter 8 will describe the research design and the research methods applied. Here, the implications for and limitations of using cross-sectional assessment data for educational effectiveness will be discussed, followed by a description of the data preparation and the data reduction procedures. A separate section is devoted to the development of a home background index, and finally the multilevel analyses steps are described. The results of the variable selection process, as well as for the factor-, reliability, and correlation analyses can be found in chapter 9, while the results of the ultimate are presented in chapter 10. Chapter 11 then covers discussion, policy recommendation, and conclusions, while the final chapter is reserved for the English and German summaries.

2 EDUCATIONAL CONTEXT IN THE GCC COUNTRIES

2.1 Introduction

In total, eight countries border the Persian Gulf. When excluding the non-Arab state of Iran, seven countries remain, all of which are subsumed under the term *Arab States of the Persian Gulf*: namely Bahrain, Iraq, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates (ARE). With the exception of Iraq, all are politically and economically united in the Gulf Cooperation Council (GCC), an institution that was established in 1981 with the objective of strengthening relations and cooperation between participating countries in various areas such as economic and financial affairs, commerce, customs, and communication, but also in education and culture (Cooperation Council for the Arab States of the Gulf, 1981). The total area of the GCC countries is about 2,573,108 km^2, and its total population is estimated to be around 54 million people. The GCC will provide the focus for the current study, as its member countries exhibit several key similarities and because internationally comparable achievement as well as background data is available for each country. The region shares social and cultural values, religious beliefs, and historical events; each country declares Arabic as their official language. In addition, all the GCC countries are classified among the 21 wealthiest nations in the world (out of 187 ranked economies) as can be derived from Table 2-1.

Table 2-1 gives an overview on selected demographics and on indicators related to primary education in the Gulf region. These include population size, gross domestic product (GDP) per capita and rank among 187 measured economies, percentage of public expenditure in education, as well as net enrollment, student-teacher ratios, and their TIMSS 2015 achievement. For reference, the GCC countries are listed along with the highest and the lowest achieving TIMSS 2015 countries participating in grade four mathematics and science (i.e. Singapore and Morocco, respectively; South Africa had a similarly low achievement, but did not participate in the science assessment).

Table 2-1: Overview on demographics and selected indicators relevant to primary education

Country	Population*	GDP per capita**		Public Expenditure in Education (%)***	Net Enrollment Ratio in Education (%)***	Student-Teacher Ratio in primary education***	Average 2015 scores****	
	2015 (in thousands)	Rank	USD				Mathematics	Science
Bahrain	1,372	#014	50,704	3	-	12	451 (1.6)	459 (2.6)
Kuwait	3,936	#005	71,887	-	92	9	353 (3.2)	337 (6.2)
Oman	4,200	#021	46,698	4	91	7	425 (2.5)	431 (3.1)
Qatar	2,482	#001	127,660	4	92	11	439 (3.4)	436 (4.1)
Saudi Arabia	31,557	#012	55,158	5	96	11	383 (4.1)	390 (4.9)
United Arab Emirates	9,154	#008	67,871	1	91	19	452 (2.4)	451 (2.8)
Singapore	5,535	#003	90,151	3	100	17	618 (3.8)	590 (3.7)
Morocco	34,803	#112	8,330	5	98	26	377 (3.4)	352 (4.7)

Notes. * United Nations, n.d.b. ** International Monetary Fund, 2017 (based on 187 economies), *** TIMSS & PIRLS International Study Center, Boston College, 2016a, **** Mullis, Martin, Foy, and Hooper (2016) for mathematics & Martin, Mullis, Foy, and Hooper (2016) for science

Since the Second World War, GCC countries have experienced tremendous transformations in almost all aspects of socio-economic and political life, with major impact on their educational systems. The GCC countries represent some of the fastest growing economies in the world, mainly driven by their high oil and gas revenues (Low & Salazar, 2011). During the last decades, the GCC countries have reached living standards and income levels equal to those of developed countries and close to all young Gulf citizens now have access to formal education. Nevertheless, with increased oil revenue, gaps between upper and lower classes have widened; wealth is now distributed mainly between the upper classes (Saif, n.d.). Furthermore, with modernization and the development of more bureaucratic structures, in most GCC countries the power and authority of local sheiks are currently decreasing; the gap is being filled by the rise of a new and growing class of educated professionals (Colton, 2011, p. 40).

The situation in the Gulf region differs quite a bit when compared to the challenges faced by other developing countries, many of which, after achieving political independence, tried to develop their own human resources because of missing financial resources and in order to become more independent from their previous colonial masters (Bahgat, 1999, p. 128). The Gulf area, on the other hand, began to accumulate a vast economic fortune in the years following the start of the Second World War by exporting their natural gas and oil reserves. In the first decades after the war, the region (with exception of Saudi Arabia and Iraq) was still under British rule which was established in the 19th century and administered via a system of tribal leadership of only ten families (Metz, 1993, p. 30). These families had negotiated commercial treaties with the British Empire against British protection and now were benefiting from this new wealth. Kuwait, one of the first countries where oil resources were discovered, gained independency from British hegemony in 1961; subsequently, the rest of the region followed suit, culminating with the independence of the United Arab Emirates (formerly Trucial States) in 1971 (Metz, 1993). Ruling families, later the rulers of the newly established Gulf monarchies, shared the wealth accumulated from oil revenues with their people and also invested in the improvement

of social services, health care, and the education system. These investments enabled the foundation of a modern schooling system, which was needed in response to a shortage of the skilled local workforce necessary to meet the requirements of modernization. In the decades after the Second World War, many new schools and later universities were built as part of the newly created welfare system – in which most social services, including school attendance, were offered free or for a minimum of charge (Bahgat, 1999, p. 129). These developments resulted in great advances in quantitative educational factors, such as increased literacy and enrollment rates and decreased student-teacher ratios; the fast expansion of the education system, however, was only made possible by the assistance of expatriate teachers from Middle Eastern Arab countries.

"Western-style mass schooling" (Ridge, 2014, p. 23) then started at the beginning of the 1970s, with the withdrawal of British dominance and the economic wealth accumulated in the region allowing countries to take the "fast track to modernization" (Bill, 1984, p. 115). In contrast to this rapid economic development, however, the culture, mentality, and attitudes of the people changed very little, which led Bill (1984, p. 115) to conclude that "modernization and economic growth raced far ahead social and political development". This imbalance between fast economic growth and social development created a special situation leading to a number of social distortions, such as a mismatch between traditional and modern schooling, an imbalance between national and foreign workers, and rising gender disparities (Bahgat, 1999, p. 129). As these developments are important for a better understanding of the factors that led and still lead to low results in international achievement tests and to quite substantial gender differences, they should be elaborated in a bit more detail in the following sections.

Traditional and modern education

Until around the end of the 19th century, the traditional form of education in the region was the *kuttab* (or *Maktab*), where a group of students were mainly taught in reciting the Qur'an, and sometimes in reading, writing, grammar, and basic arithmetic skills (Bahgat, 1999, p. 129; "Maktab," 2007). While the first modern schools were founded in Kuwait in the first part of the 20th century, the foundation of a modern school system on a larger scale did not begin until the early 1950s. As royal families and governments generally sponsored investment in school infrastructure and provided public education free of charge, they – as the funders of the education system – also could exert strict control over the institutions of learning on all levels, so that there was "little room for academic and political freedom" (Bahgat, 1999, p. 130). He concluded that this situation would result in two main characteristics of the public education in the

GCC countries: firstly, the curriculum tended to be dominated by Islamic and Arabic studies; and secondly, more emphasis was put on academic learning than on vocational and technical training in general.

Expatriate labor force

Due to the rapid economic growth, in combination with a lack of a skilled national labor pool, the whole region's economy is heavily dependent on expatriate labor force. This is especially the case for the private sector, as the public sector is preferred by the locals because of the perception of having a far greater prestige and better working conditions such as higher salaries, better job security, shorter working hours, and an earlier retirement (Randeree, 2012; Ridge, 2014). In consequence, non-nationals[1] now represent a significant share of the population, accounting for from about 33% in Saudi Arabia to nearly 90% in the United Arab Emirates (see overview in Table 2-3). A similar situation can be observed in the field of education. As no teacher education facilities were available in the early years of mass schooling, the vast majority of public school teachers had to be recruited from surrounding Arab countries, particularly Egypt, Palestine, Jordan, Syria, and Lebanon (Bahgat, 1999, p. 130; Ridge, 2014, p. 21); as a downside, this resulted in the import of influences from a variety of different curricula and mainly transferred teacher-centered approaches with a focus on hard skills such as memorization and repetition (Ridge, 2014, p. 21). The employment of non-nationals as teachers poses many challenges concerning consistency in the quality of teaching, but also in terms of adjusting qualifications to the needs of the local systems (Ridge, 2014, p. 113). Only in Oman, which always had fewer natural resources than other GCC countries and in which men had fewer employment opportunities, is the share of males in the educational sector somewhat higher (Ridge, 2014, p. 125).

From the 1990s on, the steady decline in the quality of teachers from Egypt (which was the largest group of expatriate teachers) became more apparent, and GCC countries began more intensively investing in the training of local teaching forces (Engman, 2009, p. 40; Ridge, 2014, p. 23). However, the majority of those who embarked in the field of teaching were women, as men usually had (and still have) more employment possibilities, and teaching among them is

[1] Non-nationals are "1 - persons bearing nationality of a foreign State other than the GCC State of residence, or bearing no proof of nationality from any given state, or 2 – holders of residence permit residing in the given GCC country at date of census" ("GCC: Total population and percentage of nationals GCC: Total population and percentages of nationals and foreign nationals in GCC countries," 2017)

often regarded as a "low-status" profession (Ridge, 2014, p. 98). In consequence, boys in the more segregated Gulf school systems, especially those in single-sex schools and in higher secondary education, are still mainly taught by expatriate Arab male teachers (Barbar, Gardner, & Andrew, 2016, p. 45; Ridge, 2014, p. 109).

While non-nationals dominate the workforce in most GCC countries, they are not integrated in the Gulf societies, but rather live (and often work) completely separated as an independent population, under completely different conditions when compared to residents or nationals. They only have temporary residency, have (with few exceptions) no access to citizenship of the country they are living in, and only have limited possibilities to participate in society (Fargues, 2011, p. 274). Non-nationals work under precarious situations and their wages are often very low. Usually, they are bound to specific employers and risk deportation if they don't maintain valid contracts. The highest proportion of non-nationals originally stemmed from other Arab countries but their share declined to less than 30 percent in 2002, while the proportion of Asians rose (Kapiszewski, 2006). Galal (2008, p. 250) reports that in general Arabs dominate the higher skill categories, such as technicians or mangers, while Asians dominate lower skill positions such as services, agricultural and production related jobs. The middle-skills categories (sales) are shared between both groups.

More recently, weaker revenues from natural resources in the last decade of the 20th century, coupled with higher unemployment rates, led to the launch of so-called nationalization programs in the Gulf States. These nationalization policies have the objective of reducing dependency on foreign labor by prioritizing the national population in the labor market through human resource strategies influencing "recruitment, training, career management and the design of reward systems" (Randeree, 2012, p. 6).

Education and gender gap

While the economic modernization of the Gulf requires a skilled labor force, the contribution of women in this context still is only modest. A summary of the ILOSTAT labor statistics data (The International Labour Organization, 2018) by the World Bank (2016) showed a female labor force participation rate ranging from only 20% in Saudi Arabia to 53% in Qatar for females older than 15 in 2015. The low contribution of women might be explained by the traditional nature culture of the Gulf societies, which prescribe different roles for the two sexes. Until a few decades ago, the role assigned to women by society "was being a good wife and a good mother" (Bahgat, 1999, p. 133) and consequently their work domain was focused mainly within the domestic, household area (Randeree, 2012, p. 4). Education for girls, therefore, was

not seen as a necessity by many until foreign presence in the region played an important role in opening schools for girls (Bahgat, 1999, p. 133). Later, abundant financial resources from oil sales and the region's strive for social and economic modernization led to a vast expansion of female education. However, while girls in all Gulf countries now have equal access to primary and secondary education, and in some of the countries even outnumber their male counterparts in university enrollment, they still are restricted in terms of job opportunities. Women face the most restrictions in Saudi Arabia, where practitioners of Wahhabism still teach that a women's primary responsibility is maintaining home and family life, and consider gender-segregated fields like education, nursing, and public administration more appropriate for women (Bahgat, 1999; Ridge, 2014, p. 146).

Modernization programs

With the infrastructure mainly in place, staffing issues addressed, and enrollment rates in primary education close to 100%, in the end of the 1990s, the next wave of modernization programs were launched. Targeting the quality of education, they were influenced by the participation of the region in international comparative assessments as well as by new goals for education declared by international organizations, such as the Education for All initiative (UNESCO, 2000) or the Millennium development goals (United Nations, n.d.a), introduced in the year 2000**.** As local capacity was not sufficient to undertake comprehensive educational reforms, global management consultancy firms such as McKinsey, the Rand Corporation, or the World Bank were contracted to assist in developing the necessary strategies to help GCC countries in the intended transition from resource-based to knowledge-based economies (Ridge, 2014). Based on their recommendations, the GCC countries undertook a number of various reforms and special initiatives to improve educational quality in areas such as curriculum, professional development, and the use of ICT technology in education. This included shifting from public to more independent schools, from Arabic to English as the language of instruction in science and mathematics, and from traditional teaching methods to inquiry-oriented ones (Bou-Jaoude & Dagher, 2009, p. 1). In addition, in some of the countries the time allocated for mathematics and science instruction or the teaching of computer technology skills has been extended (Al-Awadhi, 2016, p. 8; AlMaskari, AlMawali, AlHarthi, & AlRasbi, 2016, p. 12). Some more recent examples of such programs include the *Bahrain Numeracy Strategy*, with the objective to raise mathematics performance by enhancing the quality of instruction and learning and help Bahraini students develop self-confidence (Al-Awadhi, 2016, p. 8; Oxford Business Group, 2012, p. 188), implemented in 2011. In Oman, the *Cognitive development program* was inaugurated in the 2007-8 school year, with the intention "to encourage students to acquire

knowledge, improve their level of attainment in science, mathematics, and environmental geography, and enhance their study of the practical aspects of these subjects" (AlMaskari et al., 2016, p. 12). In Kuwait, a collaboration between the Kuwaiti Ministry of Education and the World Bank related to curriculum reform, teaching strategies, and teacher skills enhancement was established (National Center for Education Development, 2016, p. 6). While most recent results from international large-scale assessments show certain improvements in terms of achievement and gender equity in most countries of the region, this progress is slow and the quality of education in the region remains a major concern especially in the fields of curriculum implementation, teacher education, and in a lack of research in the field (BouJaoude & Dagher, 2009, p. 3). Ridge (2014, p. 96) sees the constant struggle between countries and even within countries among different territories "to be seen as the biggest or the best", in combination with a refusal to acknowledge any weaknesses in their countries, as the major problem that hinders substantive development in the region.

2.2 The Schooling System in the GCC Countries

The formal education in the GCC countries comprises kindergarten, stages of primary (or basic) education, intermediate (or preparatory) and secondary schooling, followed by tertiary education. All of the countries also focus on extending vocational education tracks or different specialization programs on a secondary level.

Public education

Public schooling on all levels is usually free of charge for national citizens of GCC countries. Most have a highly centralized education system, wherein the Ministry of Education is responsible for prescribing the national curriculum and for providing all necessary facilities and equipment needed for the public school sector. However, several countries in the region have started initiatives to de-centralize the school system. While Bahrain decentralized its Ministry of Education in the 1980s, granting schools more autonomy shortly thereafter, the United Arab Emirates more recently distributed the responsibility for the education to local education authorities in each Emirate and also in Oman the Ministry of Education is implementing a strategy to delegate more administrative functions to regional offices (Ridge, 2014).

The language of instruction in the public school sector in general is Arabic. Compulsory education usually goes until Grade 9 – in Saudi Arabia, even until Grade 12. While Oman provides formal education until the end of secondary school, attendance is not compulsory (Al-Ani, 2016, p. 328). An overview on the communalities and differences of the GCC primary and

secondary education cycles summarized from the *TIMSS Encyclopedia* (Mullis, 2012) can be found in Table 2-2.

Table 2-2: Overview on primary and secondary education in the GCC countries

Country	Grade											
	1	2	3	4	5	6	7	8	9	10	11	12
Bahrain	Basic Education (Cycle 1)			Basic Education (Cycle 2)			Basic Education (Cycle 3)			Secondary		
Kuwait	Primary					Intermediate				Secondary		
Oman	Basic (Cycle 1)					Basic (Cycle 2)					Secondary	
Qatar	Primary						Preparatory			Secondary		
Saudi Arabia	Primary						Intermediate			Secondary		
United Arab Emirates	Basic (Cycle 1)					Basic (Cycle 2)				Secondary		

Notes. Content summarized from Mullis (2012).
Areas with dotted pattern: Not compulsory

Private education

The Gulf region has a quite pronounced private school sector, which in some cases is supported by the ministries of education but usually is not free of charge for the students enrolled. Ardent (2015, p. 12) argued that the private school system is steadily growing as parents gain awareness and readiness to pay for the higher quality of education, more modern curricula, and stronger orientation towards the English language which are often provided by private schools. Furthermore, he stated that a high demand for private schooling is also based on the expatriate population, which often faces restrictions in enrolling their children in the public school sector. As can be derived from Table 2-3 below, a higher share of the private school sector is typically found in GCC countries with high foreign populations, such as the United Arab Emirates or Qatar.

Table 2-3: Percentage of private enrollment in primary and percentage of non-nationals

Country	Private Enrollment in Primary (%)*	Non-Nationals (% of total population)**
Bahrain	36	52
Kuwait	43	69
Oman	20	44
Qatar	63	86
Saudi Arabia	10	33
United Arab Emirates	77	89

Notes. * World Bank, n. d., ** "GCC: Total population and percentage of nationals and non-nationals in GCC countries (latest national statistics, 2010-2015)," 2015

Although private schools often have their own curricula and offer instruction in English or in the national languages of the immigrant population, they still are closely supervised by the ministries of educations, which also approve curricula and learning material.

Table 2-3 lists the percentage of students who are enrolled in private schools on a primary level, in combination with the share of non-nationals in the population.

Many GCC countries offer different types of private schooling for different purposes. In general, the following types can be distinguished (the naming follows the conventions described by Jarrar and Alharqan [2016] for Qatar):

- Independent (private Arabic) schools that are often associated with the ministries of education and follow the national curriculum. Those schools are often attended by national children of wealthier families.
- Community schools that are specific private schools for the expatriate population. They follow, to a certain extent, the curriculum of the different expatriate communities.
- International schools, which usually have the highest standards and fees and follow an "international" curriculum. The language of instruction in international schools is usually English.

2.3 The Curricula in the GCC Countries

All GCC countries have national curricula for primary and secondary education in mathematics and science (TIMSS & PIRLS International Study Center, Boston College, 2016a). For a long time, criticism was raised against curricula in the GCC countries for being outdated; it was posited that they would not prepare children for the needs of the labor market, and offered insufficient attention to analytical thinking and communication skills (Aziz, 2016, p. 39; Bou-Jaoude & Dagher, 2009, p. 3; Brewer, 2007, p. 2).

However, in the last couple of years, education became an issue of major concern in the region. Consequently, all GCC countries developed roadmaps for their primary and secondary education, including standards for mathematics and science instruction (Aziz, 2016, p. 39). Al Maskari et al., for example, report that for the Omani curriculum:

> The scope and sequence of both the mathematics and the science curricula were revised completely for Grades 1 to 10. Certain learning outcomes were moved from one grade to another. New outcomes were introduced for some grades to bring them in line with international scope and sequence. Topics covered by TIMSS

2007 also were taken into consideration (AlMaskari et al., 2016, p. 13).

Table 2-4 provides information on the coverage of the TIMSS testing framework in the GCC countries. The first three columns for each subject show how many of the 17 mathematics and 23 science topics from the TIMSS 2015 framework are covered by the national curricula in the region.

Table 2-4: Number of TIMSS topics intended to be taught by end of grade and Test Curriculum Matching Analysis

Country	Mathematics				Science			
	TIMSS Topics covered (all = 17)			TCMA	TIMSS Topics covered (all = 23)			TCMA
	Number of Topics Taught to All or Almost All Students	Number of Topics Taught to Only the More Able Students	Not included in the Curriculum Through Grade 4	Test Curriculum Matching Analysis (%of items covered)	Number of Topics Taught to All or Almost All Students	Number of Topics Taught to Only the More Able Students	Not included in the Curriculum Through Grade 4	Test Curriculum Matching Analysis (%of items covered)
Bahrain	16	0	1	98	20	0	3	96
Kuwait	17	0	0	91	23	0	0	90
Oman	8	9	0	74	12	1	10	86
Qatar	13	0	4	95	20	2	3	100
Saudi Arabia	17	0	0	100	23	0	0	40
United Arab Emirates	15	0	2	100	17	1	5	40
Gulf Average	**14**	**2**	**1**	**93**	**19**	**1**	**4**	**75**
Int. Average	**13**	**1**	**3**		**16**	**1**	**4**	

Notes. Content summarized from TIMSS & PIRLS International Study Center, Boston College (2016a) for the TIMSS topics covered and from TIMSS & PIRLS International Study Center, Boston College (2016b) for the TCMA analyses.
TCMA = Test Curriculum Matching Analysis

It should be noted that the evaluation concerning the match between national curricula and TIMSS evaluation framework is based on subjective judgement of the National Research Coordinators (NRCs) for TIMSS. On average, the regional coverage of the TIMSS domains appears a bit higher than the international average for both subjects. However, the fact that the topic coverage is particularly low for both subjects in Oman is noteworthy. A slightly different perspective is obtained by the results of the test curriculum matching analysis (TCMA). Results of the TCMA are displayed for each subject in the rightmost column. Here the NRC compared the coverage of his/her national curriculum with the TIMSS framework on a test item level, and the results again clearly show a lower coverage of the TIMSS test content for Oman. Interestingly, Saudi Arabia and the United Arab Emirates obtain only 40% coverage in the TCMA analyses although all or nearly all of the science content domain topics in general are reported by the teachers, as included in the curriculum and as already covered in their teaching.

2.4 Achievement of GCC Countries in International Large-Scale Assessments

Participation in TIMSS, PIRLS, and PISA

GCC countries have participated in several cycles of IEA TIMSS (Mullis, Martin et al., 2016), as well as in IEA PIRLS (Mullis, Martin, Foy, & Drucker, 2012). Qatar and the United Arab Emirates also participate in OECD PISA (OECD, 2016a). Table 2-5 shows an overview on the participation of GCC countries in the different assessment cycles of TIMSS, PIRLS, and PISA.

Table 2-5: Participation of GCC countries in international large-scale assessments

Country	TIMSS											PIRLS				PISA		
	1995		1999	2003		2007		2011		2015		2001	2006	2011	2016	2009	2012	2015
	G4	G8	G8	G4	G8	G4	G8	G4	G8	G4	G8	G4	G4	G4	G4	15 y	15 y	15 y
Bahrain		x	x		x		x	x	x	x	x				x			
Kuwait	x	x				x	x	x		x	x	x	x		x			
Oman							x	x	x	x	x			x	x			
Qatar						x	x	x	x	x	x		x	x	x	x	x	x
Saudi Arabia					x		x	x	x	x	x			x	x			
United Arab Emirates								x	x	x	x			x	x		x	x

Note. Content summarized from Mullis, Martin et al. (2016) for TIMSS; from Mullis, Martin, Foy, and Drucker (2012) for PIRLS; and from OECD (2016b) for PISA.

As shown in Table 2-5, GCC countries only participated sporadically in international large-scale assessments and until around 2007 mostly in grade eight. From 2011 on, however, all six GCC states participated on the primary level of the TIMSS assessment as well. In both assessment cycles and both grades, GCC countries are located on the lower end of the TIMSS scale, with the highest achievement scores usually listed for Bahrain or the United Arab Emirates. When compared to the group of countries participating in the same grade, GCC countries mainly seem to perform comparatively better in grade eight than in grade four.

Performance on primary level in TIMSS grade four

At the primary level, in both subjects, all GCC countries performed in the lowest quartile of the TIMSS 2015 ranking scales. Internationally, the results of the region are comparable to some other (predominantly) Islamic countries such as Iran (431 score points in math/421 score points in science), Indonesia (397/397), Jordan (388/-), or Morocco (377/352). An overview of the grade four mathematics and science results for 2011 and for the most recent assessment in 2015 can be found in Table 2-6. As in Table 2-1 for comparison, the table also contains the mathematics and science performance of the highest and the lowest achieving TIMSS 2015 countries

participating in both grade four mathematics and science (i.e. Singapore and Morocco, respectively).

Table 2-6: Overall mathematics and science scores of the GCC countries in 2011 and 2015

Country	Mathematics		Difference (Absolute Value)	Science		Difference (Absolute Value)
	Average 2011 Score	Average 2015 Score		Average 2011 Score	Average 2015 Score	
Bahrain	436 (3.2)	451 (1.6)	15 ▲	449 (3.5)	459 (2.6)	9 ▲
Kuwait	342 (3.6)	353 (3.2)	11 ▲	347 (4.8)	337 (6.2)	10 ▼
Oman	385 (2.9)	425 (2.5)	41 ▲	377 (4.3)	431 (3.1)	54 ▲
Qatar	413 (3.4)	439 (3.4)	26 ▲	394 (4.3)	436 (4.1)	42 ▲
Saudi Arabia	410 (5.2)	383 (4.1)	27 ▼	429 (5.5)	390 (4.9)	39 ▼
United Arab Emirates	434 (2.0)	452 (2.4)	17 ▲	428 (2.5)	451 (2.8)	23 ▲
Gulf Average	**403 (3.4)**	**417 (2.9)**	**23 ▲**	**404 (4.2)**	**417 (4.0)**	**30 ▲**
Singapore	606 (3.2)	618 (3.8)	12 ▲	583 (3.4)	590 (3.7)	7
Morocco	335 (4.0)	377 (3.4)	43 ▲	264 (4.4)	352 (4.7)	89 ▲

Notes. Content summarized from Mullis, Martin, Foy, and Arora (2012) for TIMSS 2011 mathematics; from Mullis, Martin et al. (2016) for TIMSS 2015 mathematics; from Martin, Mullis, Foy, and Stanco (2012) for TIMSS 2011 science; and from Martin, Mullis, Foy et al. (2016) for TIMSS 2015 science.
▲Results in 2015 significantly higher
▼Results in 2011 significantly higher
() Standard errors appear in parenthesis

Results from Table 2-6 show that for four of the countries, results in both subjects have remarkably improved between both assessment cycles, while achievement in Saudi Arabia and Kuwait declined in the same period. It can also be seen that the differences within the GCC region's top and low performing countries are quite large. In science, the difference between Kuwait and the two top performers Bahrain and the United Arab Emirates exceed by far one standard deviation. It also can be seen that the average achievement of the GCC countries for mathematics is about two standard deviations lower than for the TIMSS 2015 top performing country Singapore, while the achievement difference for science still amounts to more than one and a half standard deviations in favor of Singapore.

The international results for the last cycle of PIRLS in 2011 (Mullis, Martin, Foy, & Drucker, 2012) show a similar picture: All GCC countries are located in the lowest quartile of the achievement scale. However, differences within the region – e.g. between the highest performer, the United Arab Emirates, with 439 score points and the lowest performing country, Oman, with 391 points – were lower.

Since 2009, Qatar and the United Arab Emirates also participated in the OECD PISA Assessment (OECD, 2016a). Both countries are located in the lower half of the PISA performance distribution in all three subjects, with the United Arab Emirates outperforming Qatar in all subjects. This mirrors to a large extent the findings from the TIMSS assessment, especially when looking at grade eight.

3 RESEARCH ON FACTORS INFLUENCING STUDENT PERFORMANCE

3.1 Educational Effectiveness Research

About five decades of educational effectiveness research (EER) have brought the topic of educational effectiveness to a prominent position in research agendas around the world. While initial research results in this area indicated that "Schools bring little influence to bear on a child's achievement that is independent of his background and general social context" (Coleman et al., 1966, p. 325), currently there is a widespread consensus among researchers that schools influence children's development and educational outcomes in many ways (Chapman et al., 2015; Reynolds et al., 2014; Teddlie & Reynolds, 2000).

3.1.1 Strands of educational effectiveness research

Depending on the underlying research interest, three major strands of EER can be distinguished: *School Effects Research,* that studies the scientific properties of school effects; *Effective Schools Research,* that focuses on the processes of effective schooling and is initially often based on qualitative case studies of well-performing outlier schools; and *School Improvement Research,* that examines how schools can be changed and improved over time (Teddlie & Reynolds, 2000).

School Effects Research is concerned with the influence of schooling on intended student outcomes. Good and Brophy (1986) define *school effects* as what is known about the ability of schools to affect the outcomes of the students that they serve. A similar definition is given by Raudenbush and Willms (1995, p. 308), who define school effects as "…the extent to which attending a particular school modifies a student's outcome." The underlying question here is to what extent the school environment shows a separate influence on student outcomes beyond certain input characteristics of the student body. School effects essentially focus on the identification of factors which enhance effectiveness in the school environment using methodological sound approaches.

The development of the School Effects Research branch also can be seen as a reaction to the Coleman Report, which concluded that "…the inequalities imposed in children by their home, neighborhood, and peer environment are carried along to become the inequalities with which they confront adult life at the end of the school" (Coleman et al., 1966, p. 325). In addition to

the pessimistic conclusion drawn concerning the influence of school-related factors, researchers also tried to address or counter methodological concerns that were brought up regarding EER (at that time called *school effectiveness research*) from the very beginning. Teddlie and Reynolds, for example, state that the Coleman Report received many criticisms about methodological issues, including the charge "that they did not operationalize the school input variables adequately in order to properly assess the effect that schools have on student achievement" (Teddlie & Reynolds, 2000, p. 58). Researchers in this field are predominantly concentrating on general methodological and psychometric issues such as reliability, generalizability, or validity.

The second branch of EER, the Effective Schools Research, also emerged as a reaction to the Coleman Report. Research in this strand initially tried to refute results from the report, and intended to prove that schools can do and make a difference. Focus here is set on the identification of highly successful schools and students, and comparing them with comparable schools – in terms of student composition – that are less effective in terms of student outcomes. In that sense, it can be argued that "A more effective school is one in which student performance is higher than predicted by input" (Chapman et al., 2015, p. 27). The research interest in Effective Schools Research is mainly focused on identifying differences between schools in order to understand the conditions that lead to more effective schools. Research designs are usually based on qualitative case studies of especially effective schools and originally focused mainly on public schools attended by children from low socio-economic backgrounds – for example, Edmonds (1979).

School Improvement Research, however, is not primarily focused on detecting effectiveness-enhancing factors related to outcome variables, but rather seeks to develop strategies to enable schools to become more effective. The main focus here is on change processes in educational contexts that should be described and ideally improved. Here, the individual school is considered the center of the change – thus, changes and reforms need to consider the internal conditions of a school, and usually to follow a systematic approach of improvement over several years. Hopkins (2001, p. 13) defines school improvement as a "distinct approach to educational change that aims to enhance student outcomes as well as strengthening the school's capacity for managing change".

The current research project seeks to detect effectiveness-enhancing factors in the GCC countries and aims to describe the relationship among them; it therefore is based in the *school effects research paradigm*.

3.1.2 International comparative studies and educational effectiveness

The historical context described in the previous section predominantly reflects the developments of research in a Western context, where most of the research was done. Teddlie and Reynolds (2000, p. 232) argued that educational effectiveness in the past "has shown heavily ethnocentric tendencies" and they found, when evaluating the corresponding literature, that research in this field is "almost exclusively based upon scholars and researchers within the country of origin of the writer." They therefore concluded that "the area of international effectiveness research…suggests an area so far relatively undeveloped." Most of the literature in the past stems from Western countries, mainly from North America, Great Britain, The Netherland, Canada, Australia, Norway and Sweden. More than a decade later, Reynolds et al. (2014, p. 221) still emphasize the importance of the international dimension, stressing in their state-of-the-art review of EER that "An international perspective is of vital importance, since EER (Educational effectiveness research) may not mean the same thing in different parts of the world."

Thus far, only a few studies, such as the International School Effectiveness Research Project (ISERP) as described by Reynolds (2006), have explicitly adopted a research design to measure educational effectiveness. The study was conducted in nine educational systems, but among them only two from outside the Western Hemisphere – namely Hong Kong and Taiwan. In spite of major differences across countries and especially between Western and Asian school systems, Reynolds also reported important similarities in terms of the factors that are associated with good schools: "We cannot stress too highly that many factors that make for good schools are conceptually quite similar in countries that have widely different cultural, social, and economic contexts. The factors hold true at the school level, but the detail of how school-level concepts play out within countries is different between countries. At the classroom level, the powerful elements of expectation, management, clarity, and instructional quality transcend culture" (Reynolds, 2006, pp. 554–555).

Postlethwaite and Ross (1992) were among the first to use the vast range of contextual variables contained in international large-scale assessments to identify indicators associated with a kind of educational effectiveness. They analyzed data from the IEA Reading Literacy study conducted between 1989 and 1992, which included 32 educational systems from all over the world – but none of them in the Gulf area. Summarizing their results, schools associated with higher achievement tended to be well-managed, initiative-taking, well-stocked with library books, and had teachers who were more professional and used particular methods of teaching (encouraging

the students to read, emphasizing assessments, having high demands on structure, and so forth). Unfortunately, this study ignored the hierarchical structure of the data by not disentangling the effects of different educational levels (for example by applying hierarchal multi-level analyses) – a critical consideration for this type of analysis as argued by Raudenbush and Bryk (1986).

Martin, Mullis, Gregory, Hoyle, and Shen (2000) based their analyses on the TIMSS 1995 study and included data from 34 educational systems, but again, at that time, no Gulf State participated in the study. Their contribution can be seen as one of the first studies to use international large-scale assessments while concurrently taking the hierarchical structure of the data into account. Martin et al. found that factors related to the socio-economic status (SES) of the student distinguished more uniformly between high- and low-achieving schools across countries than factors that are more directly related to the school, class, and teacher level. Subsequently, a growing number of authors applied multilevel modeling techniques to account for the clustering effects of nested data when using IEA TIMSS and PIRLS data for analyses in the field of educational effectiveness (for example Kyriakides, 2006; Lamb & Fullarton, 2001; Rutkowski & Rutkowski, 2008; Schwippert, 2001; Webster & Fisher, 2000).

The above-mentioned authors, among others, focused on what occurs within schools and tried to identify "value-added" variables by investigating characteristics related to organization, form and content. Findings from previous multilevel analyses of the author using eighth Grade data of TIMSS 2007 and 2003 (Neuschmidt, Hencke, Rutkowski, & Rutkowski, 2010; Neuschmidt, Hencke, Rutkowski, & Rutkowski, 2011) indicated home background indicators and nationality status as the most important predictors of mathematics achievement in the Gulf area. In addition, different class- and school- level related variables, such as student behavior, teaching experience, and monitoring homework were found to be significant indicators predicting mathematics outcomes.

Results from PISA 2012 indicated the following major general findings on system level (OECD, 2013): A negative relation between stratification in school systems and equity; and a more equitable allocation of school resources as well as a greater degree of school autonomy in terms of curricula and assessment in high-performing countries. Results also indicated lower performance for systems with larger proportions of students who arrive late for school and skip classes. In PISA 2012, the two GCC countries (Qatar and the United Arab Emirates) that participated, together with 63 other educational systems, showed a country mean in mathematics achievement far below the OECD average – but also, interestingly, greater equity concerning

their educational outcomes. Looking at the PISA 2015 results (OECD, 2016a), a similar pattern also can be discerned for both countries in science achievement.

An investigation on educational effectiveness factors was also among the research topics presented in the *Relationships Report* published by the International Study Center for TIMSS & PIRLS (Martin & Mullis, 2013). For the 34 countries and three benchmarking participants that administered TIMSS and PIRLS to the same students, the relationship between school, teacher, and home background scales on one hand, and student achievement in the three subjects on the other, were analyzed. For this purpose, several two-level hierarchical linear models were constructed. The sampling design of most of the analyzed educational systems didn't allow for the creation of three-level models; given the usual selection of one class per school, the variance components between schools and classes could not be separated. While the authors found considerable differences across countries concerning the achievement levels between schools and in the relation of school variables to student achievement, the results between the three subjects were found to be very similar. The home resources indicator was found to be the most important predictor for achievement. After controlling for the home background on both levels, the school environment scales indicating school safety/orderliness and emphasis on academic success still played an important role in many of the analyzed countries. The most important school instructional scale was found to be the student engagement in reading, mathematics, and science. In all four GCC countries (Qatar, Oman, Saudi Arabia, and United Arab Emirates) for which data was available, student engagement was significantly associated with achievement even after controlling for the home background. Other important predictors emerging in the region were: schools are safe and orderly and school support for academic success.

It is important to note here that the concepts used to measure effectiveness might differ in different regions of the world, and may therefore not necessarily reflect the Western view which often mainly focuses on academic achievement. Harber and Muthukrishna (2000, p. 430), investigating school effectiveness in South African schools in the 1990s, for example, describe an ideological dimension of effectiveness aimed at "fostering a non-violent, non-racist and democratic society" which goes beyond dimensions of functional effectiveness that include indicators like an orderly atmosphere and businesslike behavior.

In a seminar related to TIMSS Mathematics Learning Outcomes in Doha (Qatar) the author asked representatives from Ministries of Education and National Committees for Education from Qatar, Bahrain, and Kuwait about their definition of an effective school in their country. Participants listed the following characteristics, summarized in the report from Khan (2015):

An effective school:

- Is capable of achieving its future vision for education in light of the international vision.
- Guarantees distinguished and equal educational opportunities for all and thus helps students achieve better than expected results.
- Helps the students to acquire positive trends related to citizenship.
- Cares about teachers' career development.
- Offers opportunities for participation, teamwork and fruitful cooperation amongst teachers.
- Provides modern educational resources for the students and the teachers.
- Provides diversified technological systems.
- Provides assessments and agendas.
- Caters for all students' inclinations and trends in school activities.

This list of characteristics of effective schools resulting from the TIMSS seminar also shows a certain emphasis on educational quality and equity, which reflect the main dimensions of educational effectiveness regarded in the West. However, beyond the focus on academic outcomes, respondents introduced the idea that effective schooling in the region is also required to deliver a good civic education in the sense that students should, as a result, become good citizens of their country. Such statements point to the importance of an additional function of schooling in the region: namely, legitimization of the respective system of government. The different functions of schooling are described in the section on educational quality in chapter 3.2.

3.2 Educational Effectiveness Research (EER) – Definitions

EER has gradually developed from trying to prove that "schools matter" to a more comprehensive understanding of which conditions and factors affect the effectiveness of the school as a system – and how they interrelate (Sammons, Davis, & Gray, 2015). As separate research strands focusing on *school effectiveness*, *teacher effectiveness*, *instructional effectiveness*, and so forth developed independently from each other, only to be combined in more recent years, related terminology has partially changed meanings over time. Hence, associated terms are often used in a different way by different authors: "…it is important to note that the terms 'school effectiveness', 'teacher effectiveness' and 'educational effectiveness' are used inconsistently in

the literature and that these themselves are interrelated" (Creemers, Kyriakides, & Sammons, 2010, p. 4).

Educational Effectiveness

Many authors (for example Chapman et al., 2015; Scheerens, 2004b, 2016) see the basic functioning of the educational system as an *Input – Output* model that is influenced by process factors within, and by context factors external to, the system under consideration. An example of such a model is depicted in Figure 3-1. Accordingly, education can be seen as a production process (in the field of economics, this is also known as *educational production function*), managed by malleable inputs and processes, and ultimately leading to certain output factors, which are often measured on a student level (Scheerens, 2016). More details about the Input – Output model from the economic perspective and about the transfer from economic theory to the field of education can be found by Hanushek (1986). In organizational theory, a similar kind of model is referred to as the *rational goal model*, in which productivity and efficiency are the central criteria to assess effectiveness (Scheerens, 2004a, p. 124). Other models may emphasize different aspects of effectiveness: The *open systems model* focuses on growth and resource acquisition, while the *human relations model* focuses on human resource development, and the *internal process model* on stability and control (Scheerens, 2016).

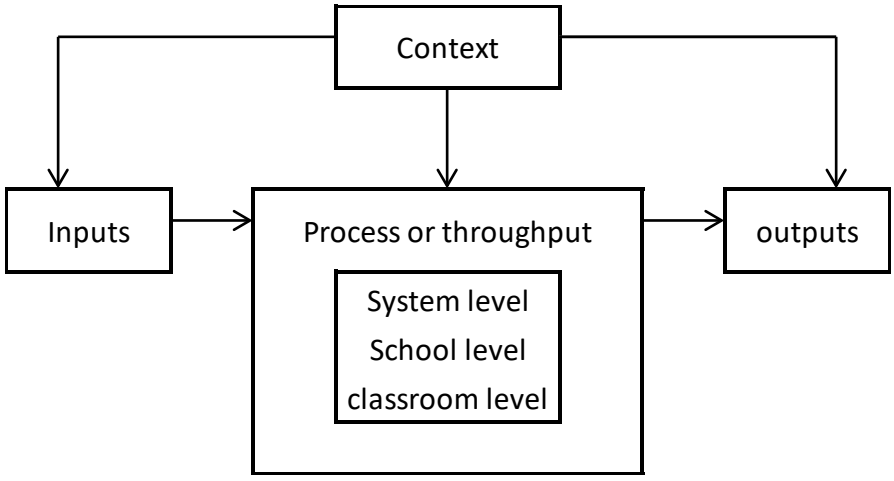


Figure 3-1: Basic system model on the functioning of education (from Scheerens, 2016, p. 6)

Following the *input-process-output* model approach, EER is primarily concerned with detecting malleable input and process variables that are associated with outcome factors of interest, often cognitive student outcomes (Creemers & Kyriakides, 2008; Scheerens, 2004b).

Scheerens therefore described effectiveness research as follows:

> The major task of educational effectiveness research is to reveal the impact of relevant input characteristics on output and to "break open" the black box in order to show which process or throughput factors "work", next to the impact of contextual conditions (Scheerens, 2016, p. 6).

The term *educational effectiveness* tries to integrate a broad range of research areas from different strands related to research on different levels, often with a focus on conditions on school level (such as school organization and policies) and classroom level (with a focus on teacher behavior, classroom instruction etc.; Chapman et al., 2015; Creemers & Kyriakides, 2008). Scheerens (2016) also includes policy-amenable conditions at the national level.

Following advances in the field of effectiveness research in education, but also due to major improvements in the scientific methods applied, it is now generally accepted that those *effectiveness-enhancing factors* work on different levels of educational systems, while still being interrelated. When viewing educational systems as hierarchically organized, educational effectiveness can be regarded as an attempt to incorporate effectiveness research from all different levels. The following broad definition of effectiveness research given by Creemers et al. (2010, p. 3) and quite similarly by Chapman et al. (2015) will be adopted for this thesis: "Education effectiveness research can be seen as an overarching theme that links together a conglomerate of research in different areas, including research on teacher behavior and its impacts; curriculum, student grouping procedures; school organization; and educational policy".

According to the level of the educational system, EER can then be categorized into the following subareas:

- *System effectiveness*, a more recent term and not yet necessarily included in all definitions concerning educational effectiveness, was stimulated by the rise of international comparative assessments. It investigates malleable conditions at the national level that can be associated with student outcomes; for example, policies regarding to school autonomy, accountability, and choice of the school by parents.
- *School Effectiveness* then points to malleable factors on the school level, such as school organization and educational policies. Creemers and Kyriakides (2008, p. 3) give the following definition: "School effectiveness here refers to the role of school processes and organization: 'the impact that school-wide factors, such as policy for

teaching, school climate, and the school's perceived mission, have on student's cognitive and affective performance."

- On classroom level, the term *teacher effectiveness* focuses on the impact of teacher background and classroom factors on student performance. The terms *instructional effectiveness* or *teaching effectiveness* are sometimes used to specifically refer to activities of the teachers in the classroom. These terms may partially be used interchangeably. In this thesis, the term teacher effectiveness will be used in the more general sense of Creemers and Kyriakides (2008, p. 3), who define teacher effectiveness as referring "to the impact that classroom factors, such as teacher behavior, teacher expectations, classroom organization, and use of classroom resources, have on student performance."

Value-added

While Chapman et al. describe EER similarly to the definitions above, they add an important *value-added* concept to the second part of their definition:

> It therefore seeks to identify and explore the factors related to teaching, curriculum, and learning environments that may explain in a statistical sense (both directly and indirectly) the variation in student outcomes, while also controlling for student intake characteristics such as socioeconomic status and prior attainment/prior ability (Chapman et al., 2015, p. 30).

The focus here is on *effectiveness-enhancing factors* that are *purged* from any context factors such as the student's background. The term *value-added* is borrowed from the discipline of economics; when transferred to education, this concept basically means "a measure of the relative gain in achievement made by pupils" (Teddlie & Reynolds, 2000, p. 264). The underlying idea is that a school or educational system is not so much responsible for the absolute level of the student achievement, but rather for the progress students make within the educational system. Consequently, any context factors that influence student achievement need to be disentangled as much as possible from the educational factors which often constitute the main focus of interest. According to this conceptualization "A more effective school is one in which student performance is higher than predicted by input" (Chapman et al., 2015, p. 27).

However, the extent to which a disentanglement of school instructional factors from background factors is possible depends on different factors such as the data design (longitudinal or

cross-sectional), the availability of context factors regarded as important – such as prior ability or socio-economic status of the student, and the methodology used for the separation of educational factors from context factors. Different value-added approaches and the model used for the current study will be discussed in more detail in section 8.2.2.

The OECD, after seeking expert input from the field, defined the value-added component of a school as "the contribution of a school to students' progress towards stated or prescribed education objectives (e.g. cognitive achievement). The contribution is net of other factors that contribute to students' educational progress" (OECD, 2008, p. 17).

Effectiveness and Quality Criteria

In the most general sense, *effectiveness* refers to the level of goal attainment; school effectiveness, therefore, refers to the school's degree of achieving its educational objectives. This definition, however, needs a clarification concerning the objectives of a school, who defines them, and how they can be measured. Usually, two general dimensions of effectiveness are discussed: a quality dimension and an equity dimension.

Educational Quality:

According to Heid (2000), quality as such is not an objective and observable property of an object. Assigning a certain level of quality, therefore, only can be based on a subjective evaluation process. Heid argues that this evaluation depends on explicit and implicit decisions about certain criteria to evaluate an objects' nature, and that these decisions are made by those who claim to ensure and establish quality (Heid, 2000, p. 41). This means that it is not possible to define a uniform definition of educational quality, as due to the different interests of stakeholders involved in education, quality only can be defined on the base of a certain perspective (Harvey & Green, 1993; Terhart, 2000). Harvey and Green (1993) define five major differing conceptualizations or categories of quality: quality as *exceptional*, quality as *perfection* or *consistency*, quality as *fitness for purpose*, quality as *value for money*, and quality as *transformation*.

In EER, and hence in the scope of this thesis, the focus is mainly set on the aspect of the *transformation of the participant*. In that sense, "A quality education is one that affects changes in the participants and, thereby presumably enhances them" (Harvey & Green, 1993, p. 24), and these changes are usually measured against certain output criteria on a student or school level. In consequence, quality here can be seen as the discrepancy between a desired outcome or char-

acteristic and a certain status or input condition based on a certain evaluation criterion. Educational quality therefore depends on the objectives that an educational system or a school is supposed to fulfill, as they will define the output criteria. In this sense, the quality of structures (such as curriculum or opportunity to learn) and those of processes (which also could be evaluated on their own) here are rather seen as *effectiveness-enhancing factors* determining the outcome quality (Creemers, 1994) .

Although historically there were several distinct modes of teaching and education (see also section 2.1 about the education in the Gulf Area), a modern kind of school system has become prevalent in current global trends, exhibiting similar characteristics and objectives. Adick (1992, p. 244) calls this the "universalization of modern schooling" and described universal common characteristics such as: a differentiated school system which distinguishes between classes, levels, and so forth; teaching according to a prearranged curriculum, professionalized staff teaching at scheduled time intervals; and state-controlled regulated educational practices in schools.

According to Adick, the objective of modern education systems can be seen as fulfilling certain qualification, selection, and legitimization functions:

> The acquisition of sanctioned knowledge, rewarded with a certificate, becomes a form of cultural capital. This allocation of chances for a better life by means of the school seems to be basically legitimate in the sense that everybody believes in it. And what is even more challenging for analysis, this model of schooling is universally accepted (Adick, 1992, p. 244).

Fend (2006, p. 54) supported this notion of education systems being part of a universal project of modernity, and described four different social functions of an educational system and thus of a school: The *qualification* of the students; their *allocation* into the employment system, respectively into a social stratum; *enculturation* (referring to the reproduction of cultural capabilities and cultural comprehension of the world and the person); and the *legitimization* of the respective system of government.

The necessity of the qualification of the student body is obvious: for a society, it is important to have qualified members to make its economy competitive, and for an individual qualification

provides a better chance for good work conditions and high salaries. Fend (2006, p. 51) described job-related skills and knowledge therefore as the important educational outcomes on the level of the student body.

Nevertheless, questions remain regarding which criteria should be used to measure educational quality, and finally to determine the effectiveness of an institution or educational system according to the definitions listed above. While certain indicators, such as transitions to certain kinds of secondary education or university or the number of grade repetitions were initially used, it later was argued that decisions about promotion and referrals are influenced by other factors than education in a school or classroom alone (Creemers & Kyriakides, 2008). Therefore, cognitive criteria were preferred – mainly achievement in basic school subjects like mathematics, reading, or science. Creemers and Kyriakides (2008, p. 20), when reviewing the effectiveness literature, consequently stated: "The majority of current studies collected data from national tests in subjects areas like mathematics and languages." However, it can be argued that students in modern societies will need to learn more than basic skills in core subjects, leading in the direction of higher-order learning and metacognition. Levine and Lezotte (1990, p. 70), when reviewing achievement criteria to measure effectiveness, found that most tests assess "fragmented, lower order skills." While they acknowledged that these rather mechanical skills – such as basic computational skills – need to be mastered, especially in primary grades, they also argued for the necessity of including measures on higher order learning and thinking skills such as reading comprehension and mathematics problem-solving. They regarded an exclusive focus on "low-level learning" as harmful, as such a focus could result in stressing factors and practices that might unfavorably influence students' later achievement. However, the author also agrees with Creemers and Kyriakides (2008, p. 21) that basic learning and basic knowledge are required before higher-order learning and thinking skills can be developed. In consequence, especially in primary education, a certain focus on these basic cognitive outputs is still valid in modern societies.

Taking the different main tasks of schools listed above into consideration, the extension of education beyond the acquisition of cognitive knowledge and skills cannot be denied. Thus, social skills, problem-solving skills, and personal competences – such as responsibility and initiative-taking – are also regarded as increasingly important (Creemers & Kyriakides, 2008; Raven, 1991). Delors (1996, p. 212) for example, in his report for UNESCO, emphasized the importance of a good civic education "in the struggle against exclusion of all those who for socio-economic or cultural reasons find themselves marginalized in present-day societies." *Becoming*

a good citizen, but more in the sense of showing loyalty to the State and its leaders – as described by Fend (2006) as the social function of the educational system to legitimize the respective system of government – also was mentioned as an important quality criterion for an effective schooling in the Gulf Area (Khan, 2015).

While schools likewise can contribute to non-cognitive outcomes, studies have shown that the impact of education on these domains, which are usually less prioritized in the curricula, is often rather small (Gray, 2004; Opdenakker & van Damme, 2000). Moreover, research shows that *affective* and *cognitive* outcomes do not necessarily concur. Affective outcomes here are used in the sense of Knuver and Brandsma (1993, p. 190) as the students' attitudes towards school and learning. Their study on the relation of cognitive and affective outcomes indicated a reciprocal relationship wherein higher cognitive scores increase motivation and well-being, which in turn increase cognitive results (Knuver & Brandsma, 1993). Isac (2015, p. 139), who investigated effective citizenship education, likewise reported that for non-cognitive outcomes schools hardly would make any difference.

It is therefore argued here that, as similarly concluded by other researchers, using achievement measures in basic subjects (and consequently the approach chosen for this thesis) still has some justification in EER, especially on primary level.

The Equity Dimension

Apart from a perspective of measuring educational quality as achieving good results in certain outcome areas (described as quality or excellence), the question of the extent to which educational systems are able to reduce the differences or variance between different subgroups of students, independent from their antecedent conditions, can also be asked. In many educational systems, a certain compensation for different and non-malleable context conditions, in the direction of more equal opportunities in the labor market, is seen as an important function of the educational system. The OECD (2012) argues that both the quality and the equity dimension need to be regarded in order to obtain a high-performing education system, providing empirical evidence using the PISA 2009 (OECD, 2010a) results. They define the equity dimension as follows: "Equity in education means that personal or social circumstances such as gender, ethnic origin or family background, are not obstacles to achieving educational potential (fairness) and that all individuals reach at least a basic minimum level of skills (inclusion)" (OECD, 2012, p. 9). However, as discussed by Schwippert (2001, pp. 27–30), due to limitations in resources and a limited amount of time teachers will have to decide about the distribution of time and

attention to different students based not only on their motives and beliefs, but also social expectations and curriculum guidelines. Consequently, teachers will have to find a balance between the quality and the equity dimension when allocating their time. Heckhausen (1981) distinguished different kind of allocation strategies: the *need principle* [Bedürftigkeitsprinzip] where the focus of time allocation is on students showing a certain deficit with regard to an educational objective, the *justness principle* [Prinzip der Billigkeit] where the support is related to the achievement level of a student, and the *equality principle* [Gleichheitsprinzip] with equal allocation of time and attention to each student. Ultimately, the relation between these principles specifies the quality criterion applied.

In the early period of school effectiveness research, the equity dimension dominated in part, and strong movements tried to investigate *inequalities* among different student groups (see Jencks, 1972 and Edmonds, 1979) and launch school improvement projects especially for low-SES students ('the urban poor'). Unfortunately, results in this regard proved to be rather modest. Teddlie and Reynolds (2000) summarized research giving evidence for the existence of differential effects for schooling in terms of prior attainment, socio-economic indictors, gender, and ethnicity (known as *differential effectiveness research*); it seemed, however, that research results were rather inconclusive and especially did not clearly indicate that more effective schools – as defined in the classical sense – would contribute to a closing of the achievement gap. Summarizing the knowledge base, Kyriakides (2004) concluded that effective schools are able to promote learning of their students but may not have a special impact on disadvantaged students. On the other hand, there are certain rather consistent findings that "Children from disadvantaged backgrounds are likely to be more affected by their schools than other groups across all schools" (Chapman et al., 2015, p. 96). It should be noted, however, that there is still not sufficient understanding regarding which effectiveness factors may be responsible for these differential school effects.

Summarizing the above-mentioned findings, it can be concluded that dimensions of both quality and equity should be regarded for further projects in EER.

3.3 General Effectiveness Factors

3.3.1 Introduction

The main purpose of many educational effectiveness studies is to describe characteristics and processes that "add value" to student outcomes in order to help researchers and policy-makers

understand and finally overcome weaknesses in educational systems. It is therefore not surprising that a rich body of research concerning this area has been accumulated over time. One of the first to explicitly list a set of effectiveness-enhancing factors (the so-called *five-factor model* or *five correlates of school effectiveness*) was Edmonds (1979). Teddlie and Reynolds (2000) later expanded the list of basic effectiveness-enhancing factors based on an evaluation of comprehensive reviews of several hundreds of school effectiveness studies collected by Levine and Lezotte (1990) and by Sammons, Hillman, and Mortimore (1995). They identified the following nine global effectiveness-enhancing factors, which are also well summarized by Reynolds, Sammons, Fraine, Townsend, and van Damme (2011, pp. 17–18):

1. An effective educational leadership
2. A focus on academic outcomes and on maximized learning time
3. A positive school culture that involves a shared vision, an orderly climate, and a positive reinforcement
4. High expectations of students and staff
5. Monitoring progress at school, classroom and student level
6. Parental Involvement
7. Generating effective teaching through maximizing the learning time, grouping strategies, benchmarking against best practice, and adapting the practice to student needs
8. Professional development of staff
9. Involving students in the educational process

Although slightly different terms are sometimes used, other authors reviewing educational effectiveness factors (such as Marzano, 2003; Marzano & Kendall, 2006; Scheerens, 1992) report similar factors. Some authors add a few factors they regard over and above the before mentioned ones as essential. Scheerens (1992) for example, added *external stimuli to make schools effective*, *physical and material school characteristics*, *teacher experience*, and *school context characteristics*, while Cotton (1995) additionally regards *District-school interactions*, *special programs*, and *Equity* as important factors.

Many of those *global* factors also could be identified in the school effectiveness analysis conducted by Martin and Mullis (2013), which was based on 34 countries and three benchmarking participants administering PIRLS and TIMSS to the same fourth grade students. They concluded from their analysis that "…an effective school was safe and orderly, supported academic

success, had adequate facilities and equipment, was staffed with well-prepared teachers, had well-resourced classrooms, and provided effective instruction" (Martin & Mullis, 2013, p. 7). Albeit not all of the factors listed by Teddlie and Reynolds (2000) could be supported, these findings, from an international large-scale assessment administered in a large variety of different countries, in general provide support for earlier analyses with regard to effectiveness-enhancing factors functioning on a global level.

However, during the last decades, increasing interest has been devoted to investigating how different effectiveness-enhancing factors work depending on the context of the school under consideration, leading to the so-called *context-specific* models of educational effectiveness (Reynolds et al., 2011). Several authors examined the processes of effective schooling in schools with different average levels of socioeconomic status (SES; Hallinger & Murphy, 1986; Rowan & Denk, 1984; Wimpelberg, Teddlie, & Stringfield, 1989). These authors, among others, found that the level of SES does indeed influence the processes in schools. For example, it seems that parents from low-SES communities often prefer an emphasis on social and vocational education, while parents from high-SES areas put a higher emphasis on academic goals. In addition, low-SES schools in general experienced less parental involvement. This situation, in turn, was hypothesized to influence the activities and curricula offered to students. High-SES schools were often found to be more academically oriented, with curricula more specifically designed to promote cognitive learning. Teachers of higher SES schools were found to have higher expectations of students' academic success. Hallinger and Murphy (1986, p. 349) argue that "The combination of infrequent home-school contact and low academic expectations make the typical low-income school a less effective environment for learning cognitive skills." Their research reveals that instructionally effective schools are influenced by their environment, and adapt their strategies and processes accordingly. Effective low-SES schools isolated themselves from their environmental norms, and focused on the mastery of basic reading and mathematics skills. They developed a system of rewards intended to build up the academic self-esteem of their students. Principals exerted a strong administrative leadership, setting high standards for students and teachers. In contrast, effective high-SES schools were in general associated with a more open environment of high expectations. A high visibility of parents applying pressure for children to succeed changed the role of the principal to rather one of mediating the demands and expectations of the community (Hallinger & Murphy, 1986, p. 350).

Another important aspect that influences school effectiveness-enhancing factors is the national and cultural context. International large-scale assessments such as TIMSS, PIRLS, PISA, or the international school effectiveness study ISERP allow for the investigation of the international

dimension of educational effectiveness (see also section 3.1.2). Interestingly, findings suggest that some factors "travel" across countries, depending on the cultural context, while others don't. Reynolds (2006) for example, summarizing major findings from the ISERP study, found that many general effectiveness factors regarding classroom management, instruction, and climate did explain variation in student achievement in diverse countries. In particular, Reynolds found that specific teacher behaviors – such as *clarity*, *questioning*, *high expectations*, *a commitment to academic achievement*, and *lesson structuring* – could partially explain differences between more and less effective schools across the world. On the other hand, it seemed that certain school factors, such as the *quality of the principal*, while being an important factor in all countries under investigation, travelled conceptually – meaning that the leadership style mattered by context. For example, Reynolds reported that leadership is more directive in Asian cultures, while it is more lateral/ vertical in the Western societies.

The subsequent sections describe major factors that were identified as being associated with student achievement and indicate empirical evidence from previous studies, reviews, and meta-analyses. While some of the factors are operating from outside (extrinsic), and thus are susceptible to policy interventions, others are inherent in nature (intrinsic) and thus cannot be easily altered. While EER is often more interested in malleable factors on school and classroom levels, both groups of factors are interlinked and both are important in predicting achievement. All will be discussed in the following sections. Two factors (*time on task* and *opportunity to learn*) that are for several effectiveness frameworks considered as essential elements on each educational level (eg., Creemers, 1994; Creemers & Kyriakides, 2008; Scheerens, 1992) are discussed across all levels at the beginning of sections 3.3.2 and 3.3.3. Subsequently, intrinsic student-level factors will be discussed in section 3.3.4, followed by class-level factors in section 3.3.5 and school-level factors in section 3.3.6. The chapter will be concluded by a short overview on context-level conditions for effective schooling in section 3.3.7. All of the factors reviewed in the sections below constitute the basis for the conceptual framework of this research project that will be developed in chapter 6.

3.3.2 Time on task

Time on task refers to the time students are willing to spend on learning and on educational tasks" (Creemers & Kyriakides, 2008, p. 100). This and associated concepts are identified differently by different scholars; for example, *academic learning time* by Creemers (1994, p. 28) or Scheerens and Bosker (1997), or *effective learning time* by Scheerens (2016, p. 112).

Time on task dependents on student motivation and expectation but also on the amount of time offered for learning to students by the school and especially by the teachers. The general concept of time on task has received criticizism, for example, by Gage (1978, p. 75): "because of its psychologically empty and quantitative nature." However, the author agrees here with Creemers and Kyriakides (2008) that these criticisms don't affect the concept of time on task itself; rather, they imply that in addition to the time factor, the question of which activities are offered and what learning processes are taking place needs to be considered. Consequently, this factor is closely related to the factors described in the subsequent sections: *opportunity to learn* and *quality of teaching* – or as Creemers and Kyriakides (2008, p. 100) state: "It is also important to note that time on task refers to the time during which students are really involved in learning, provided that this time is filled with opportunities to learn."

In his definition of academic learning time, Creemers (1994, p. 29) identifies four different aspects showing the different levels on which the variable is operating and its relation to the concept of opportunity to learn. He distinguishes between the *allocated time* (learning time allocated by teachers), *time on task* (the time students are really involved), *student error rate* (level of difficulty of tasks), and *task relevance* (relevance to a certain part of the curriculum). This emphasizes again that concepts of time on task and opportunity to learn are operating closely together.

At the student level, the conceptualization of time on task is somewhat challenging, as direct observation is usually not possible, or at least difficult. Therefore often proxies are used, such as the time spent on homework (Cho, 2010; Kyriakides, 2005; Kyriakides, Campbell, & Gagatsis, 2000; Neuschmidt & Aghakasiri, 2015), the time spent on private tutoring (Cho, 2010; Kyriakides et al., 2000; Kyriakides, 2005), or on learning related out-of-school activities (Cho, 2010). Additionally, indicators related to student absence are used (de Jong, Westerhof, & Kruiter, 2004).

The amount of time spent on homework is a proxy which is used in many educational effectiveness studies for time on task, while in other studies it is used for opportunity to learn. This concept merits further discussion. Moreover, the empirical evidence of relations between the amount of time spent on homework and higher educational outcomes is rather mixed. Cooper, Robinson, and Patall (2006), in their meta-analysis, found some evidence of a homework-achievement correlation for secondary schools in the United States (Cooper, Lindsay, Nye, & Greathouse, 1998) and came to similar conclusions in a separate study. Neuschmidt and Aghakasiri (2015) indicated a significant relation between amount of homework and achievement

in Oman. Conversely, other authors found no correlation (Kyriakides, 2005) or indicated contradicting results, or even negative correlations on student level, in certain models (Cool & Keith, 1991). Looking at international large-scale assessment data, mixed results can also be found: Based on the PISA 2012 results, the OECD (2014a) reported that for most of the countries spending more time on doing homework tends to be associated with higher PISA scores. They also indicated, based on analyses of PISA 2009 data, that the effect decreases with the amount of time spent, reporting that after around four hours additional time spent on homework, it only had a "negligible impact on performance." However, Dettmers, Trautwein, and Lüdtke (2009) analyzing PISA 2003 data from 40 countries, could not establish a clear-cut relationship between homework time and achievement in their multilevel analyses. In TIMSS 2011, the relation between the amount of homework and achievement are reported to be more "mixed"; this can be explained by the different objectives homework can have: While in some cases it is given to students in order to keep up with their classmates, in other situations it as given for practice or as an enrichment exercise. However, it was found for most countries that in the 8th grade, students who reported doing homework for over 45 minutes, but below 3 hours, achieved the highest mathematics and science achievement on average (Martin et al., 2012, p. 418; Mullis, Martin, Foy, & Arora, 2012, p. 402).

It becomes apparent that the objective of homework assigned by teachers differs between student groups, grades, and possibly subjects – leading to varying results in relation to student achievement.

Other aspects of a more methodological nature should also be considered. Cool and Keith (1991) for example, raise questions about the validity of the homework variables in use. They conclude that a homework indicator that is "based on a single general question about normal homework practice, is probably an unreliable measure of true homework practice" (Cool & Keith, 1991, p. 40). Trautwein (2007) also argues that it is important to clearly distinguish between effects on an individual level, as discussed here, and those on a classroom level – a distinction which he sees as unfortunately not having been taken into account by many studies. Moreover, some researchers argue that the effect of homework on achievement might be attributable to a "common cause," thus possibly decreasing once the models control for variables such as motivation, prior ability, quality of instruction, tracking, or home background (Cool & Keith, 1991; Dettmers et al., 2009; Trautwein, 2007). Other researchers indicated that at least on a class level, other factors – such as the frequency of homework or the number of tasks – might be more important than the amount of time spent on homework alone (de Jong et al., 2004; Trautwein, Köller, Schmitz, & Baumert, 2002).

In summary, it can be concluded here that homework as an indicator for *time on task* might not be a reliable measure, and thus should be avoided if possible.

Concerning other out-of-school activities, those related to activities within schools are especially found to have a relation to student achievement. For example, Anderson, Wilson, and Fielding (1988) had 155 5th Grade students record their outside-school activities for a period of between eight and 26 weeks. They found that "reading books" was the best predictor for student's reading ability. Similarly, Mullis, Martin, Kennedy, and Foy (2007), analyzing the PIRLS 2006 data, found that "On average internationally, and in most countries, students who reported reading novels and short stories most frequently had higher average achievement than those who read less frequently"; Won and Han (2010) reported associations between reading behavior and mathematics achievement using TIMSS 2003 data. In contrast, non-academic out-of-school activities, such as "listening to music", "watching television", or "playing computer games" were repeatedly found to be negatively associated with academic performance if an extensive amount of the daily leisure time was spent in such activities (Anderson et al., 1988; Martin et al., 1997, 1997; Mullis et al., 1997).

At the class level, students' time on task, which is defined as the time students spend actively learning, will next to student compositional factors and classroom environment related factors also be determined by the actual time spent on teaching by teachers (the *instructional time*) and is in general closely related to classroom management (see also section 3.3.5.2). In this regard, effective teachers are characterized according to their ability to direct their classrooms and the environment therein; teaching environments that are effective, therefore, are characterized as those in which "academic activities run smoothly, transitions are brief, and little time is spent getting organised or dealing with inattention or resistance", as per Brophy and Good (1986, p. 109). On the other hand, the extent to which students are engaged in the activities led by their teachers, or rather distracted by off-tasks activities such as social interaction, is also an important question. With the exception of studies which make use of classroom observations, the measurement of student attentiveness is not strictly feasible; therefore, analyses have been more focused on the investigation of the relationship between instructional time and achievement. Findings in this regard are not unambiguous: as previously mentioned, the question of how instructional time is used is likewise important; this, in turn, depends on additional factors such as the opportunity to learn (for example the quality of the curriculum and instructional materials) and the quality of teaching (hence the use of instructional approaches). Lee and Barro (2001), analyzing cross-country achievement data after controlling for a variety of school resources, found inconsistent results for the relation between school-term length, while Wößmann

(2003), in a similar study using TIMSS 1995 data, found significant (albeit small) effects. Lavy (2010), however, analyzing the PISA 2006 database and additional Israeli data, reported modest to large effects associated with one more hour of weekly instruction on average. Using data from TIMSS 1995, Martin, Mullis, Gonzalez, Smith, and Kelly (1999) reported that in high-performing countries, students tend to spend more time in schools and have more instructional time than in lower-performing countries. Relevant analyses were carried out by Sandoval-Hernández, Aghakasiri, Wild, and Rutkowski (2013) on PIRLS 2006 data from 45 countries. While the authors didn't find a consistent relation between the yearly overall schooling time and reading achievement, they found a far stronger relation in many countries when correlating solely the *effective teaching time* (the time the teacher spent to instruction as opposed to time spent on administration and other tasks) with student achievement. This finding again gives clear indication that the amount of time is not necessarily a factor on its own, but rather should be regarded in conjunction with other important, interrelated factors, such as the opportunities to learn and the quality of teaching. It should be noted that for the current analyses, due to the absence of suitable data, only indicators for the overall available time can be created, but not specifically for the amount of time the teacher actually focuses on instruction.

At the school level (often based on policies implemented on a regional or national level) the time for learning is mainly determined by the time scheduled for instruction, depending on the duration and amount of lessons per subject, and the school days per year. It should be noted that the prescribed time for learning might differ significantly from the actual amount of time students are taught because of external circumstances such as unplanned school closings, for example due to severe weather conditions, civil unrest, teacher absenteeism, etc.

3.3.3 Opportunity to learn

Opportunity to learn considers the fact that students need opportunities to acquire knowledge and skills, in addition to the time spent on tasks alone.

Opportunity to learn, therefore, is related to the actual content that is taught and learned – based on the curriculum, which is usually defined on country or regional level. The content and skills defined in the curriculum have to be incorporated in the curricular material (such as textbooks) and need to be presented by the teachers. Opportunity to learn, therefore, can be generally understood as the alignment of classroom practices with the curriculum and concentrates on the extent to which those practices cover the tests designed to monitor performance (Scheerens, 2016, p. 55). The concept was introduced in the early IEA studies FIMS and SIMS, which were

conducted in 1964 and 1980-1982, respectively. Opportunity to learn has applicablility relating to different levels of education. On a contextual level (usually a national or regional level), decisions about general learning objectives and content are made; this is usually referred to in IEA studies and elsewhere as the *intended* curriculum. A level below, the *implemented* curriculum refers to what is actually taught in the schools/classrooms, largely impacting the opportunity to attain the goal specified in the curriculum. Finally, the *attained* curriculum relates to the formal learning experiences of students. The attained or *experienced* curriculum refers to the knowledge and skills achieved. More information regarding these curriculum concepts can be found for example in Travers & Weinzweig's *Studies in mathematics education series: Vol. 11* (Travers & Weinzweig, 1999).

The concept of opportunity to learn, as relating solely to the content of education, has been expanded in more recent policy debates, especially in the U.S., by integrating process indicators looking at how the content was presented and who presented it (McDonnell, 1995). In doing so, the concept became partly mixed with other dimensions – being associated with, for example, the quality of education and time; the focus therefore has somewhat shifted towards accountability and policy issues.

In more recent applications, opportunity to learn is defined as "The opportunities which schools provide students to learn what is expected of them," especially regarding their learning and progress concerning information for which they will be held accountable (Herman, Klein, & Abedi, 2000). Following this concept, opportunity to learn would comprise the following categories according to Boscardin et al. (2005, pp. 309–311):

- Curriculum content with the dimensions content coverage (the extent to which students cover the curriculum for a certain grade level or subject), content exposure (the time devoted to instruction and the depth of teaching), and content emphasis (defines the topics that are selected for emphasis and the emphasis on lower or higher order skills)
- Instructional strategies including the quality of instructional delivery (presentations of the lessons)
- Instructional resources (whether there are appropriate resources to prepare students for success)
- General assessment preparation

While all categories listed above are considered to be important determinants of educational effectiveness, this thesis will rather follow the originally defined concept of opportunity to learn, similarly to Creemers (1994) or Creemers and Kyriakides (2008), and will cover categories added later – such as *instructional strategie*s and *assessment preparation* – instead under the header of *Quality of teaching*.

Opportunity to learn, in the original sense, is usually measured by checking whether topics presented in a test were also present in the students' education. The IEA study TIMSS asks about the perceived preparation level of teachers concerning various topics presented in the assessment, as well as for specifics regarding when and for how long a certain topic was taught/ introduced to the sampled students. Likewise, emphasis given to each subdomain is measured. As teachers do not always follow the curriculum prescribed, classroom observation would be a more valid technique to assess the content coverage, according to Creemers (1994).

The concept of opportunity to learn has been included in several IEA studies (Comber & Keeves, 1973; Postlethwaite & Wiley, 1992), but also in other research projects. For example, Boscardin et al. (2005) and Wang (1998) found the opportunity to learn to be related closely to achievement in different subjects. Jones, Davenport, Bryson, Bekhuis, and Zwick (1986), when reanalyzing the High school and Beyond study data, found that the level and number of courses were strongly related with improvements in student outcomes, especially in mathematics, even after controlling for student background factors and aptitude. These conclusions, as well as the findings from the IEA SIMS study, led McDonnell (1995, p. 308) to conclude that "…curriculum exposure could be an effective lever in efforts to improve student achievement and to distribute learning opportunities more equitably."

Opportunity to learn can also be regarded from a perspective of social equity. A curriculum may differentiate between different student groups when implemented via tracking on school level or ability grouping on class level. This, in turn, can limit or enhance access to the content to learn – producing different learning opportunities for different groups of students. In countries with a tracked student system, opportunities to learn have been found to be more closely related to student achievement (Creemers & Kyriakides, 2008); Oakes (1990) also found a strong relation between social groups and course level. Minority students, for example, were usually placed in low-track classes, leading to a lack of equal opportunities for different student groups. Oakes (1990) opines that these differences in opportunity limit instructions, and therefore argues against tracking systems and ability grouping.

In summary, it can be concluded that the concepts of time on task and opportunity to learn are closely interrelated, and that the definition of the latter concept in particular varied over time and among researchers. However, there seems to be common agreement that, in addition to the amount of time students are actively involved in the learning process, certain opportunities also need to be available to allow for effective learning. Moreover, factors related to the quality of the instruction and the learning environments are also fundamental to an effective learning process. These dimensions will be discussed in subsequent sections of this chapter.

3.3.4 Student-level factors

The following section discusses two important student factors that (contrary to time on task or opportunity to learn) are intrinsic in nature and therefore cannot be easily and directly altered by educational policies: *aptitude* and elements of the *affective domain*. These factors are nevertheless interrelated with more extrinsic factors that are malleable, and there are indications that changing external factors also might have an impact on the intrinsic factors discussed below.

3.3.4.1 Aptitude

The term *aptitude* is described and used a bit differently by different authors. Usually, it comprises a component of both general intelligence and prior knowledge, indicating what the student already knows about a certain subject. Often used interchangeably with the terms *ability*, *prior knowledge*, or *prior achievement,* aptitude is often seen as an important factor with impact on achievement and is regarded as an important controlling variable in EER to disentangle effects of the home background from the effects of schooling (Teddlie & Reynolds, 2000, p. 264). Consequently, there is strong empirical evidence for relations of aptitude with achievement (Reynolds, 1991; Reynolds & Walberg, 1991).

Aptitude is often conceptualized as a kind of test score from a (more or less) standardized test, or as another indicator of achievement in the early stages of learning. Carroll (1963), in his model of school learning, conversely conceptualizes aptitude as the amount of time needed to learn under optimal instructional conditions. It is evident that students with a higher aptitude need less time to make educational progress. A possible explanation for the association of aptitude with higher achievement gains are given for example by Hegarty-Hazel and Prosser (1991b) for physics and for Hegarty-Hazel and Prosser (1991a) for Biology. They concluded from their studies that prior knowledge led to an adoption of more effective study strategies, and therefore to higher achievement in physics and biology.

3.3.4.2 Affective factors

The *affective* domain refers to a wide range of beliefs, feelings, and moods beyond cognition. While there is a vast body of research in this area, strong theoretical foundations seem to be lacking, and different and interrelated concepts are neither clearly defined nor distinguishable from one another. This could be due to the fact that concepts in the affective domain are more difficult to depict and to measure when compared with cognitive factors (McLeod, 1992); however, they are acknowledged as central concern in the field of teaching and learning.

Largely adhering to the categorization and definitions of McLeod (1992), who tried a reconceptualization of the research on affect in mathematics, the term *affec*t here is used in a more general sense, as a superordinate concept comprising more specific dimensions such as *beliefs*, *attitudes*, and *emotions.* In the context of education, researchers mainly focus on attitudes towards certain subject areas, often mathematics.

Beliefs

Fishbein and Ajzen (1975, p. 131) define the term *belief* "as the subjective probability of a relation between the object of the belief and some other object, value, concept, or attribute." Thus, beliefs refer to an individual's understanding of the relation between him- or herself and his or her environment. The process of developing subject-related beliefs is assumed to be strongly influenced by the cultural setting and the context in which learning takes place (Schoenfeld, 1989).

Beliefs can be categorized according to the object of the belief, for example beliefs related to a subject (such as mathematics), beliefs about one's self, beliefs about teaching and learning, and so forth. Research related to subject-related beliefs and beliefs about the self have in particular received considerable attention in the past. Beliefs about the self mainly include *self-concept* (the individual's perception of self) and *self-confidence* – with the latter regarded as a component of a more general self-concept (Reyes, 1984, p. 559). In this area, substantial gender differences have been found, as reported by McLeod (1992). In terms of learning mathematics, for example, the author indicates that boys in general are more confident than girls, even when girls are performing higher.

Attitudes

Attitudes "refer to affective responses that involve positive or negative feelings of moderate intensity and reasonable stability" (McLeod, 1992, p. 581). This definition is in agreement with

Koballa (1988), who reviewed different definitions of the concept of attitude and described the common underlying element as a favorable or unfavorable feeling towards a specific object. Attitudes are often surveyed by means of questionnaire items which ask questions regarding whether respondents like or dislike a certain subject, or are curious about or bored by it. While conceptually, attitude is closely related to *value*, the latter is seen as more broad in nature, and more persistent (Koballa, 1988). On the other hand, the concept of attitude is also closely related to belief and some authors incorporate beliefs – as a component – into a more general concept of attitude (McLeod, 1992) .

Another closely related and partly overlapping concept used in educational research is that of *achievement motivation,* which is regarded as being affected by some of the components described above, such as attitudes towards learning and self-confidence. The following important components related to motivation are often distinguished: *intrinsic values* or *interest* (in which an activity is done because it is enjoyable), *extrinsic motivation* or *utility value* (in which something is done because it leads to a desirable outcome), and *ability belief* (which refers to self-concept and the attribution of failure and success to individual's ability; Mullis, Martin, Foy, & Arora, 2012; Wigfield & Eccles, 2000).

Emotions

Emotions seem to be less researched in education, probably largely because researchers were more interested in stable factors that easily can be measured by questionnaires. Emotions are seen as a type of affective response that may vary quickly, and therefore are less stable when compared to beliefs and attitudes (McLeod, 1992, p. 578).

Mc Leod summarized the body of research in regard to mathematics education by elaborating on three major facets linking the different affective responses of students:

> First, students hold certain beliefs about mathematics and about themselves that play an important role in the development of their affective responses to mathematical situations. Second, since interruptions and blockages are an inevitable part of the learning of mathematics, students will experience both positive and negative emotions as they learn mathematics; these emotions are likely to be more noticeable when the tasks are novel. Third, students will develop positive or negative attitudes toward mathematics (or parts of the mathematics curriculum) as they encounter the same

> or similar mathematical situations repeatedly (McLeod, 1992, p. 578).

There is a strong body of research investigating the association between affective factors and academic outcomes (Papanastasiou, 2000; Papanastasiou & Zembylas, 2002; Reyes, 1984; Wang & Staver, 1996). However, the influences are understood to be bi-directional, with affective factors and achievement affecting each other. Reyes (1984), summarizing the body of research related to self-concept and achievement, reported consistent, positive correlational associations; he indicated support for causal effects of self-concept on achievement, but partly also indicated certain support for the opposite direction. Consequently, Papanastasiou (2000) refers to student's perception about the value of learning mathematics as both, an input and an outcome variable.

For TIMSS, Mullis, Martin, Foy, and Arora (2012, p. 326) also confirmed that "Each successive TIMSS assessment has shown a strong positive relationship within countries between student attitudes toward mathematics and their mathematics achievement." However, from previous cycles, across countries, the tendency of some of the highest-performing countries (especially in East Asia) to have the smallest percentage of students reporting positive attitudes towards learning mathematics persists. The same basic findings emerged from the analyses of the TIMSS science outcomes (Martin et al., 2012).

Similarly, in his synthesis of over 800 meta-analyses, Hattie (2009) also reported positive relations between affective factors and achievement. While medium-effect results were reported for motivation and self-concept, they were a bit lower for attitudes towards mathematics and science.

3.3.4.3 Social background of the students

Not only school-related factors, but also out-of-school factors, and especially the family or home background of the student should be considered when investigating EER – as a substantial amount of the time spent outside schools is shaped by a child's family context. Parents or guardians necessarily influence their children's opinions and attitudes towards education and learning. Moreover, they also directly influence their opportunities to learn. It was found that socio-economic characteristics, like economic and cultural resources, or prestige indicators such as the parental profession and the education of the parents (among others), are important predictors for educational aspiration, later competencies, and later academic success in school in terms of educational attainments more generally (Sirin, 2005).

When explaining success or failure based on student's family background, important theoretical contributions can be found in cultural and social reproduction theories (for example Bourdieu & Passeron, 1977) and social action theories (for example Boudon, 1981). These are based on an underlying sociological concept in which societies can be described as social structures, which are then stratified into groups (or classes) based on certain similarities of their members. Members of a group share common traits, and might fill specific positions within the society. Each individual, family, or group can be classified within a given society and class based on certain dimensions, according to their control over attributes of social value – such as wealth, prestige, or power. The relative position of the person or group within the hierarchical social structure can thus be defined as the *socioeconomic status* (SES; Mueller & Parcel, 1981, p. 14).

It is widely believed that members of a certain social class will reproduce the class itself: as cultural values, norms, and attitudes of parents and the wider family context are, to a great extent, passed on to the child in a process that is partly intentional and partly subconscious. Consequently, this kind of *cultural reproduction* leads to the process of transferring aspects of society (the social class) from one generation to the next. This process of sociocultural reproduction has been described by Bourdieu and Passeron (1977), among others, and can be regarded as an important factor which influences student learning beyond influences from the formal education system.

While Marx (2012), writing from an economical perspective, differentiated between only two *classes* which are distinguished from each other in terms of their access to or lack of the *means of production,* in turn defining their access to power or the lack thereof, Bourdieu (1986) used the term *social spaces* instead of classes – and his conceptualization regarding *capital* is more refined. He argued that the functioning of the social world only can be fully understood if not only the *economic* form of capital, but also more immaterial resources, which he defined as *cultural capital* and *social capital,* are also recognized (Bourdieu, 1986, p. 46). The following sections provide a short overview of the conceptualization of the different forms of capital described by Bourdieu.

Economic capital

For Bourdieu (1986), *economic capital* comprises the economic resources to which an individual has access. This concept includes material resources, such as income, as well as material goods and assets that can be easily converted into money. Variables referring to the possession of household items, for example, could serve as indicators of the family's economic capital. While Bourdieu regarded economic capital as being at the "root of all the other types of capital"

(Bourdieu, 1986, p. 54), he also argued that economic capital has no influence independent from other forms of capital. Thus, the availability of economical capital allows parents to pay for better schools and extracurricular activities for their children (Graaf, Graaf, & Kraaykamp, 2000, p. 93).

Cultural Capital

Cultural capital, on the other hand, refers to informal interpersonal skills, habits, manners, linguistic styles, tastes, and lifestyles. Bourdieu (1986, p. 47) here distinguished between three interrelated states or types of cultural capital. Firstly, the *embodied* cultural capital describes the persistent attitudes of the mind and body which depend on class and society. It also comprises knowledge – either consciously acquired, or inherited by socialization – of culture and tradition. Secondly, *objectified* cultural capital refers to material objects, such as paintings or musical instruments, to which society allocates value and esteem. *Objectified* cultural capital can be easily transferred to other persons or exchanged to a form of economic capital. Finally, *Institutionalized* cultural capital represents an institution's formal acknowledgement of an individual's cultural capital (e.g., academic qualifications or credentials).

Cultural capital, in its embodied state, can be transmitted via cultural socialization processes to later generations. Accordingly, parents that embody a higher cultural capital will be more able to socialize their child with forms of communication, attitudes, and behavior, which often better fits learning behavior in school. Bernstein (1971) also emphasized the importance of cultural capital for successful school career. He argued that different classes also embody different *linguistic codes* which can help to explain achievement inequality between different population groups. Vis-à-vis education, the assumption is that students from middle classes can handle more elaborated codes – and consequently are more likely to perform better in the education system, because schools are relatively anonymous institutions that need to use more elaborated code, as they are concerned with the introduction of new knowledge which goes beyond existing shared meanings (Atherton, 2011).

Social capital

Social capital, according to Bourdieu (1986, p. 51), describes the social network of a person (which can be institutionalized or informal) and his or her group relations. Network connections, to Bourdieu, are the product of a constant effort; consequently, the transfer or reproduction of social capital needs expenditure of time, effort, and economical capital.

As per Bourdieu, the different types of capital can be transformed into each other to a certain extent, but at the risk of some loss and at the expense of time and energy. Economic capital, for example, can be transferred to cultural capital by investing in the education of the next generation. The next generation, in turn, may – via better positions and higher salaries – be able to convert this cultural capital back into economic capital. Similarly, a higher cultural capital could lead to certain behaviors and communication skills allowing for the development/extension of an individual's network; this, in turn, could give access to certain positions and professions which otherwise could not be obtained. Economic capital and time can also be invested to expand an individual's network, which might be of later benefit, thus increasing economic capital.

In the opinion of Bourdieu, the reproduction of social injustice is happening in a less obvious way than for Marx, via the investment of economic capital of one generation in the education (in other words, in cultural capital) and social capital of the later one. This reproduction process leaves social classes segregated and impermeable; consequently, different schooling opportunities remain.

In explaining influences stemming from student background on achievement and education aspirations, Boudon (1974), a prominent representative of the *social action* theories, also provided significant contributions. Boudon regarded the stratification of the society as both the cause and consequence of differences between members of society – which would also affect their education. He summarized this concept in this way: "The lower the social status, the poorer the cultural background – hence the lower the school achievement, and so on" (Boudon, 1974, p. 29). Boudon called this the *primary effect* of origin. However, he also defined another important component that would affect the students' educational opportunities: the *secondary effect* of origin. Boudon argued that decisions concerning a specific transition from one level of the education system to the next – independent from the actual student achievement – is also dependent on an individual's evaluation of economic and social costs and benefits, which in turn is dependent on the social status of a family. For the upper classes, the relative costs for higher or prolonged education are lower, while the benefits are regarded as higher. In contrast, lower class families would need to spend far more in terms of effort and resources to select higher tracks and longer education. Accordingly, Becker and Lauterbach (p. 19) concluded that unequal educational opportunities in the different social strata are based on an evaluation of advantages (benefits) and disadvantages (costs) of further education and higher education.

Coleman (1988), another important author who elaborated on the theoretical construct of social capital, aimed at integrating elements of the two aforementioned strands which describe social

action as being either nearly entirely formed by the social context (Bourdieu) or, on the other hand, as being shaped by independent actors who act according to self-interest in order to maximize utility (Boudon). Coleman's definition of *social capital* reads: "Social capital is defined by its function. It is not a single entity but a variety of different entities, with two elements in common: they all consist of some aspect of social structures, and they facilitate certain actions of actors … within the structure" (Coleman, 1988, p. 98). He distinguished three dimensions of social capital: the *level of trustworthiness* of the social environment, through obligations and expectations held by their members; the *information-flow capability* of the social structure; and the *implementation of effective social norms and sanctions*. He stressed the unique characteristic ability of social capital to benefit not only the individual but also the "public good" (Coleman, 1988, p. 119).

The different theories discussed above indicate that mechanisms explaining the influences of student background on achievement are manifold and work, in part, indirectly.

Empirical Evidence

As expected from theory, many different aspects – mainly of economic and cultural capital – have been found to be associated with student achievement outcomes.

Sirin (2005), for example, reviewed in a meta-analysis literature regarding the relation between SES and academic achievement, including 74 independent samples with altogether more than 100,000 students. He found medium to strong relations between various SES variables and measures of educational achievement. Likewise, Martin and Mullis (2013), in their educational effectiveness analyses on 34 countries and benchmarking entities who participated in TIMSS and PIRLS 2011 with the same students, reported that their home resources for learning variable (an index based on typical variables serving as SES indicators such as the parental education, the highest parental education level, or the number of books at home) was the strongest predictor for achievement in all subjects in nearly all of the countries (Martin & Mullis, 2013, pp. 136–137). Similarly, the OECD reported for PISA: "A consistent finding throughout PISA assessments is that socio-economic status is related to performance at the system, school and student levels" (OECD, 2016a, p. 205).

However, while often inspired by Bourdieu's theory of capital, researchers use quite different variables to measure the student's social background, and disagreement about the conceptual meaning of SES remains (Sirin, 2005, p. 418). According to Sirin, there is nevertheless some agreement that parental income, parental education, and parental occupation can be seen as the

main components of the conceptualization of SES. While current research focuses on the effects that are attributed to a student's learning environment, the influences of the home background on student performance must still be considered and extracted by forming a theoretically well-founded, regionally-appropriate indicator – as this procedure alone will allow investigation of the influence of school factors on students' achievement by disentangling school effects from the effects of the students' home background. A detailed approach regarding the question of how an indicator of student background which is more in line with the theories discussed can be conceptualized will be described in section 8.4.

The societal structure of GCC countries

That social stratification in the GCC countries is, to a certain extent, based on different criteria than in the West needs to be taken into account. Instead of classification based on the availability of certain forms of capital, the main principle for social stratification in the GCC countries is the affiliation to the ruling family (Colton, 2011, p. 1). While a system of ruling families that was based on societal norms and power structures already existed in the 19th century, the author argued that the current unique position of the Gulf State rulers (colonial influences notwithstanding) can mainly be attributed to their economic power, stemming predominantly from revenues of the oil industry. Colton (2011) asserts that while the leaders of the GCC countries are not dependent on their citizens for income anymore, they still require that their people regard them as legitimate rulers. In order to keep stability and maintain rule, therefore, they distribute much of their wealth in the form of employment and other gratifications – especially to those individuals and groups closest to them. This system, however, discriminates against the non-national population; consequently, the question of whether an individual is regarded as a national of the country of residence becomes one of the most important factors in determining the individual's place in society. Together with the implementation of so-called *nationalization policies,* these societal norms may signal to national youth that they are entitled to a job "by virtue of their nationality" (Ridge, 2014, p. 151); this, in turn, could also be an important explanatory factor when considering why education is viewed as less important for a larger share of the national population, and is consequently also less valued than elsewhere.

Within the national population, further divisions can be made by religion, tribal connections, and regional location – with each group and geographic location having a different proximity to the ruling family (Colton, 2011). Saif (n.d., p. 24), in his detailed class analysis of the Middle East, distinguished between the following traditional classes: a ruling class, a bureaucratic class, the bourgeoisie, the clerics, the traditional working class, the peasants class, and the nomadic

class. However, he also stated that these classes partly overlap vertically, between different ethnic groups, and horizontally, by occupation and capital formation. Moreover, Saif states that with the frequently-changing political and economic situation, certain changes in social stratification also took place, and classes were reordered accordingly. For example, wealth is now distributed mainly between the ruling, the bureaucratic, and the cleric class, at the expense of the lower middle class and the peasants (Saif, n.d., p. 24) and the power of local sheiks are currently decreasing (Colton, 2011, p. 40). However, for entrepreneurs and merchants as well, the accumulation of wealth still is primarily determined by connection to the power centers of the state – instead of by innovation in industrial or productive development (Farsoun, 1997, p. 19).

More so than in Western societies, stratification by gender must also be taken into account in the Gulf region, as for a long time good education for girls was not seen as a necessity. While in the past couple of years women have received more equal access to primary and secondary education, in certain (especially technical) fields they still lack equal opportunities in higher education and even more in the working environment.

The lowest social class consists of non-nationals, who are farthest from the ruling family and therefore only hold minimum rights and benefit least from the welfare states. Immigrant workers are only measured by their economic value for the society, with those having higher education and skills granted more rights and privileges (Colton, 2011, p. 40).

Nevertheless, it should be kept in mind that the region is currently undergoing rapid transformation away from the oil industry and towards diversification; the objective of building Western-style "knowledge societies" is recognized by state leaders as one of the main drivers for further economic development and for participation in the globalized competitive market (Alshumrani, Alromi, & Wiseman, 2014; Hvidt, 2016). This transformation process can also be expected to further impact social stratification and the value of education in the region. Governmental interest in Western-style education is currently increasing, particularly regarding private schooling that follows "international curricula"; meanwhile, parents are becoming more aware of the higher quality of private schools, for which they are willing to pay higher fees (Ardent, 2015, p. 12).

Some similarities with educational contexts of the West can be noticed, particularly for those societies with a high share of immigrant labor. Ridge, Farah, and Shami (2013), for example, investigated male dropout rates from secondary schools in the United Arab Emirates. They

found many similarities with the conditions for dropouts all over the world, listing a low-socioeconomic background with poorly educated parents, and a lower amount of economic and learning resources (next to poor-quality teaching), as the main reasons for student dropout. They also found that the employment situation of the father had great impact on the dropout rate. Hence, Ridge et al. (2013, p. 14) concluded: "The lower levels of educational degrees attained by the parents of dropouts are likely to have been transmitted to the dropouts in the form of attitudes towards schooling or simply modeling of the parents."

Smits and Huisman (2012), who studied the dropouts on primary school level in six Arab countries (Algeria, Egypt, Morocco, Syria, Tunisia, and Yemen), additionally included traditional SES factors such as measures of wealth, education, occupation, and family composition, as well as wider context factors such as the district level of modernization, educational facilities, and patriarchy in their analyses. While they also stated that "the status attainment process is at least partly driven by the same factors" as in other countries (Smits & Huisman, 2012, p. 16), they still noted a lower relative importance of those factors in the analyzed Arab countries. Instead, a higher variation was explained by the context in which the children live. They concluded that "in the Arab world, the environment where children are born determines their educational chances to a much larger extent than in Western countries" (Smits & Huisman, 2012, p. 17).

Interestingly, when looking at the TIMSS Home Resources for Learning scale, which was constructed as an indicator for SES, it can be seen that the variance explained in terms of mathematics and science achievement in Arab countries is indeed generally among the lowest of all participating countries. While the international average is approximately 10% explained variance for mathematics and close to 11% for science, the range of the Gulf countries' explained variance for mathematics is only between 1% (in Saudi Arabia) and 9% (in the United Arab Emirates). For science, the range of explained variance is between 2% (for Kuwait and Saudi Arabia) and 8% (in the United Arab Emirates). These comparisons might indicate that SES indicators which are based mainly on economic and cultural capital might work to a lesser extent in more traditional Gulf societies with a lower share of non-nationals. It can be concluded from the aforementioned section that careful analysis of student background in the GCC countries requires that additional components, such nationality status and gender, as well as the broader context of the student background beyond the family, be taken into account.

3.3.5 Class-level factors

While EER was primarily focused on a school level, later research based on multilevel models (for example, Hill & Rowe, 1996) showed that often a far greater proportion of variance can be explained at a classroom level. More recent analyses therefore give investigation of classroom-level factors a more central role, integrating findings from the formerly rather independent strand of *teacher effectiveness research*. Similarly, integrated educational effectiveness models frame the classroom level as the main determinant of educational outcomes, with the other levels of education holding less central positions – rather, as responsible for setting the pre-conditions for learning. As Creemers (1994, p. 5) stated: "From a theoretical and empirical point of view, the classroom is the predominant place in the school where learning and teaching takes place, and in this way the classroom level is more important for learning and outcomes than other levels in education." Characteristics of teachers and their behavior, as well as their instructional methods, were especially identified as the main factors associated with student outcomes (Brophy & Good, 1986); these, in turn, may influence other factors such as the classroom climate. However, in order for teaching and learning to take place in an effective environment, further favorable conditions and factors on a class-, school-, and context level will be necessary: for example, a clearly specified curriculum along with favorable policies for its implementation, and availability of corresponding instructional resources.

3.3.5.1 Teacher Background Factors

While *instructional effectiveness* (or *teaching effectiveness*) focuses on the teaching process as such, *teacher effectiveness* in the more narrow sense rather endeavors to identify specific background characteristics that are associated with teaching quality and, accordingly, with student achievement. Scheerens (2016, pp. 60–62), in his evaluation of the knowledge base regarding EER, distinguished the following important types of personal characteristics: personality traits, formal qualifications and experience, subject matter knowledge and knowledge about teaching and learning, and pedagogical content knowledge. Mayer, Mullens, and Moore (2000, iv) distinguished similar teacher characteristics as a part of school quality: teacher academic skills, teacher experience, teaching assignment (matching the formal qualification described by Scheerens), and professional development. These characteristics will be discussed in the following sections.

Personality traits and academic skills

While different teacher personality traits have been analyzed in the past, only limited empirical evidence regarding associations with student achievement could usually be found (Scheerens, 2016). However, Darling-Hammond (2000) reported some evidence that teachers' verbal ability might be related to student achievement; Mayer et al. (2000, p. iv) likewise concluded in their review that "students can learn more from teachers with strong academic skills."

Formal qualification and experience

Findings about associations between formal qualification and student achievement seem to be somewhat inconsistent. While Goldhaber and Brewer (2000) found differences in student achievement in mathematics in the U.S. between students who were taught by teachers who were fully certified and those who were not formally qualified, they could not confirm similar findings for other subjects. Darling-Hammond (2000, p. 8) reported that most of the studies in the U.S. about teacher certification in different subject fields "found higher ratings and greater student learning gains for teachers who have more formal preparation for teaching." TIMSS 2011 results showed that teachers who majored in education had the highest associations with mathematics and science achievement, while subject-specific orientations seemed to be less important, at least in the earlier grades (Martin et al., 2012; Mullis, Martin, Foy, & Arora, 2012). Analyses from Blömeke, Olsen, and Suhl (2016) related to teacher quality in TIMSS 2011 countries showed that the formal teacher education tended to be the strongest predictor of student performance across countries in their analyses and that it was most important for the Western Asian/ Arab region.

Results from meta-analyses conducted by Hanushek (1995) generally indicated a larger impact of teacher education in developing countries when compared to the U.S. Scheerens (2016, p. 61) explained this effect as being due to larger variations in teacher education in those countries – in contrast with more pronounced uniformity in teacher education in Western countries. In the Gulf area, large variations in terms of teacher background may be expected due to the high share of expatriate teachers from various different nations.

Regarding teacher experience, it can be expected that higher levels of experience due to a more advanced subject matter, and especially pedagogical knowledge gained by teaching practice and professional development, would lead to higher student learning gains. These effects have indeed been detected by some authors, for example Harris, Chapman, Muijs, Russ, and Stoll (2007) and Nye, Konstantopoulos, and Hedges (2004). The TIMSS 2011 data also showed that

on average, across countries, achievement in mathematics in both grade four and grade eight was highest for students who were taught by teachers with 20 years or more of experience (Mullis, Martin, Foy, & Arora, 2012, p. 292). Similar results were reported for science, albeit the effect was less pronounced in grade eight (Martin et al., 2012, p. 297). Other authors did not always find significant effects; according to Darling-Hammond (2000), effects rather seem to be curvilinear – with teachers with five to ten years of experience often having the strongest impact in relation to student outcomes.

Subject Matter Knowledge and Pedagogical Knowledge

Subject matter knowledge and *pedagogical knowledge* are variables that are frequently assessed to explain teacher effectiveness. Subject matter mastery is seen as a basic requirement for good teaching, and some authors, such as Monk (1994), indeed confirm positive correlations between the coursework taken by teachers and student achievement – to a certain extent. In general, however, findings are often neither as strong nor as consistent as could be hypothesized. Ashton and Crocker (1987), summarizing different studies, found positive relations – with small effect sizes, generally – in only 5 out of 14 studies. Darling-Hammond (2000), in her review of studies regarding the correlation between courses taken by teachers and student achievement, concluded that beyond a certain level which would satisfy the demand of the curriculum, the effect of additional courses becomes smaller. Findings related to pedagogical knowledge seem to be slightly more consistent and stronger. Ashton and Crocker (1987), for example, identified positive relationships between professional education and student performance in four out of seven studies, while Evertson, Hawley, and Zlotnik (1985) reported positive effects of teachers enrolled in formal education in 11 out of 13 studies. Similarly, Monk (1994) reported that "teacher education coursework" was positively associated with student outcomes, being – at times – even more influential than preparation of other subject matter. While both dimensions of teacher knowledge were historically regarded as independent domains, more recent research (e.g., Baumert et al., 2010; Hill, Ball, & Schilling, 2008; Tatto et al., 2012) covers both dimensions simultaneously, based on Sulman's idea of *pedagogical content knowledge* as the content knowledge that deals with the teaching process, including "the ways of representing and formulating the subject that makes it comprehensible to others" (Shulman, 1986, p. 9).

Professional Development

Professional development is usually offered by policymakers and educational reformers in order to improve teacher knowledge, skills, and practice with the intention of eventually improv-

ing student achievement. While short and event-like professional development programs in particular often fail to change teachers' attitudes and teaching practices, there is significant evidence that highly intensive, inquiry-based professional development might change teachers' attitudes towards reform, as well as their preparation and teaching practices (Supovitz, Mayer, & Kahle, 2000). In addition, it seems that professional development experiences need to be longer in length to trigger some effect. For example, Supovitz and Turner (2000) found in their analyses that only teachers given trainings of more than two weeks reported above-average changes in teaching practices and classroom culture. Yoon, Duncan, Lee, Scarloss, and Shapley (2007), who reviewed more than 1,300 studies relating teacher professional development to student achievement, found only few studies conducted with sufficient scientific rigor so as to be worth pursuing. All nine remaining studies conducted in primary education indicated a moderate association between professional development and student achievement in three different subject areas. They found positive significant effects especially for professional developments with a training duration of more than 14 hours: that is, encompassing much shorter durations than those reported by Supovitz and Turner. Blömeke et al. (2016), who did a comparative analysis concerning factors related to the quality of instruction, found that professional development activities were particularly important for the Asian and Arab countries.

In addition to the factors listed above, the *gender* factor merits brief discussion in this section. There are diverging beliefs about the importance of teacher gender for student learning. Some authors, such as Brophy (1985), believed that teacher gender has no impact on student achievement. Others reported a "math anxiety" of female teachers that could lead to lower achievement, especially among female students (Antecol, Eren, & Ozbeklik, 2012, p. 1). Still others assert that it might be helpful if the gender of students and their teachers match, for example to give a "male role model" to underachieving boys (Carrington & Skelton, 2003, p. 254). In general, the empirical evidence either does not seem to support significant differences (Ehrenberg, Goldhaber, & Brewer, 1995) or supports only slight differences in favor of female teachers, and predominantly for higher educational levels (Nixon & Robinson, Michael, D., 1999). Results from a randomized experiment conducted by Antecol et al. (2012) on the effects that female teachers of mathematics had on test scores of primary students indicated that, instead of the teacher's gender, rather the teacher's academic background seems to matter.

3.3.5.2 Effective Instruction

While for earlier phases, much of the research on teaching effectiveness was focused on teacher background characteristics and personality traits (with rather limited success), in later phases

the so-called *process-product studies* (Scheerens, 2016, p. 52) emerged. Here, more attention was paid to the relation between observed teacher behavior in the classroom and student achievement (Brophy & Good, 1986; Levine & Lezotte, 1990; Mortimore et al., 1988; Scheerens & Bosker, 1997). Including research conducted in a variety of contexts and countries, Chapman et al. summarized the rather generic results from this period as follows:

> Effective teachers emphasise academic instruction as their main classroom goal, have an academic orientation, create a business-like, task-oriented environment, and spend classroom time on academic activities rather than on socializing, free time, etc. (Chapman et al., 2015, pp. 101–102).

Recently, more reviews and research projects have put a stronger focus on teaching strategies that were developed from constructivist learning theories – such as the teaching of higher-order thinking skills or self-regulated learning – with "a strong re-statement of the fact that teaching is about facilitating learning, by considering learning activities and student engagement" (Scheerens, 2016, p. 57).

The following section gives a short summary of key quality factors that play an important role in instructional effectiveness, mainly based on the most recent reviews of key factors in instructional effectiveness research by Muijs et al. (2014) and Scheerens (2016).

Classroom Management

Teacher effectiveness research has repeatedly found that the way a classroom is managed is an important precondition for effective instruction. By helping to limit misbehavior and distraction, which influence the attention students pay to the lesson content, classroom management has consequences on the *time on task* (Brophy & Good, 1986; Doyle, 1985; Muijs & Reynolds, 2000). Research has shown that teachers have to establish and enforce clear rules and procedures for student behavior, especially at the beginning and the end of the lesson, as well as during transition periods; in addition, rules and procedures need to be explicitly and clearly communicated to the students. According to Brophy and Good (1986) and Doyle (1985), effective teachers are therefore able to manage classrooms in such a way that activities run smoothly, transition periods are short, and not much time is spent on organization or dealing with misbehavior.

Clear and structured teaching

Research has shown that learning gains are usually higher in classes where most of the lesson time is led by teachers – as opposed to students working on their own (Chapman et al., 2015, p. 102). This does not imply advocacy for rote memorization or drills. Instead, the teacher is expected to actively transfer the content to students in a clear, structured way, rather than relying on textbooks or similar material to do so. This kind of teaching is also known as *direct teaching*. Doyle (1985) considers the following features of direct instruction important:

- Clearly formulated teaching goals
- The material to be followed is split into smaller tasks and taught in an appropriate order
- Clear explanations about what students are supposed to learn
- Regular questions to monitor students' progress
- Sufficient time for students to practice
- Working with a skill until it is overlearned by the students
- Regular reviews and holding students accountable for their work

Furthermore, it is also important that the teacher outlines the content, summarizes important subparts, and reviews key findings at the end. New knowledge should be linked to prior knowledge, while the main ideas of new material also need to be linked to one another (Brophy & Good, 1986; Chapman et al., 2015; Muijs et al., 2014). Moreover, the main concepts should especially be presented with a certain degree of redundancy and clarity (Scheerens & Bosker, 1997).

It seems that for a deeper understanding of the presented material, good questioning strategies – that attempt to involve students in the class discussion and check their understanding – are important. It was found that most questions should call for explanations (process type) instead of a single response (product type). The cognitive level should be mixed and adapted to the skills that need to be mastered, and teachers should provide swift and substantive feedback to students on the accuracy of their answers (Brophy & Good, 1986; Chapman et al., 2015; Mortimore et al., 1988; Muijs et al., 2014; Reynolds et al., 2014).

Strong empirical support for structured whole-class teaching and associated student achievement was provided by several intervention programs, such as the Missouri Mathematics Effectiveness Project (Good & Grouws, 1979), or classroom observation studies like the Junior School Project (Mortimore et al., 1988) or the Gatsby Mathematics Enhancement Programme (Muijs & Reynolds, 2000). It seems that direct instruction methods are especially helpful for student groups with low socio-economic background and low attainment (Chapman et al., 2015).

The main reasons behind the strong effects associated with a structured direct teaching approach seem to be that teachers have more contact with each individual student when compared to individual settings, and that students' *time on task* is higher. In addition, the teacher can more easily detect distraction due to a lack of understanding or boredom, and change and vary activities accordingly (Chapman et al., 2015).

The strong empirical evidence for direct instruction does not imply that group work or seatwork should be regarded as ineffective. The advantages of group work lie mainly in the cooperative aspects, and thus on the contributions it can make to the development of students' social skills. As the knowledge base of a group is most likely larger than the knowledge of an individual student, group work also allows for the solving of more complex tasks. Additionally, the combination of group work and individual practice with direct instruction methods may be an important feature of an effective lesson – as the former allows student to review and practice what they have learned during the lesson (Creemers & Kyriakides, 2008). However, to be effective, tasks for group work or individual seatwork need to be clearly explained to the students, and the teacher must monitor and help the students during those periods (Muijs & Reynolds, 2000).

Activation and self-regulated learning

While there is strong empirical evidence relating to direct instruction methods in the field of instructional effectiveness, it should be noted that related studies usually focused on a limited number of core subjects, and tested students' basic skills. Particularly in light of what is known as the *cognitive revolution* (Scheerens, 2016), recent educational effectiveness research takes a broader view on education, with a more student-centered focus on self-regulated and life-long learning. Constructivist learning and teaching approaches emphasize the active role of students in constructing knowledge. The main underlying concept of constructivism is the assumption that there is no strict separation between subject and object; consequently, the perception of the reality is always seen as influenced by presumptions of the observer (Gruehn, 2000, p. 53). This

implies that each form of knowledge needs to be constructed by the learner through the activation of his cognitive structures. In consequence, learning strategies and the reflection on those strategies are seen as important as mastering the content itself. For teaching, this implies that the learning environment should be engaging for students and allow for the exploration of real-life content – or at least simulated environments. More modern teaching strategies, influenced by constructivist ideas, also try to offer students several different opportunities for active learning, comprising varying facets such as cooperative learning, discovery learning, peer-tutoring, and student experiments, embedded in a challenging learning environment (Cobb et al., 1991; Scheerens, 2016, p. 44). Teaching strategies that stimulate students to be cognitively active (also called *cognitive activation*) are tasks that require higher-order thinking skills. These are intended to enable students to really understand what was taught to them, and to use mistakes as future learning opportunities (Klieme & Rakoczy, 2003, p. 335). Under a constructivist paradigm, therefore, teachers are more in a role of facilitator or coach, supporting students to implement their own strategies which can help them solve various kinds of problems, and as a result, help them to organize their own learning (Creemers & Kyriakides, 2008, p. 109). The strategies necessary for students to develop such kinds of learning are often summarized under the concept of *self-regulated learning*, which, according to Pintrich (2005, p. 453), can be defined as "an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment." *Cognition* here refers to the "cognitive information-processing strategies that are applied to task performance, for example attention, rehearsal, elaboration" (Chapman et al., 2015, p. 109). Another noteworthy term in this context is *meta-cognition,* which refers to the instrument that controls the elements in the definition above, and is also referred to as *thinking about thinking*, or *higher-order thinking* (Chapman et al., 2015, p. 109). *Metacognition* in this sense "forms the basis of the process of self-regulated learning" (de Boer, Donker-Bergstra, & Kostons, Danny D. N. M., 2012, p. 8). Veenman, Van Hout-Wolters, Bernadette H. A. M., and Afflerbach (2006, p. 9), reviewing the research in this area, concluded that three fundamental principles are needed for a successful metacognitive instruction:

- Embedding metacognitive instruction with the subject matter taught to allow connections between both dimensions
- Engaging students in the application of meta-cognitive principles by developing an understanding of their usefulness
- Assuring long-term training of the metacognitive skills with regular reviews

The body of empirical evidence about the association of cognitive activation strategies with student learning gains is growing. Klieme and Rakoczy (2003, p. 336) found in an analysis of the TIMSS-Video Study (Stigler, Gallimore, & Hiebert, 2000), wherein cognitive activation was recognized as one out of three major dimensions summarized from the observer's ratings, a correlation with student outcomes based on the TIMSS assessment. Hattie (2009), who conducted perhaps the most comprehensive review in the field of educational effectiveness by synthesizing the findings of more than 800 meta-analyses, confirmed the classical findings, but furthermore stressed that the emerging relations between constructivist approaches and achievement with problem-solving skills and meta-cognitive strategies are important.

However, it should be noted that the quantitative and self-assessed questionnaire data available from the TIMSS assessment is only for limited use in assessing constructs related to cognitive activation. In order to asses these constructs more comprehensively, qualitative approaches would be more appropriate.

High Teacher Expectations

According to Teddlie and Reynolds (2000), high expectations of teachers for students can be seen as one of the most important factors in EER; this area has been a focus of research for several decades. This factor emerged in virtually all larger empirical studies and reviews in Great Britain (eg., Mortimore et al., 1988; Rutter, Maughan, Mortimore, Ouston, & Smith, 1979), The Netherlands (Scheerens, 1992, 2000), and the United States (Levine & Lezotte, 1990; Sammons et al., 1995; Teddlie, Kirby, & Stringfield, 1989). By the 1960s, Rosenthal and Jacobson had already described an effect in which an a priori positive expectation of a student by a teacher later might be confirmed via a "self-fulfilling prophecy" – also known as the *Pygmalion effect* (Rosenthal & Jacobson, 1968). It was found that teacher expectations may affect students in a variety of ways, such as communicating their expectations to students; paying more attention to, and spending more time with, high-expectancy students; criticizing, and giving lower-level academic tasks to low-expectancy students; and so on (Muijs et al., 2014). It is therefore important that teachers be made aware of the importance of showing a positive attitude and high expectations also for disadvantaged or less capable student groups, and of the importance of relying on objective achievement measures, thus continuously questioning and mitigating stereotyping and snap judgments. Of course, high expectations of teachers alone will not be sufficient; they also need to show corresponding attitudes and to clearly communicate these expectations to the students. Sammons et al. (1995, p. 39) stated that "…even if teachers do not believe success is possible, conveying conviction that achievement can be raised can

have a powerful effect" and further that "reinforcing this success through praise (…) is a key opportunity for communicating high expectations."

Assessment and Feedback Strategies

The assessment and monitoring of student progress also are important factors in EER and already belong to the effectiveness-enhancing correlates identified in Edmonds' (1979) so-called *five factor model* (see section 3.3.1). Assessments should be *formative* – meaning that the results are used to influence decisions about subsequent steps in instruction, instead of *summative,* wherein the intention is rather that of a final judgement. Data from formative assessments should enable teachers to identify their students' needs, but also allow for the evaluation of the impact and quality of their own teaching practice. Additionally, positive effects on student motivation (as it shows that students are interested in their progress) are seen, and assessments can be used to analyze students' progress (Creemers & Kyriakides, 2008; Teddlie & Reynolds, 2000). There is empirical evidence that more frequent formative testing may lead to learning gains. It seems, however, that positive effects decrease beyond about one to two tests per week (Bangert-Drowns, Kulik, & Kulik, 1991; Black & Wiliam, 1998). However, the effect of the assessments themselves seem strongly related with the feedback strategies and the use of the test information. It could be shown that assessment strategies are especially helpful if the feedback comes promptly, and contains information in some way about the correct response (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991). In addition, effective performance feedback should not be judgmental, but rather identify learning gaps, ideally helping to identify means or techniques to bridge these gaps (Scheerens, 2016). Procedures of formative assessment, feedback, and corrective measures also play major roles in the *mastery learning* framework (Bloom, 1968), an important construct of a structured teaching practice similar to *direct teaching* described in section 3.3.5.2.

Adaptive Teaching

There are strong indications for positive effects relating to both traditional teaching approaches and more modern approaches, influenced by the constructivist paradigm; it seems that an appropriate mix of the different methods is important. Therefore, it should be considered that each of the teaching practices and strategies are not effective on their own, but rather become effective when integrated as a product of varied strategies employed by the teacher to keep students engaged (Muijs & Reynolds, 2000). Consequently, choice of material and teaching strategies need to be adjusted to the characteristics of the students, for example, based on their ability levels or motivational profiles (Scheerens, 2016, p. 21). Instead of any single teacher behavior

being found to be strongly related to achievement, rather several smaller correlations were found, thus indicating that "effective teaching is not being able to do a small number of 'big' things right but is rather doing a large number of 'little' things well" (Reynolds et al., 2014, p. 212).

As evidenced above, effective teaching therefore not only depends on teachers' behavior, but also on their background characteristics, such as pedagogical content knowledge; their beliefs, as well as their expectations about their students, are also important.

Differential Effectiveness

Traditionally, teacher effectiveness research focused on rather generic teacher factors related to cognitive student outcomes, often as measured by standardized achievement tests. More recently, researchers have begun to investigate differences in teacher effectiveness, especially in regarding certain core areas of the curriculum, the student background composition (SES, ability, and personal characteristics), and different teacher roles. Based on these investigations, researchers have also claimed to develop more appropriate differentiated models of teacher effectiveness (for example, Campbell, Kyriakides, Muijs, & Robinson, 2003). However, empirical findings in differential effectiveness research seem to produce rather heterogeneous results, with main empirical evidence found in the areas of curriculum and student background composition. Muijs, Campbell, Kyriakides, and Robinson (2005, p. 65), for example, indicated some differential effectiveness between subjects such as English and Mathematics, although (as they concluded) they were built "upon strong generic similarities." They reported findings from an effectiveness project in England indicating that subject knowledge mattered less for numeracy than it did for literacy, and that differentiating tasks by ability seemed to be more important for literacy. The authors also reported findings from a comparison of different content domains (number, calculation, and measures, shapes, and space), based on a reanalysis of data stemming from a study of teacher effectiveness in mathematics conducted by Muijs and Reynolds (2003). The findings here showed: that varied teaching was less related to achievement in the number domain when compared to other domains; that a high pace with immediate feedback was most important for calculation; and that clear explanations, and asking students to explain their answers, were most strongly related to achievement gains in measures, shape, and space. Yet stronger research evidence is available in terms of SES and ability level. In general, it seems that low-SES students are more strongly affected by instructional quality than high-SES students. They need more control and a more structured approach, more feedback, and the curriculum material needs to be presented in smaller packages (Brophy & Good, 1986). However, it

seems that differences in effective teaching practices in regard to student background "are often matters of degree (e.g., extent of structure and praise) rather than pointing to a complete disjuncture between teaching methods or curricula" (Muijs et al., 2005, p. 65).

3.3.5.3 Classroom Climate

The classroom climate can be defined "as the general atmosphere in the classroom" (Scheerens, 2016, p. 43), and is developed via a dynamic relationship between teachers and students within their learning environments during the school year (Fraser, 1994).

Originally, teacher effectiveness research often focused on investigations of management techniques that are important for creating a good classroom climate, such as business-like and supportive style of teacher-student interactions; achievement orientation; high teacher expectations; and clear disciplinary rules (Campbell et al., 2003; Chapman et al., 2015; Muijs et al., 2014; Scheerens, 2016). These styles and techniques were discussed in the previous sections. Pointing in a similar direction as *classroom climate,* but sometimes defined more broadly, the term *classroom environment* – regarded as relating to the behavior of all the different stakeholders influencing classroom instruction – is used (Chapman et al., 2015, p. 103). The concept of classroom environment takes not only teacher-student interactions and student-student interactions, but also students' treatment by the teacher, competition between students, and classroom disorder into account. While the first two elements are seen as important components in the measurement of classroom climate in the more narrow sense, the remaining elements refer to approaches of teachers to create an efficient and supportive learning environment, which has proved to be an important additional factor in teacher effectiveness research (Walberg, 1986).

The psychological learning environment formed by a social group was also considered to be an important factor influencing student outcomes and attitudes, according to Walberg's theory of educational productivity (Walberg, 1971). Secondary analyses by Haertel, Walberg, and Haertel (1981), based on studies conducted in four different countries, supported Walbergs' model by finding positive associations between student perceptions of their class environment and learning outcomes in eight different subject areas in the sub-dimensions of cohesiveness, satisfaction, task difficulty, formality, goal direction, democracy, and material environment. In a study based on 1,955 students participating in the U.S. assessment of science, Walberg, Fraser, and Welch (1986) also found that, even when predictor variables were controlled for, the class environment (among 9 out of 11 other predictors related to his theory of educational

productivity) was significantly related to student outcomes. There is some indication that classroom climate also might influence student achievement indirectly, mediated by instructional quality and instructional time (Reynolds & Walberg, 1991).

Classroom- (and school-) climate will not only depend on the behavior and beliefs of the teacher. Of course, also the behavior of the students contributes to the climate; accordingly, student composition is important. Willms (1992, p. 41), for example, states that schools with high-ability or high-SES students have associated contextual advantages: "On average they are more likely to have greater support from parents, fewer disciplinary problems, and thus a climate conducive for learning." Supporting empirical evidence was found by Opdenakker, van Damme, Fraine, van Landeghem, and Onghena (2002), who found in their analysis of Flemish data that learning climate correlated with group composition, albeit sometimes showing additional effects on achievement even after controlling for the composition effects.

3.3.6 School-level factors

In earlier periods of EER, much attention was given to factors on a school level that might influence student achievement; with the development of multilevel modeling techniques, however, school and class level could be separated, with many studies indicating larger effects in relation to student outcomes on class level. Nevertheless, school-level effects are important, as they influence student learning by "establishing high expectations for educational experiences and by setting the context within which quality interactions can occur" (Mayer et al., 2000, p. 37). Thus, school-level factors can be seen as important preconditions for classroom learning – even if they often affect students more indirectly, and the effects are mostly small. The following sections describe some main factors that are theoretically and empirically associated with effective schools.

3.3.6.1 Professional Leadership

In EER, leadership has always been regarded as an important effectiveness-enhancing factor essential for student success (Brookover, 1979; Edmonds, 1979; Mortimore et al., 1988; Purkey & Smith, 1983; Rutter et al., 1979; Sammons et al., 1995).

While the importance of the principals' role (or that of his or her team) as such is not doubted, it seems that the question of which leadership styles and personal characteristics are best associated with effective schools depends on patterns of school organization, the development level

of the school, and other contextual factors. Bossert, Dwyer, Rowan, and Lee (1982, p. 38) consequently conclude that "No single style of management seems appropriate for all schools. For example, reviews of the successful schools literature intimate that principals must find the style and structures most suited to their own local situation."

However, certain characteristics of successful leadership are reported more consistently in the literature. According to the findings, leadership should be *firm and purposeful,* implying that the principal understands the school's needs, is actively involved, and is the key agent in initiating change processes. Leadership also should be *participative*, meaning that successful principals share leadership with other members of the senior team or with teachers and involve their staff more generally in decision-making. The principal is also supposed to be the *leading professional,* implying his or her involvement in and knowledge about what goes on in the classroom, including the curriculum, teaching strategies and the monitoring of progress (Mortimore et al., 1988; Rutter et al., 1979; Sammons et al., 1995). As with most of the school factors, it is assumed here that the impact of principals on student outcomes will work rather indirectly "by influencing school and staff culture, attitudes and behavior which in turn, affect classroom practices and the quality of teaching and learning" (Sammons et al., 1995, p. 22). The importance and shaping of effective leadership styles also seem to depend on the cultural context. For example, in their study about school effectiveness research in nine different countries, Reynolds, Teddlie, Stringfield, and Creemers (2002, p. 255) found that the effectiveness of a school depended more on the leadership of the principal in English-speaking countries, whereas this was less the case in the non-English speaking societies (Hong-Kong, Taiwan, The Netherlands, and Norway) in their study. According to Reynolds et al., the latter educational systems were so ordered and well-engineered that individual leadership characteristics mattered less than other system variables.

3.3.6.2 Productive School Climate and Culture

There is quite a body of research investigating *school climate:* a description for the general atmosphere of the school. School climate is often regarded as a factor that is partly malleable by the actions of the school leader, but also emerges from interactions between staff and students, and the students themselves (Scheerens, 2016, p. 89). The concept of *school culture* is similarly defined, but is based more on norms and values. Maslowski (2001, pp. 8–9) defines school culture as "the basic assumptions, norms and values, and cultural artefacts that are shared by school members, and which influence their functioning at school." Factors reflecting school

climate and culture have emerged in virtually every review or study about school effectiveness research. Important factors relating to school climate and culture are detailed below.

Orderly atmosphere and a positive disciplinary climate

Many authors regard an orderly environment as an important precondition for effective learning (Levine & Lezotte, 1990; Mortimore et al., 1988; Sammons et al., 1995). It is easy to understand that teachers are not able to maintain student attention and engagement without an orderly environment, and that lesson time most likely could not be efficiently used. TIMSS 2011 data (Mullis, Martin, Foy, & Arora, 2012) supports this notion. Over all participating countries, schools where principals reported "Moderate Problems" with school discipline and safety had, on average, 45 points (close to half a standard deviation) lower mathematics achievement in grade four than schools where discipline and safety were creating "Hardly any Problems". Martin and Mullis (2013) conducted a school effectiveness analysis based on 32 countries participating in TIMSS and PIRLS in mathematics, science, and reading. They found that the factor *schools are save and orderly* was positively related with achievement in at least one subject in 15 countries, even after controlling for the home background. For the participating GCC countries, significant associations emerged in Oman, Qatar, and in the United Arab Emirates for at least one subject, but not in Saudi Arabia.

Shared vision, staff cohesion, and collaboration

Schools have frequently proved to be more effective when staff are committed to a school-wide mission focused on academic improvement, and when a consensus is put into practice through consistent and collaborative ways of working and decision-making (Levine & Lezotte, 1990; Sammons et al., 1995, p. 23). Rutter et al. (Rutter et al., 1979, p. 192) pointed out that the atmosphere of any school "will be greatly influenced by the degree to which it functions as a coherent whole, with agreed ways of doing things…". Several studies, reviews, and school improvement programs have given empirical evidence that a consensus on values and goals, grounded in common and agreed-upon approaches to school life, is related to higher academic outcomes (Levine & Lezotte, 1990; Purkey & Smith, 1983; Rutter et al., 1979). Mortimore et al. (1988, p. 224), for example, found positive associations with school learning in schools where teachers followed consistent approaches in using school curriculum guidelines; Rutter et al. (1979, p. 121) reported that students are more likely to maintain guidelines of behavior if they understand that standards of discipline are based on "general expectations set by the school."

Collegiality and collaboration between staff can be seen as important conditions for achieving the consensus and the implementation of common approaches described above (Sammons et al., 1995). For example, regular meetings of teachers may be helpful in improving cohesion and collaboration among teachers. Having a participative approach, including staff members in decision-making processes, and creating a sense of "ownership" is also regarded as important (Mortimore et al., 1988). There also needs to be some constancy in the staff composition over time: as Purkey and Smith (1983, p. 443) state, "Frequent transfers are destructive and likely to retard, if not prevent, the growth of a coherent and ongoing school personality." Issues with staff constancy may be an important concern in the Gulf Area, potentially hindering the creation of a productive school climate and finally having some consequences on school learning. In the GCC countries, teacher attrition and a teacher turnover rate is generally high. Reasons given in the literature include that the teaching profession is seen as a low-status profession (Ridge, 2014, p. 135); consequently, national male teachers in particular seek out other employment activities or promotions as soon as possible. Conversely, the situation for expatriate teachers is difficult, as they usually earn lower wages and do not have the same rights as national teachers – a fact which, mediated by different personal, economic and sociocultural factors, again leads to high attrition rates (Demirjian, 2015; Ridge, 2014).

Teddlie & Reynolds provide a concise summary of the important factors needed for establishing and maintaining a good school culture:

> The generation of a learning community amongst staff in which all members share good practice, act as critical friends and engage in a process of mutual education and re-education is clearly essential in the continuation of a positive school culture over time, as well as in its creation (Teddlie & Reynolds, 2000, p. 148).

3.3.6.3 Concentration on teaching and learning

A focus on the importance of academic goals and processes and high academic emphasis has repeatedly been shown to exhibit correlations with school effectiveness. While many factors in this area may play an even more important role on classroom-level, the school – usually via policies, regulations, and priorities – often sets the standards and examples for classroom practices.

Maximization of learning time, opportunity, and quality

A number of studies have shown positive relations between the maximization of learning time and student outcome measure. The number of instructional days per year, the length of a typical school day, and the time allocated to specific subjects are all linked to the concepts of *time on tasks* and *opportunity to learn*, considering that maximizing these variables will offer students more instructional time. However, not only is the amount of time available of importance, but also the quality. As stated by Carroll (1989, p. 27): "time as such is not what counts, but what happens during that time." While the curriculum might be defined at a contextual level, schools may often set conditions for the implementation of the curriculum by setting preconditions with respect to the quality of instruction and the opportunity to learn. Creemers (1994) gave the following examples: setting rules about textbooks, curricular material, grouping procedures, and teacher behavior; implementation of an evaluation policy (for the quality of instruction); and rules about the development of a school working plan, and how to follow the curriculum (for the opportunity to learn). A broader discussion on the concepts of time on task and opportunity to learn can be found in sections 3.3.2 and 3.3.3, respectively.

Academic emphasis

An achievement-oriented school focusing on the mastery of academic content can contribute to student learning, as demonstrated in school effectiveness research (Levine & Lezotte, 1990; Scheerens, 1992). However, Levine and Lezotte (1990, p. 14) cautioned that an emphasis on mastering central learning skills, in the absence of other effectiveness-enhancing factors, might not be successful – and rather should be "viewed as a building block antecedent to rather than a 'guarantee' of effectiveness."

Martin and Mullis (2013) found in their school effectiveness analyses of TIMSS and PIRLS countries that the factor *schools support academic success* was positively related with achievement in at least one subject in 10 out of 32 countries, even after controlling for the home background. In the Gulf countries under consideration, *schools' support for academic success* was a significant factor in at least one subject for Oman, Saudi Arabia, and the United Arab Emirates, but not for Qatar.

3.3.6.4 Parental involvement

Some scholars advocate for good home-school relations and for encouraging a stronger parental involvement of parents in children's learning and school activities. However, present research

doesn't point toward a common agreement on the level and type of involvement that would work best; accordingly, empirical findings are mixed. Purkey and Smith (1983), in their comprehensive review of school effectiveness literature, found only a few studies where parental involvement was related to academic outcomes. Mortimore et al. (1988, p. 226), on the other hand, reported positive benefits in schools where parents helped in the classroom and with school trips, and where regular progress meetings were provided. It can be assumed that the effect of parental involvement depends on many different factors, such as the age of the child, the management and monitoring activities of the school in regard to the home-school relations, or the student composition in terms of socio-economic background of the families.

3.3.6.5 School resources

Resources at the school level would involve buildings, libraries, heating/cooling and lighting, and general instructional material, but also teacher-related resources, such as teacher salaries and the student-teacher ratio. While resource characteristics have often been studied as malleable input variables, empirical evidence is rather contradictory. There is some indication that favorable physical characteristics – such as the amount of light, fresh air, and an acceptable level of noise – are positively linked to educational outcomes (Chan, 1979; Scheerens, 1992); Greenwald, Hedges, and Laine (1996) also concluded, based on their reanalysis of data from several education production functions, that certain school resources – such as school and class size, but even more the per-student expenditure – are positively related to achievement. In general, it seems plausible that a basic level of physical resources might be needed to successfully implement instruction, but that beyond certain thresholds better equipment, more space, and other physical resources would not contribute much more in regard to effective instruction. This could help explain the fact that school input variables indicate more significant positive relation to learning outcomes in developing countries. For example, Scheerens (2000), in his review of school effectiveness studies, reported that the availability of textbooks showed significant association with achievement in 19 out of 26 studies, and that the availability of a school library was significantly correlated with outcomes in 16 out of 18 studies. However, concerning the availability of certain facilities, such as libraries, gymnasia or computer laboratories, researchers repeatedly pointed out the question of *how* the facilities are used is more crucial than their mere availability (Scheerens, 1992, p. 92). It can be concluded that physical resources on their own will probably not constitute an important effectiveness-enhancing factor, but that to a certain extent, they might be a precondition for effective instruction, with perhaps greater importance in developing countries.

3.3.6.6 Other factors

Certain factors already described at the classroom level are more effective if they are an integrated part of the school culture, and thus consistently expressed or implemented on all levels. In this sense, for example, *high expectations* would need to be part of a general culture in a school; consequently, high expectations for students would likely be associated with a staff group "who have themselves high expectations of what is possible from them to achieve from the principal or headteacher" (Teddlie & Reynolds, 2000, p. 149). Similarly, the monitoring of student progress should be organized centrally, and accompanied by the implementation of a monitoring and evaluation system at the school level to better manage school institutions – which is also seen as a characteristic of an effective school (Teddlie & Reynolds, 2000, p. 150). Moreover, it is important that staff development becomes an integral part of school activities, and that there is a close synchronization of the school's mission and priorities with the staff development activities (Teddlie & Reynolds, 2000, p. 150).

3.3.7 Context-level factors

Factors beyond the school level are expected to influence student outcomes indirectly, via school and classroom level. However, certain factors, such as a centralized testing system, might also affect students directly. Although educational policies and guidelines (for example, concerning curricula, grouping procedures, teacher education or testing systems) are expected to define important conditions for the dependent educational levels, empirical evidence in this regard is still rather scarce. The reason might be that political and structural approaches in different countries are quite various and related studies can easily become quite complex.

International comparative assessments often show large achievement gaps between developing and developed countries, but also between different developed countries and even between different regions or school tracks within the same country. However, upon review of the literature regarding explanations of academic differences in achievement between countries, and especially those related to international comparative assessments, it seems that the conditions for learning and teaching on the level of the educational system – and thus the structure of the school system – do not contribute much in explaining those differences (Baumert, Bos, & Watermann, 2000; Fend, 2004; Lankes, Bos, Mohr, Plaßmeier, & Schwippert, 2003; Schümer, 2001). The high performance of East Asian educational systems in particular in international large-scale assessments such as TIMSS and PISA, as well as the performance of countries like

Finland, boosted the interest of researchers and policy-makers in detecting factors of educational success in those countries. There is consequently a growing body of EER in these countries; Chapman et al. (2015, p. 280), however, have viewed country-specific and regional studies (e.g., in East Asia) as "limited in their utility by the variation in their methodology, sampling, data collection methods, and analytical techniques."

Maybe the issue at hand is not so much about differences in organizational structures in and of themselves, but rather about their degree of success in implementing the effectiveness-enhancing factors described in the previous sections most consistently and comprehensively in an educational system. Chapman et al. (2015) identified a number of teaching behaviors, widely practiced in the East Asian region, that are also labelled as *effective teaching* in the Western literature. These behaviors include: a high level of academic engagement; whole-class interaction, and more time on task; teaching with variation; a brisk teaching pace; appropriate teacher questioning; more opportunity to learn; and regular homework, with timely feedback.

They concluded in their review of the effectiveness literature from the East Asian region:

> In spite of working in centralized systems amidst a trend of decentralization in most nations of the region, school leaders are able to put teaching at the heart of their school lives and provide strong support to teachers and their professional development, most of which is based within schools to suit teachers' convenience. Teachers in East Asia have adopted a set of effective methods and are able to apply them proficiently, which partly explains the sustained phenomenal success of East Asian learners in international assessments (Chapman et al., 2015, pp. 279–280).

Clearly, these countries manage to implement certain conditions of effective education listed as important by Creemers (1994) better than many Western countries:

- Educational policy that focuses on the effectiveness through evaluation procedures and through promotion of effective grouping procedures, curricula, and teacher behavior
- Availability of an indicator/ evaluation system
- A training and support system promoting effective schools and effective instruction
- Implementation of guidelines for the time schedules of schools and their supervision

- Guidelines and rules for the development of the national curriculum and school working plans

3.4 School Effects

3.4.1 Existence of school effects

Spurred by the seminal reports of Coleman, much research concerning school effects has been undertaken, and their existence for different cognitive and non-cognitive outcome criteria under different schooling conditions can be taken for granted (Chapman et al., 2015; Sammons, 1999; Teddlie & Reynolds, 2000). Having reviewed the body of effectiveness research literature, Sammons (1999, p. 76) found that "Evidence for the existence of school effects has been found across all phases of schooling and for a variety of usually academic educational outcomes." However, psychometric properties of school effects, such as their magnitude or validity, do not depend only on the outcome criteria under consideration and the phase of schooling, but also on the conceptualization of a school effect measure, and thus from the corresponding measurement decisions taken. Teddlie and Reynolds (2000, pp. 65–66) and more recently Chapman et al. (2015, pp. 30–31) describe the main conceptualizations as follows:

- Absolute effect of schooling
  This effect describes the overall effect of attending a school versus "control groups" not attending a school. These effects only can be studied in systems where education is not compulsory or if students have dropped out or for other reasons had no schooling over a period of time.

All following effects are *relative*, meaning that they compare different kinds of schools or students among each other.

- School effects in the form of a *gross mean achievement score*
  This is also described as the *unadjusted average achievement* (Teddlie & Reynolds, 2000, p. 66) of all students in a school. The obtained raw scores are not adjusted for any kind of student intake and thus allow for unfair comparisons of schools showing very different background conditions.

The subsequent definitions, which were summarized from Chapman et al. (2015, pp. 30–32) and from Teddlie and Reynolds (2000, pp. 66–69), are applied in more recent effectiveness

research. They are based on so-called *value added* models. Those approaches try to disentangle student background factors, student aptitude, and non-educational context factors from the educational factors under consideration.

- The mean progress that students make over a given period of time compared with prior attainment
  These effects can be calculated from a predicted score based on regression analysis that control for prior attainment and student background characteristics.
- The impact of schooling on the *average* achievement of all students in a school, adjusted for prior attainment and socio-economic status.
- Type A and Type B effects as defined by Raudenbush and Willms (1995, pp. 309–310):
  Type A is defined as "the difference between a child's actual performance and the performance that would have been expected if that child had attended a 'typical school'". Hence this effect is measuring the total impact on a student of attending a certain school. Effects here include next to factors within the control of the school also external factors such as the influence of the social and economic context of the community surrounding the school, the composition of the student body, etc.
  Type B effects are a subset of Type A effects and include only those influences of schooling that are directly related to factors in the control of the school. Type B effects describe: "the difference between a child's performance in a particular school and the performance that would have been expected if that child had attended a school with identical context but with practice of 'average' effectiveness". School effectiveness research according to Raudenbush and Willms (1995) therefore should focus specifically on Type B effects.
- Measurement of the impact of different schools on student performance over time
- Relative size of school effect, measured by the intra-school correlation in multilevel models
  Here the variance is partitioned into different levels, for example into a student and a class level. A contextualized *value-added* model (see section 3.2 for more information on this concept) here typically controls for prior attainment, student background and measures of intake composition. This model then can be compared in terms of explained variance to a model containing additionally to the input and context measures

also the school instructional factors. This is also the approach that will be followed in this thesis.

- Differential effects
  Here the effect of individual school instructional factors is analyzed for different student groups (gender, SES, etc.) by fitting random slopes in multilevel models.
- Individual school effects based on residual estimates for each individual school in a multilevel model
  These effects of residual estimates in multilevel models controlling for prior attainment and background characteristics are used to identify whether students in some schools make significantly more or less progress than predicted.

3.4.2 Magnitude of school effects

Findings from studies related to the magnitude of school effects are rather heterogeneous and depend on many factors, such as the operationalization of the constructs, the statistical methodology used to analyze the effects, the level of analysis (for example student or school), the outcome criterion used, sample specifications, etc. Additionally, research indicates that school effects differ in different contexts and for different groups of students.

One important aspect related to the magnitude of a school effect is its operationalization, which has an effect on the *construct validity* (meaning the appropriateness of inferences made on the basis of the measurement) of the effect. The validity might be compromised if the construct is misspecified. *Internal validity* (concerning the question of whether a causal conclusion based on a study is supported) issues can occur if variables associated with the school processes under consideration are not controlled for appropriately. In addition to factors such as SES status and student aptitude that have a strong impact on student learning gains, or factors such as *teaching to the test*, random fluctuation of the student body and so forth may be a threat, especially for cross-sectional studies which are administered at only one point in time – thus confounding effects found from within the schooling system. Both SES, and to a certain extent a proxy of students' aptitudes – which are assumed to be the major factors in this regard – are (to the best knowledge of the author) taken into account for the current study (see section 8.4); the TIMSS data does not provide information about other confounding factors like the fluctuation of the student body. Threats to the *statistical conclusion validity* (defined as threats to the drawing of valid conclusions about covariation between variables based in statistical evidence) also need to be considered. Incorrect conclusions about insignificant results (also called a Type II error)

could emerge in case of small sample sizes. On the other hand, there is a chance that a certain effect could be mistakenly regarded as significant when it is not (Type 1 error); this especially might happen if researchers are "fishing" for significant associations between variables, demonstrating the necessity of basing statistical analyses on a strong theoretical framework. While sample sizes of participating schools and students should be sufficient to keep the Type II error low (see Table 7-1), the current study tried to reduce the risk of Type 1 errors by rooting the study in well-established constructs of EER.

Important for a correct specification of school effects are an appropriate selection of all relevant input variables, and that the educational input factors are clearly disentangled from external context conditions. The choice of outcome measures is also important. For example, tests measuring generalized academic abilities will most likely be less sensitive than tests based on what has been actually taught in class. Additionally, the size of a school effect will depend on the cognitive domain chosen as outcome criterion, and again might be different if non-cognitive indicators such as classroom behavior, absenteeism, or attitudes towards learning are used. Hierarchically, the level on which a certain indicator is included might affect the results as well. Thus, a correct model specification is essential. Unfortunately, even recent integrated comprehensive school effectiveness models differ quite a lot in the selection and specification of indicators and in the assumed interactions between the different parts of the model, which hampers comparison of their results.

It is generally accepted that both prior achievement of students and family socio-economic characteristics should be included in models of educational effectiveness (Sammons, Mortimore, & Thomas, 1996). Researchers such as Willms regard prior performance as the more important factor. Willms (1992, p. 58) concluded from his work in Scotland on analyses related to estimation of school effects that "If the analysis does not include measures of prior performance, the estimate of effects will be probably biased. Measures of family background add marginally to the degree of statistical control." Sammons et al. (1996, p. 23) suggested that the measure should ideally be collected at the point of entry to school or at the beginning of a relevant phase (such as primary or secondary). This however poses a problem for cross-sectional studies, wherein often, no measure about prior achievement is available. Additionally, under some conditions, researchers even advise against the inclusion of measures for prior achievement, especially if the prior achievement measures are too close to the point where the school effects are measured – thus endangering that part of the variance if factored out not only due to background factors, but also due to the effects of schooling or interaction between school and background (Teddlie & Reynolds, 2000, p. 96). Especially for studies of primary-level

school effects where typically no or only unreliable measures of prior achievement are available, they still regard a measure of family background as the best control variable.

Usually, magnitudes of school effects are measured as the variance accounted for in the student outcomes that can be attributed to the school. Although this is the "scientifically accepted method for analyzing school effectiveness studies" (Teddlie & Reynolds, 2000, p. 97), this approach is controversially discussed, as variance components between levels of analyses might be confounded and because the variance on higher levels of the educational system are generally lower than on the lower levels. Therefore, Rutter (1983, p. 4) argued that "Family variables will usually have a greater 'effect' than school variables. But this does not necessarily mean that schools have a lesser influence than families on achievement." Similarly, Bosker and Scheerens (1989, p. 745) concluded that independent variables closer to the outcome measures – like time on task – would explain more variance than, for example, school-level characteristics such as the leadership style of the director. Teddlie and Reynolds (2000, pp. 102–104) therefore discussed alternative estimates of effect sizes, which are mainly based on the percentage of standard deviations that might be more appropriate for describing the magnitude of school effects in certain models.

Results from research considering the magnitude of school effects

A detailed summary about research related to the properties of school effects is provided by Teddlie and Reynolds (2000), complemented by Reynolds et al. (2014) and Chapman et al. (2015). While there are considerable differences in findings related to the magnitude of school effects, some general conclusions can be drawn. Studies often show higher effect sizes in lower grade levels than in upper grades (Teddlie & Reynolds, 2000). The effects tend to be larger for disadvantaged groups, especially for students with low socio-economic background or with low prior attainment (Chapman et al., 2015). Dar and Resh (1994, p. 9) explained these findings with the *differential sensitivity hypothesis*, stating that students with low socio-economic status profit more from schooling, "since coming from a poor social milieu makes them more school-dependent." Also, it seems that schools usually have a larger effect related to cognitive outcome variables compared to non-cognitive outcomes (Opdenakker & van Damme, 2000; Thomas, 2001). Possible explanations given for these differences include that non-cognitive outcomes receive less emphasis in the curricula, that measurement of non-cognitive outcomes might be less precise, or that students focus more on non-cognitive activities outside schools (Reynolds et al., 2014). Findings also often suggest that the size of school effects differs among subjects. Effects in general seem to be higher for mathematics and science, subjects that are usually

mainly learned in school, compared to, for example, language – where the influence of the home background in particular tends to be stronger (Brandsma & Knuver, 1989; Teddlie & Reynolds, 2000). Besides, the degree of the sensitivity of the assessment used as outcome criteria is expected to influence the magnitude of school effects, with higher effects expected if the test is measuring instruction in the classroom more specifically. Hill and Rowe (1996, p. 27) concluded from findings of their research and from similar studies of other authors that "the greater the sensitivity of outcome measures to the curriculum and to teaching and learning as experienced by students in classes and schools, the greater the proportion of variance explained by effects at these levels."

In longitudinal study designs, higher effects are often reported compared to studies using cross-sectional designs (Teddlie & Reynolds, 2000; van de Gaer et al., 2009). Hill and Rowe (1996, pp. 27–28) explained lower effects in cross-sectional designs with the risk that school effects in such cases might be "unobserved, under-estimated, or partialled out through the statistical adjustments used to control for differences in intake characteristics."

In general, after controlling for intake factors, research findings indicate a level of around 5 to 15% of variance on school level (Brandsma & Knuver, 1989; Chapman et al., 2015; Teddlie & Reynolds, 2000). Scheerens (1992, p. 70), when summarizing the results from several school effectiveness analyses, came to a similar conclusion: "When we look at school effectiveness research in The Netherlands, for example, it seems that the average variance between schools amounts to 11 or 12 per cent of the total variance. This percentage hardly deviates from the results of the American and British studies discussed…" Bosker and Witziers (1996), in a statistical meta-analysis of 103 school effectiveness studies, also confirmed these findings. The explained variance for school-level factors, on average, was found to be 8% once results were adjusted for student background. They generally found larger school effects in developing countries, followed by studies from North America, the UK, then the Netherlands; the lowest effects for industrialized nations was in the Pacific Rim.

Although the effects measured in terms of explained variance are usually small, there is increased recognition that they should be considered as important as they might affect a large number of students and might accumulate over time. While effects measured on a teacher- or class-level are generally larger than general school effects (Hill & Rowe, 1996), it should be considered that students spend several years in school and experience usually a number of different teachers; thus, the accumulation of effects over time can be large. Reynolds et al. (2011, p. 13) concluded in their *State-of-the-art review of EER*: "Even rather small school effects are

considered important because they might be cumulative, they may refer to a large number of students, and they may make a difference to outcomes that shape later life chances."

3.4.3 Consistency of school effects

The consistency of school effects describes the correlation between different outcome measures at one point in time. In general, the measurement of a construct should be improved if different measures of that construct are collected (Teddlie & Reynolds, 2000, p. 116). Transferred to EER, this would mean that a conclusion regarding a school's effectiveness could be taken with more confidence, or would be more valid, if different outcome measures would be used to measure its effectiveness. Consequently, numerous scholars call for the inclusion of various measures of different types of outcomes when judging a schools' effectiveness. In addition to a differentiation between the levels of cognitive measures (*basic skills* versus *higher order thinking skills*), these could also include measures of attendance, attitudes, discipline, or affective variables such as satisfaction (Good & Weinstein, 1986; Levine & Lezotte, 1990; Teddlie & Reynolds, 2000).

Research concerning the consistency of effects across outcome measures is a bit inconclusive. While small to moderate correlations between different cognitive measures are found more consistently in primary education, findings reported for secondary schools are more heterogeneous. Bosker and Scheerens (1989), analyzing results of school effectiveness studies in The Netherlands and Great Britain, reported a correlation of 0.7 to 0.75 across subjects on the primary level and a broader range of 0.45 to 0.75 for the secondary level. Sammons et al. (1996), who summarized the results related to the consistency of school effects from several British studies in secondary schools, indicated similar correlations in the range of 0.4 to 0.5. Inconsistencies in school effects across subjects are mainly explained by differences in teaching quality, by differential departmental effectiveness, or by a stronger focus of a school on a specific subject (Luyten, 1994, p. 213; Reynolds et al., 2014, p. 207).

As most effectiveness studies have focused on one or two cognitive achievement criteria, less evidence can be reported concerning the consistency of non-cognitive outcome measures or between cognitive and non-cognitive measures. Research findings related to the consistency between cognitive and non-cognitive measures are gaining interest, as a positive association would support the hypothesis that both dimensions are self-complementary, while a negative correlation would require a choice regarding the school objectives, which in consequence would result in a trade-off between different educational goals in schools (van der Wal & Waslander,

2007). Several authors (Brandsma & Knuver, 1989; Knuver & Brandsma, 1993; Mortimore et al., 1988; Rutter et al., 1979) discussed these issues. While earlier studies, such as that implemented by Rutter et al. (1979), found a stronger correlation between social outcome factors (attendance and delinquency) with academic outcome measures in British secondary schools, most later studies found only weak or no relation between the cognitive and the non-cognitive dimensions. Using data from 50 British primary schools, Mortimore et al. (1988) concluded in their analysis that the cognitive and the non-cognitive dimensions of effectiveness are rather independent. Knuver and Brandsma (1993), based on their research in 212 Dutch primary schools, found weak but never negative correlations between affective measures (attitudes, motivation, self-concept, and well-being) and cognitive outcomes (language and mathematics). They carefully concluded that "effectiveness in the cognitive and affective domain can go together, at least in such a way that effectiveness in the one domain does not hinder the effectiveness in the other domain." (Knuver & Brandsma, 1993, p. 202)

3.4.4 Stability of school effects over time

A growing body of research concerns the investigation of the stability of school influences over consecutive years. Teddlie and Reynolds (2000, p. 122) concluded in their evaluation of the research body that "Early estimates of the stability of school effects where relatively low, while more recent studies, with more advanced methodologies, have yielded higher estimates." Empirical findings concerning the stability of effects were first summarized by Bosker and Scheerens (1989), who reported the correlations listed in Table 3-1 below (which, however, do not include information regarding the statistical significances of the reported results).

Table 3-1: Range of stability estimates (correlation coefficients) for school effects taken from Bosker and Scheerens (1989)

	Primary	Secondary
Across years	.35 - .65	.70 - .95
Across grades	.10 - .65	.25 - .90

The table shows Pearsons'r correlations expressing the extent to which school effects correspond between two different school years and between different grades in primary and secondary schools.

In general, Thomas, Sammons, Mortimore, and Smees (1997) reported similar results. They found relatively stable effects over a three year period, ranging from 0.82 to 0.88 on total GCSE (General Certificate in Secondary Education) performance scores, but with more fluctuation for specific subjects.

Reynolds et al. (2014) argued that stability over several years is extremely unlikely, as changes in school policies, a new director, or changes in staff and student body might influence the system. For example, based on their research conducted in London inner city schools, Mortimore et al. (1988) showed a substantial increase in the influx of disadvantaged students within a three year period of time; in other cases, educational reforms are assumed to have influence related to school effects over time. Teddlie and Reynolds (2000, p. 123) concluded that there is a "fair degree of stability in secondary schools' effect on overall measures of academic achievement (…) over time…" They saw a similar trend for basic skills in primary education, albeit the correlations in general were smaller. Researchers recognize that the generalizability of school effects over time can only be studied properly in longitudinal research designs, as processes related to effectiveness might change, which in turn can lead to underestimations of school effects in cross-sectional studies. Thus, researchers argue that judgements concerning the effectiveness of a school should be based on data from several years and a range of different aspects and outcome criteria (Thomas et al., 1997; Thomas, 2001).

3.4.5 Differential effects

While a school's overall effect relates to the impact of that school for an "average" student, there is a possibility that school effects vary across different groups of students or across different units within schools. The question of whether schools might be more effective for specific student groups is seen as especially relevant for the equity dimension of school effectiveness research; the extent to which the effects of schooling should be/ need to be the same for all students of a certain school also merits investigation (see also the passage about the *equity dimension* in section 3.2).

While a number of studies have investigated the differential effects of schooling mainly with a focus on prior achievement, gender, SES, and ethnicity, findings remain inconclusive. For example, in terms of gender, some studies related to differential effects produced no clear evidence of such effects (Brandsma & Knuver, 1989; Mortimore et al., 1988; Thomas, 2001), while others found some evidence in this regard (Strand, 2010). Concerning prior achievement, some authors found differences among schools concerning the association of prior achievement

with later achievement (Strand, 2010; Thomas, 2001). In regard to ethnicity, some authors pointed to differential school effects (Nuttall, Goldstein, Prosser, & Rasbash, 1989), while others did not found any such differences (Mortimore et al., 1988; Strand, 2010). Several studies reported small differential effects related to socio-economic status (Strand, 2010; Thomas, 2001), while Mortimore et al. (1988), for example, found no evidence of differential effects related to social-class background. Teddlie and Reynolds (Teddlie & Reynolds, 2000) concluded after evaluating the literature that at secondary level there is some evidence of differential effects related to SES, ethnicity, and prior achievement, while less evidence for such effects exists at the primary level.

3.4.6 Composition effects

It has been frequently demonstrated that individual background factors are important correlates for school learning, and that in many countries student background is strongly associated with school outcomes (Baumert, Watermann, & Schümer, 2003; Martin & Mullis, 2013; Sirin, 2005). Consequently, a differential increase of student achievement can be expected depending on his or her background, which accumulates over years of schooling; Baumert (2006, p. 101) called this the *individual Matthew effect* [individueller Matthäuseffekt].

However, it was found that beyond the effect at the student level, additional effects related to the *student composition* (proportion of students with a certain characteristic) appear, which in turn can have an additional influence on an individual's development. Baumert called composition effects that originate in differences in student composition *institutional Matthew effects* [institutioneller Matthäuseffekt] (Baumert, 2006, p. 101).

It should be noted that these kinds of effects are not defined uniformly among scholars. While at times they are described as *compositional effects* or *contextual effects*, other authors – for example, Gorard (2006) – use the term *school-mix effects*. As the term *contextual effect* is often used more broadly, and also describes other differences between schools concerning grade levels, governance structures and so forth, in this thesis the term *compositional* will be used for these effects, according to the definition of Harker and Tymms (2004, p. 183) who described them as "the statistical estimate of the additional effect obtained by the aggregated variable at the school level over-and-above the variable's effect at the individual level."

Since the Coleman Report stated that "the social composition of the student body is more highly related to achievement, independent of the student's own social background, than is any school factor" (Coleman et al., 1966, p. 325), school composition has been a major research topic. In

addition to the composition of socio-economic variables, other compositional effects have also been analyzed, including the mean prior achievement, the gender composition (proportion of boys and girls), and the proportion of ethnic minorities.

Different explanations for such effects can be found in the research literature. Willms (1992, p. 41), for example, stated that schools with high social class or high ability intake are advantaged, as in such schools there is likely greater support of parents and a more orderly atmosphere, conducive to learning. He also argued that those schools might attract more talented and motivated teachers, and that positive peer effects among students might occur. Similarly, Harker and Tymms (2004) categorized the reasons for such effects under the main headers of *peer effects*, *teaching effects*, and *facilities effects*.

However, although many indications are given for the existence of composition effects (Harker & Tymms, 2004; Reynolds et al., 2014; Teddlie & Reynolds, 2000), there is still no clear consensus about the nature of those effects. It is often argued that composition effects might occur only as artifacts in inadequately specified models (Harker & Tymms, 2004; Hauser, 1970; Nash, 2003). Harker and Tymms (2004) call this the *phantom effect*. Baumert and Schümer (2001), when analyzing the German PISA data, initially found large social composition effects. However, once they controlled for prior achievement on an individual level, and as an aggregated variable on school level, they found that the effect of the social composition reduced to only one-eighth of the original value. Based on such findings, Baumert (2006, p. 109) concluded that composition effects generally will be overestimated in cases where no indicators of prior achievement are included in the analysis models. This is a problem, particularly for cross-sectional studies where indicators of prior student achievement are usually not available. Similarly, Verhaeghe, van Damme, and Knipprath (2011) argued that in cross-sectional studies, cumulative effects of school composition might be confounded with both effects from processes during the time of schooling and preexisting effects. They concluded that these effects can only be measured properly in longitudinal designs. This argumentation is supported by Reynolds et al. (2014), who reported that most of the longitudinal studies investigating school composition have found larger significant composition effects only on the first measurement occasion.

In spite of the controversial later discussion regarding the nature of composition effects, Teddlie and Reynolds (2000) concluded in their summary of the knowledge base that the existence of composition effects can be demonstrated for all three strands of school effectiveness research: school effect studies, effective schools research, and school improvement studies. In their more

recent *State-of-the-art review of educational effectiveness research*, Reynolds et al. (2014) confirmed these findings. They stated that the majority of studies analyzing composition effects generally found positive effects in relation to achievement for all students attending schools with a high average achievement, a high proportion of girls, and a high average SES, while no clear effects could usually be determined concerning the ethnic composition, once the data was controlled for socio-economic composition.

This section can be concluded by a summary of the main topics detailed above, provided by Chapman et al. (2015, p. 46):

> In summary, the accumulation of research evidence in the EER tradition in a wide range of international contexts confirms that effectiveness is best seen as a dynamic, retrospective, and relative concept that is time- and outcome-dependent, and influenced by the sample of schools studied and the availability of data on relevant predictors (especially the choice of prior attainment measures, and both individual student and composition of intake measures), as well as the adequacy of the statistical modeling approaches used.

4 MODELS OF EDUCATIONAL EFFECTIVENESS

4.1 Theoretical Foundations

EER can be regarded as a rather independent field of study belonging to the discipline of education, but integrating contributions from several different disciplines: mainly economics, sociology, and psychology. Early educational effectiveness studies, as conducted by Coleman et al. (1966) or Hanushek (1986), were based on educational production functions originating from the field of economics. The assumption was that measureable input variables could be related to student achievement. Although there were no consistent findings giving strong and consistent evidence that such input variables alone would be a good measure to predict student achievement, the basic assumption that student performance is a function of controllable and uncontrollable variables is still valid in modern EER. Later on, important contributions to education came from the field of sociology. *Status attainment* theories, which posited that the social status of parents affects educational achievement of their children, expanded the focus to a variety of family background variables which could be included in effectiveness studies (Teddlie & Reynolds, 2000, p. 303). Additionally, the sociological perspective contributed to the measurement of effectiveness by highlighting the role of certain process variables emerging from organizational theories, such as school climate, culture, and structure (Chapman et al., 2015, p. 150). Teddlie and Reynolds (2000, pp. 304–306) listed other important theoretical concepts borrowed from the field of psychology. These comprise the student locus of internal or external control (related to the belief of an individual of the extent to which he/she can control outcomes in life), the teacher and student expectations for student performance (which is based on experimenter biased effects first studied by Rosenthal, 1968) and academic self-concept (which can be defined as the person's personal belief about him- or herself; see also section 3.3.4.3). They elucidated further that teacher effectiveness research, which only at a later stage was integrated into the EER framework, contributed key concepts of effective classroom behavior, such as: quantity and pacing of instruction, opportunity to learn, time allocation, classroom management, active teaching, whole-class teaching versus small group instruction, redundancy, clarity, praising, classroom-climate, etc.

The empirical findings from EER studies, and the theoretical constructs borrowed from other disciplines, expanded the original economics-driven input output paradigm. In this way, a blending of disciplines led to the configuration of a more comprehensive framework: namely, the *instructional effectiveness theory*.

In the beginning of the 1960s, Carroll (1963) developed his model of school learning, the most adopted theory of instructional effectiveness. He stated that the learning rate can be considered as a function of five elements: *aptitude* (amount of time to learn needed under optimal conditions), *ability to understand instruction*, *perseverance* (amount of time students are willing to engage), *opportunity to learn* (time for learning), and *quality of instruction*. The relationship between Carroll's main elements were further elaborated by Bloom (1968) in his framework of mastery learning. The different components of Carroll's and Bloom's models were subsequently extended by other authors, who added contextual, organizational, and further instructional factors. As a result, the integration of findings from different effectiveness research strands led to the development of more comprehensive models of educational effectiveness.

Most of these models combine the following features, albeit with slightly different foci :

- They are based on input-process-output and try to explain the complex interaction between the factors involved.
- They recognize that the influences on student achievement are multilevel (meaning that they have effects on different levels, such as student, class, or school level).
- They combine the organizational/structural and the learning/teaching orientation of educational effectiveness, thus acknowledging that while the teaching and learning process might be at the center of the consideration, effective learning and teaching requires schools to function as organizations.

The following section will describe three integrated effectiveness models which all draw upon the concepts and core elements described in section 3.3, and endeavor to provide further clarification and detail regarding their relationships, in order to explain educational effectiveness. Those models were consulted in more detail as a basis for the framework of this study, as all of them attempt to integrate findings from different effectiveness research strands, and are well-established in the literature; however, each has a slightly different focus, and as such may have the potential to contribute important elements to the framework of the current study.

4.2 Integrated Models of Educational Effectiveness

While most integrated effectiveness models combine organization with the learning/ teaching orientation of effectiveness, the focus in these models is often slightly different. Scheerens' (1992, p. 14) model of integrated effectiveness, for example, places more emphasis on the functioning of the school as an instructional system. Other models, such as Creemers' (1994, p. 119)

integrated model of educational effectiveness, rather focuses on the observable processes related to teaching and learning. Creemers' model focuses predominantly on the classroom level, where teaching and learning mainly occur. Other educational levels are rather seen as preconditions for effective teaching and learning. More recent research findings showed that the nature of educational effectiveness might be more complex than taken into account in the rather "static" integrated models of effectiveness (Creemers & Kyriakides, 2006). Findings supported the notion that educational effectiveness may change over time, and may depend on factors such as the outcome being measured, the current situation and context of the school, and the characteristics of the students being considered. These findings, together with a better elaboration of factors that are important for the quality of instruction and the importance of school- and classroom climate factors (among others), led to the development of a new group of models: the "dynamic" models of educational effectiveness (Creemers & Kyriakides, 2008). While these models are assumed to describe the nature of effectiveness more precisely, the fact that they make high demands of the data that would be needed to empirically support them should also be considered. The framework of the current research project, therefore, will mainly adopt Creemers' approach and focus more on the rather stable and assumingly generic factors of educational effectiveness. Creemers' model was chosen as a basis for the current framework over other comparable comprehensive models, such as Scheerens' model (1992) or the approach taken by Shavelson, McDonnell, and Oakes (1989), as it focuses predominantly on the classroom level – where teaching and learning take place. Moreover, in order to elucidate educational outcomes, Creemers' model focuses on what it defines as requisite aspects of learning theory: namely, *time*, *opportunity*, and *quality*. Creemers' model is also well-established in the research community, and his approach will be described comprehensively in section 4.2.1. Additionally, the more complex dynamic model will be consulted to complement Creemers' base model with more recent research findings. A detailed description of the dynamic model can be found in section 4.2.3. While not included in the above-mentioned models, input and context factors are also regarded as important; accordingly, Scheerens' approach will be consulted in this regard, and his model will be shortly described in the subsequent section (4.2.1). Finally, research related to more recent models of instructional effectiveness (Helmke, 2009; Klieme & Baumert, 2001; Nilsen & Gustafsson, 2016; Seidel & Steen, 2005) will be considered and later used to complement the research framework of the current study.

4.2.1 Scheerens' Model

Scheerens' (1992) model of integrated effectiveness is a multilevel model that tries to explain effectiveness from an economical and organizational perspective, in contrast to Creemers (see

the next section) who focuses more on the processes in the classroom. According to Scheerens (1992, p. 29), in this sense, "Organizational arrangements are seen as both direct and indirect 'causes' of the performance of pupils." His model, which is depicted in Figure 4-1, combines the Input-Process-Output dimension with a dimension describing the hierarchical levels (context, school, classroom, and to a certain extent the student level) by setting the economic productivity function into a multilevel environment.

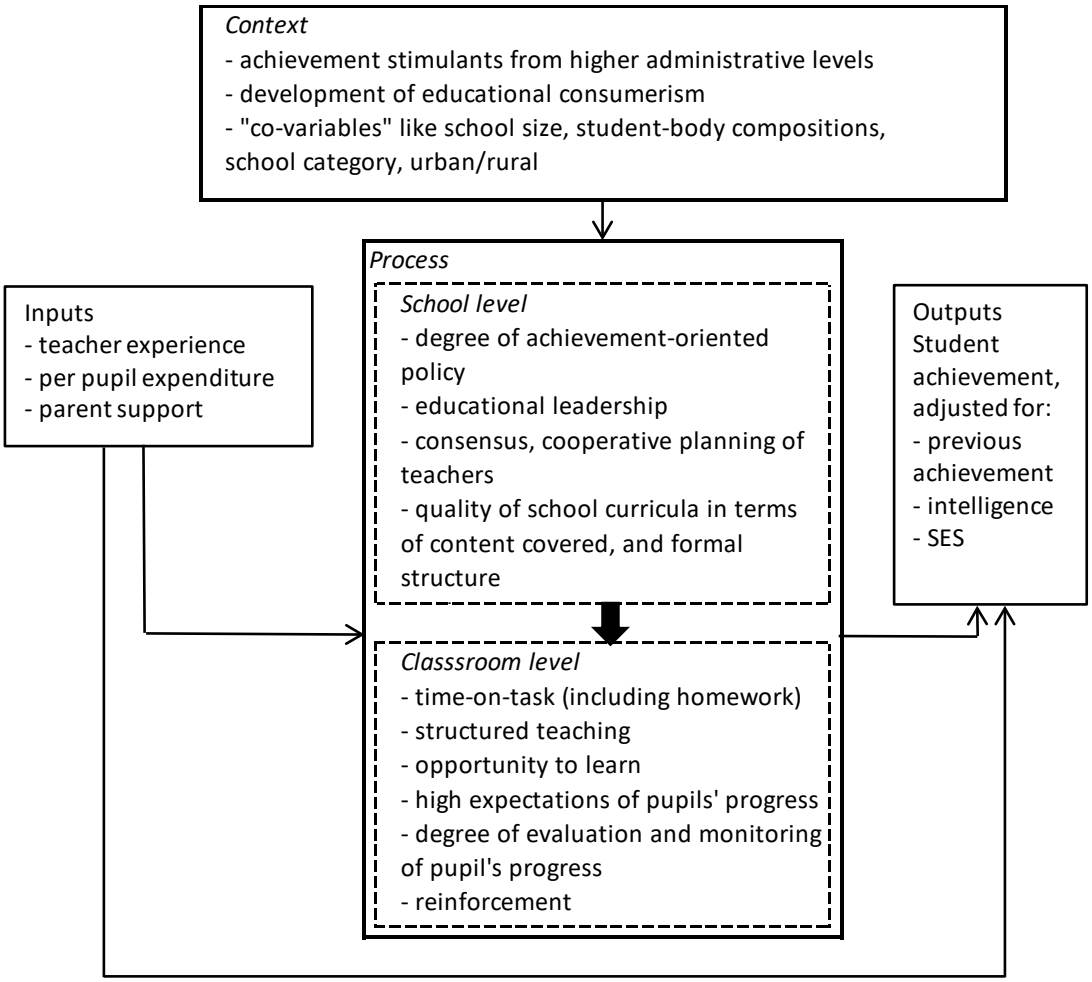


Figure 4-1: Integrated model of school effectiveness from Scheerens (1992, p. 14)

Student outcomes are assumed to be influenced by input, process, and context variables. School- and classroom-level processes mainly concern attitudes, climate variables, and teaching practices. Scheerens additionally defined an input cluster on school level, which, in addition to monetary resources, also contained teacher experience and parental support as important input factors. In addition, school and classroom processes – and, consequently, student outcomes – are seen as being influenced by higher administrative levels as well as by school context variables, such as the student-body composition, or school category.

4.2.2 Creemers' Model

Creemers' comprehensive model of educational effectiveness (Creemers, 1994) was based on empirical research on effective instruction and student learning and took earlier models, such as the model of school learning from Carroll (1963), into consideration. Creemers distinguished between four different levels of education (context, school, class, and student); his model is therefore multilevel in nature. However, Creemers placed strong emphasis on the teaching and learning process in the classroom, as stated below:

> From a theoretical and empirical point of view, the classroom is the predominant place in the school where learning and teaching takes place, and in this way the classroom level is more important for learning and outcomes than other levels in education. (Creemers, 1994, p. 5)

Higher levels are rather seen as providing the conditions for teaching and learning which influence outcomes (usually indirectly) by influencing the factors and elements at the classroom level. Outcomes, therefore, are seen as the combined effect of educational levels, also including influences from the student level, as it is finally the student who decides how much time and attention he or she will spend on learning.

An additional aspect of the model is that it distinguishes three major components (*time*, *opportunity,* and *quality*), which are assumed to influence outcomes across the different educational levels. While these components emerged from Carroll's model, Creemers elaborated especially on the quality of instruction, which he splits into curriculum related factors, grouping procedures, and teacher behavior. Quality of instruction is seen as either influencing learning outcomes directly or indirectly, by influencing the components time and opportunity at classroom level. He also distinguished clearly between time on task and opportunity to learn, and differentiated between the available time and opportunity provided by the school environment and the teacher, and the time and opportunity which is actually used by the student.

Additionally, Creemers introduced the four formal principles of *consistency, cohesion, constancy*, and *control* to describe the assumed joint impact of the various model factors on student outcomes. These principles concern the relationships between the different model factors. *Consistency* is based on the assumption that the effectiveness of the educational levels increases when the factors at these levels are aligned with each other. This might lead to a synergistic effect which exceeds the effectiveness of the separate components (Creemers & Kyriakides,

2008, p. 44). *Cohesion* is created if staff members show consistency in their effectiveness characteristics. *Constancy* is attained if consistency and cohesion persist over a longer period of time – meaning that effective instruction should be provided throughout the students' years of schooling. *Control* finally refers to the evaluation of student outcomes and teacher behavior, the maintenance of an orderly climate, and to teachers holding each other responsible for effective instruction. An overview of Creemers model is given in Figure 4-2. More detail on each of the different educational levels is provided in the sections below and detailed information about all effectiveness factors is provided in Table 4-1.

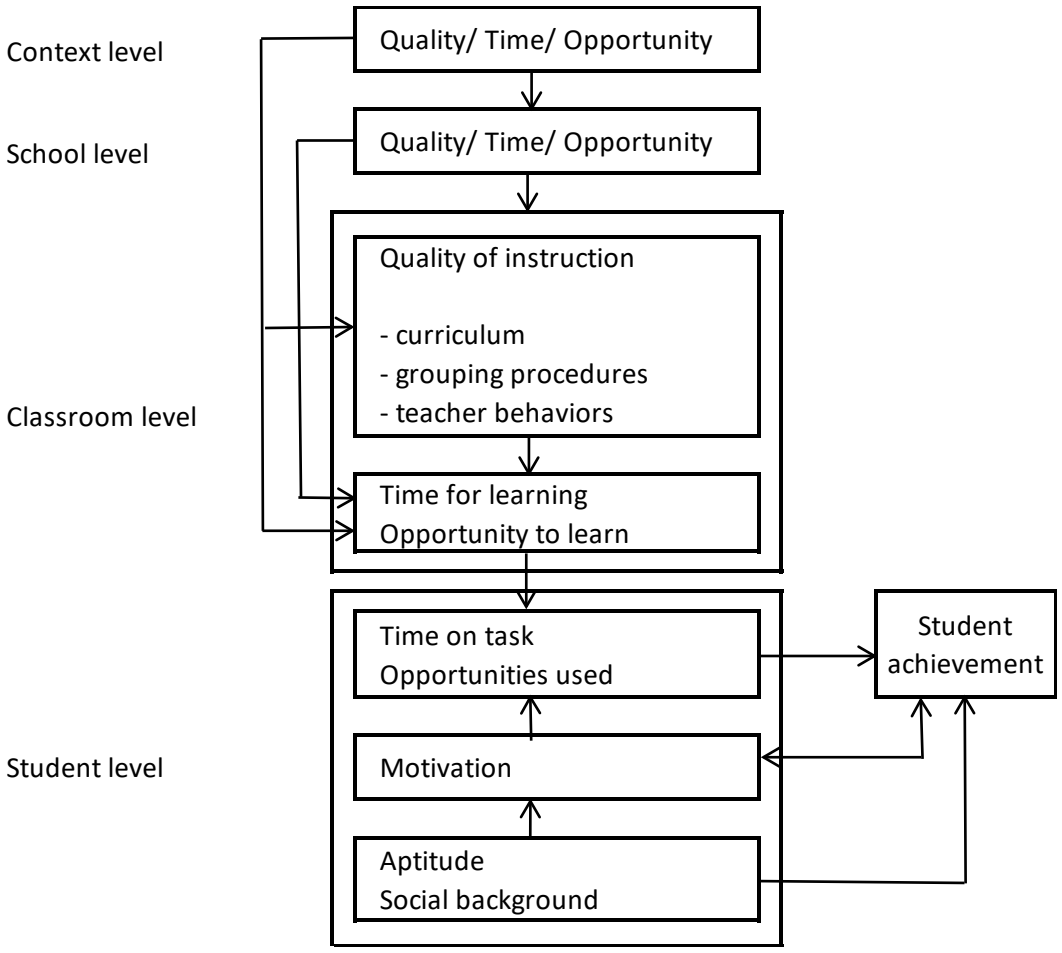


Figure 4-2: Creemers' comprehensive model of educational effectiveness – overview (taken Creemers, 1994, p. 27)

4.2.2.1 Context level factors

The highest educational level identified by Creemers (1994) is the context level, which defines certain preconditions for the levels below, the school level and the classroom level. Similarly to the other levels below, factors on the context-level are categorized into *quality*, *time*, and

opportunity. Creemers (1994, p. 122) lists the following important conditions affecting the levels described below:

- in terms of *quality*: a national policy regarding the effectiveness of education and the availability of an indicator system, a national policy regarding evaluation, a national testing system, training and teacher support systems, and the funding of schools based on outcomes (school accountability).

- In terms of *time*: national guidelines in regards to the schedules of schools and the supervision of the maintenance of these schedules.

- In terms of *opportunity*: national guidelines and rules related to the national curriculum.

4.2.2.2 School level factors

At the school level, Creemers (1994, pp. 120–121) focused on conditions for important class-level factors which are relevant for effective instruction in terms of *quality of instruction*, *time*, or *opportunity to learn*. Their influence is mediated by the time on task and by the opportunities used by the students. Concerning the quality of instruction, Creemers distinguished further between educational and organizational aspects. With regard to the educational aspects, rules and agreements concerning the instructional process at the classroom level (for example, related to curricular materials used, grouping procedures, or teacher behavior) and evaluation policies to prevent and correct learning problems are seen as most important. Concerning the organizational aspects, school policies on the supervision of school staff and the professionalization of teachers, as well as the establishment of a school culture supporting effectiveness, are seen as most important. Conditions for time at the school level are related to the time schedule of subjects and topics, and to the maintenance of an orderly and quiet atmosphere which is conducive to learning. The development and availability of a curriculum or similar, consensus about the mission of the school, and rules and agreements about the implementation of the curriculum are important prerequesits for the factor opportunity to learn on school level.

4.2.2.3 Class level factors

Creemers considers the learning processes taking place in the classroom as the main factors determining educational outcomes ” (Creemers, 1994, p. 5). Consequently, he regarded the instructional conditions as an essential component, having elaborated on the quality of instruction, which he split into three major areas: curriculum, grouping procedures, and teacher behavior (see Creemers, 1994, pp. 118–120). The *quality of instruc*tion also influences *time for*

learning and *opportunity to learn*, and thus exerts a direct, as well as indirect, effect on student learning.

Curriculum refers to the material used by teachers and students in the instructional process such as textbooks. The curriculum should be developed according to well-stated and clear educational goals, and should serve as a guideline for the other areas – the grouping procedures and the teacher behavior. Accordingly, the implementation of the curriculum by the teacher is in a central focus. Grouping procedures which are based on mastery learning and follow the curriculum should be accompanied by evaluation, immediate feedback, and individually adjusted instruction in order to detect and overcome deficiencies in student's learning. Grouping also influences the allocation of time and the opportunity to learn. Teacher behavior is differentiated into two important components: management behavior, to control the class and thus maximize learning time and opportunity; and the instructional behavior related to effective teaching. The latter is elaborated in more detail in section 3.3.5.2.

4.2.2.4 Student level factors

In line with the literature on educational effectiveness, Creemers' model also lists individual factors such as motivation, aptitude, and background as being important determinants for academic outcomes on student level (see Creemers, 1994, p. 118). Concerning motivation, Creemers assumed a reciprocal effect. He claimed not only that motivation has an effect on academic outcomes, but also that academic outcomes might affect motivations and attitudes towards learning. As defined in section 3.3.4.1, aptitude is an indicator of what a student already knows, and includes general ability or intelligence and prior learning. The social background factor reflects the SES of the student, which is seen as one of the most important factors in explaining student outcomes (see section 3.3.4.3 for more detail).

Time on task in his model is specified as the time students are willing to spend in school learning. However, the time needs to be filled with opportunities to learn, such as learning material or experiences and exercises. Creemers described the learning opportunities as the "instructional operationalization of the objectives of education" or the "content coverage" of the curriculum (Creemers, 1994, p. 118).

Table 4-1: Factors of Creemers' comprehensive model of educational effectiveness – detailed version (from Creemers, 1994, p. 119)

Levels	Components		Characteristics of the components	Formal criteria
Context	Quality		Policy focusing on effectiveness Indicator system/policy on evaluation/ National testing system Training and support system Funding based on outcomes	Consistency Constancy Control
	Time		National guidelines for time schedules Supervision of time schedules	
	Opportunity		National guidelines for curriculum	
School	Quality (educational)		Rules and agreements about classroom instruction Evaluation policy/evaluation system	Consistency Cohesion Constancy Control
	Quality (organizational)		Policy on intervision, supervision, professionalization School culture inducing effectiveness	
	Time		Time schedule Rules and agreements about time use Orderly and quiet atmosphere	
	Opportunity		School curriculum Consensus about mission Rules and agreements about how to implement the school curriculum	
Classroom	Quality of instruction	Curriculum	Explicitness and ordering of goals and content Structure and clarity of content Advance organizers Feedback Corrective instructions	Consistency
		Grouping procedures	Mastery learning Ability grouping Cooperative learning highly dependent on Differentiated material Evaluation Feedback Corrective instruction	
		Teacher behavior	Management/orderly and quite atmosphere Homework High expectations Clear goal setting Restricted set of goals Emphasis on basic skills Emphasis on cognitive learning and transfer Structuring the content Ordering goals and content Advance organizers Prior knowledge Clarity of presentation Questioning Immediate exercises Evaluation Feedback Corrective instruction	
	Time for learning Opportunity to learn			
Student	Time on task Opportunities used Motivation Aptitudes Social background			

Creemers' approach is regarded as one of the most influencial theoretical constructs in the field (Teddlie & Reynolds, 2000). Its validity was examined by several authors (for example de Jong et al., 2004 or Kyriakides, 2006). De Jong et al. (2004), who conducted a study of mathematics in the first year of primary education in the Netherlands, found that the amount of time spent, the opportunity to learn, and the quality of instruction were strong predictors of achievement. Analyses undertaken by Kyriakides (2006) using IEA TIMSS 1999 data resulted in a number of variables related to the three main factors of Creemers' model, but he only could explain a small percentage of unexplained variance. However, up to present, educational studies testing Creemers' model outside of the Western world are still rare; consequently, Kyriakides (2006, p. 528) noted the necessity of further analysis of data from international comparative studies, in order to investigate the validity and generalizability of Creemers' model.

4.2.3 The dynamic model of educational effectiveness

More recent research provides evidence that the relationship between different effectiveness-enhancing factors might be more complex, and that some important additional components are still missing in the previous models. Based on the weaknesses and gaps detected during empirical testing of Creemers' comprehensive model, Creemers and Kyriakides (2008) developed Creemers' model further, in a more dynamic direction.

Similarly to the integrated models of educational effectiveness, the dynamic model takes into account the fact that influences on student achievement are multilevel, and distinguishes between four different levels (context, school, class, and student). Like Creemers' model, the dynamic model also emphasizes factors related to teaching and learning on class level. It is assumed that higher-level factors may influence teaching and learning situations both directly and indirectly, via policies and regulations. The model also emphasizes the dynamic processes and conditions associated with teaching and learning. The model therefore claims, for example, that policies on higher levels are evaluated over the years and related to the particular weaknesses that occur in a school (Creemers & Kyriakides, 2008, p. 78). Stronger emphasis is also put on the development of a school learning environment (SLE) that promotes educational outcomes. The model tries to link the areas of school effectiveness with school improvement. Consequently, only changes in those areas where the schools face specific weaknesses are regarded as important in terms of being altered to improve school effectiveness. These areas should be known to the school by means of regular school evaluation, and measures need to be taken to remedy the detected weaknesses. Thus, here the model incorporates elements of other organizational theories, such as *contingency theory* (Donaldson, 2001) and *cybernetics* (Ashby, 1961;

Stacey, 2007). An overview on the main characteristics of the dynamic model can be found in Figure 4-3.

When compared to the previously described models, an important distinction is that different dimensions are used for measuring how the identified effectiveness factors work. This implies that the factors should not only be examined by measuring their frequency, but also need to be investigated in terms of their *quality* or how they are functioning. Here, Creemers and Kyriakides (2008) saw factors as multi-dimensional constructs and measured them in five different dimensions: *Frequency*, *Focus*, *Stage*, *Quality*, and *Differentiation,* which are described in the subsequent paragraphs.

Frequency refers to the quantity of an activity associated with an effectiveness factor. However, in the dynamic model the association does not necessarily need to be a linear one. The frequency of "personal monitoring", for example, might exhibit a curvilinear relation with outcomes (Creemers & Kyriakides, 2008, p. 84).

Focus relates to two different aspects: first, the specificity of the activity as such can be more general or more specific; second, the specificity and number of the purposes for each activity. Creemers and Kyriakides stated, for example, that a policy on parental school involvement might be very specific (parents may visit schools only at specific hour), but at the same time it is multi-purpose (parents may visit schools to exchange information about children and to assist teachers inside and outside classroom). Curvilinear associations between specificity and number of purposes may also be expected in such cases (Creemers & Kyriakides, 2008, p. 85).

Stage refers to the duration a factor remains active, and relates to the principle of *constancy* in Creemers' model. While measuring the stage gives information about the continuity of a factor, it is worthwhile to consider that the activities associated with these factors might change in the process of self-evaluation processes and subsequent redefinition of the policies.

Quality denotes the construct validity (the properties of a construct), but also describes the extent to which staff make use of the policies and documents available to ensure the quality of instruction.

Differentiation refers to the extent that activities associated with an effectiveness factor can be seen as generic for all student groups. Put differently, differentiation refers to the adaptive implementation needed for different student groups (such as SES, thinking styles, motivation, or prior knowledge).

However, while the model allows for a detailed description of the complex nature of educational effectiveness, measuring each factor in five different dimensions makes the model quite complex.

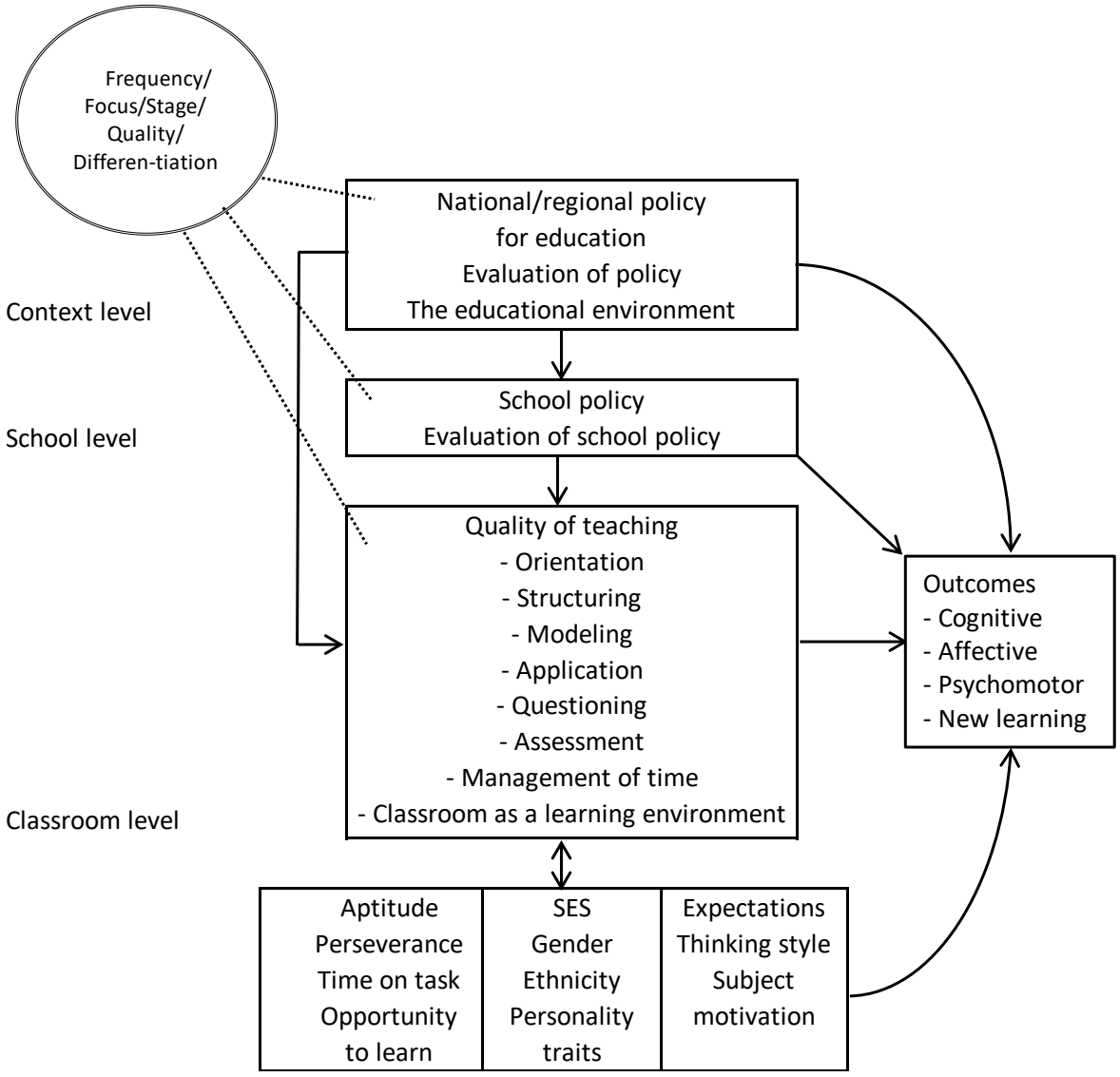


Figure 4-3: Main Characteristics of the Dynamic Model (from Creemers & Kyriakides, 2008, p. 150)

4.2.3.1 Context factors of the dynamic model

Creemers and Kyriakides' (2008) dynamic model doesn't focus on a specific structure of an educational system, but rather focuses on policies affecting learning inside and outside the classroom and on their regular evaluation. In particular, context and school level policies that are related to teaching practices and to the school learning environment are regarded as essential in affecting teaching practice in classrooms and, in turn, student learning outcomes.

Similarly to Creemers' model, also the dynamic model focuses on relevant factors in teaching and learning in regard to the dimensions of *quantity* (time), *quality*, and provision of *learning opportunities*. Secondly, the evaluation mechanisms of the national educational policies are assumed to contribute to the improvement of educational effectiveness on a system level.

4.2.3.2 School factors of the dynamic model

While school factors are assumed to influence student outcomes partly directly, they are expected to influence them mainly indirectly – via influence on the classroom level and especially on teaching practices. Elements that provide the conditions for the same essential concepts of quantity and quality of teaching, and the provision of learning opportunities that were used to define class-level factors, are especially emphasized. The model therefore highlights two aspects that are assumed to affect learning and teaching and, consequently, the student outcomes: school policies regarding teaching, and school policies regarding the creation of an effective school learning environment. Policies here do not only comprise formal documents and guidelines, but "mainly refer to the actions taken by the school to help teachers and other stakeholders have a clear understanding of what they are expected to do" (Creemers & Kyriakides, 2008, p. 118).

Altogether, on school level Creemers and Kyriakides regarded the subsequent four important factors in the model:

- School policy for teaching and actions taken for improving teaching practice;
- School policy for creating an SLE and actions taken for improving the SLE;
- Evaluation of school policy for teaching and of actions taken to improve teaching, and
- Evaluation of the SLE

4.2.3.3 Class factors of the dynamic model

Similar to Creemers' integrated model, the dynamic model also focuses on the classroom environment, referring to factors that are related to teacher instruction and associated with student outcomes. Only observable factors (teacher behaviors) are regarded in their model meaning that explanatory factors such as teacher beliefs and knowledge are not taken into account.

The model distinguishes between eight main instructional factors: *orientation*, *structuring*, *questioning*, *teaching/modeling*, *application*, *the teacher's role in making the classroom a learning environment*, *time management*, and *classroom assessment* which are described in more detail in Table 4-2.

Table 4-2: Main elements of the dynamic model on the teaching level (summary taken from (Chapman et al., 2015, p. 116)

Factors	Main elements
Orientation	- Providing the objective of a specific task/lesson/series of lessons - Challenging students to identify the reason why an activity is taking place in the lesson
Structuring	- Begining with overview and/or review of objectives - Outlining the content to be covered and signalling transitions between lesson parts -Drawing attention to and reviewing main ideas
Questioning	- Raising different types of questions (i.e. process and products) at approriate difficulty level - Giving time for student to respond - Dealing with student responses
Teaching/modelling	- Encouraging students to use problem-solving strategies presented by the teacher or other classmates - Inviting students to develop strategies - Promoting the idea of modelling
Application	- Using seat work or small group tasks in order to provide needed practice and application opportunities - Using application tasks as starting points for the next step in teaching and learning
Time management	- Organizing the classroom environment - Maximizing engagement rates
Making the classroom a learning environment	- Establishing on-task behavior through the interactions promoted (i.e. teacher-student and student-student interactions) - Dealing with classroom disorder and student competition by establishing rules, persuading students to respect them, and using the rules
Classroom assessment	- Using apppropriate techniques to collect data on student knowledge and skills - Analyzing data in order to identify student needs, and reporting the results to students and parents - Evaluating own practice

In contrast to earlier effectiveness models, teaching in this instance does not focus only on the acquisition of basic skills through approaches such as direct teaching, but rather follows a more integrative approach which also covers new goals of education associated with theories of teaching in line with constructivism (see also section 3.3.5.1), as research indicates that both strategies might be equally effective (Louis et al., 2010). Louis et al. therefore suggested to

combine both approaches into an overarching construct, which they called "focused instruction" (Louis et al., 2010, p. 39).

4.2.3.4 Student factors of the dynamic model

Firstly, the dynamic model includes all the student level factors of Creemers' model (i.e., aptitude, socio-economic background, motivations, time on tasks, and opportunities used). Additionally, the dynamic model includes personal characteristics of students that were found to be associated with learning gains. In general, the dynamic model distinguishes between two main categories of factors: 1. socio-cultural and economic background variables emerging from a sociological perspective, and 2. background variables emerging from a psychological perspective. Figure 4-4 shows the student-level factors of the dynamic model and their assumed interrelations.

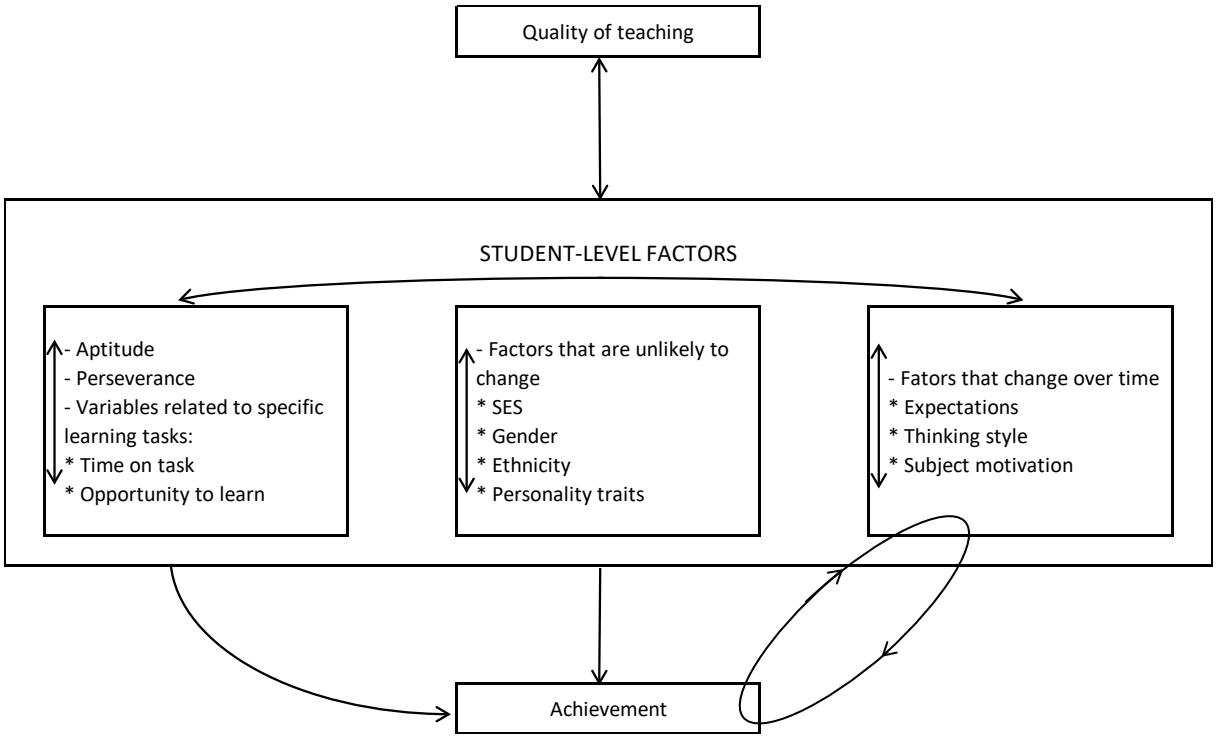


Figure 4-4: Factors of the dynamic model operating at student level and their assumed interrelation (from Creemers & Kyriakides, 2008, p. 94).

The socio-cultural and economical background variables contain SES, ethnic background, and gender. From the psychological perspective, the model adds aptitude, motivation, and expectations. Concerning motivation, the main focus is put on its conception as perseverance and subject-related motivation (de Jong et al., 2004; Kyriakides, 2005).

A new asset of the model is the addition of students' personal characteristics, such as *personality traits* and *thinking style*. They are seen as important variables that teachers need to take into account in order to be able to differentiate the teaching practice accordingly, and thus to respond to the different needs of the students to improve effectiveness. Creemers and Kyriakides (2008) perceived some of the student-level factors as being more stable, while others, such as motivation, are more susceptible to interventions, and consequently might show reciprocal effects with students' achievement gains.

The empirical validity of the dynamic model was tested in several studies. Creemers and Kyriakides (2010), for example, showed that school factors can be classified according to the five dimensions of the dynamic model and that most of the factors are associated with different learning outcomes. Several studies and meta-analyses gave empirical evidence for the association of the teacher factors identified in the model with student outcomes (Blömeke et al., 2016; Kyriakides & Creemers, 2009; Scheerens, Luyten, Steen, & Luyten-de Thouars, 2007).

Based on a larger review of school effectiveness research, Scheerens (2013) studied the extent to which school effectiveness research studies were based on theoretical constructs. He found that out of 109 studies, only 11 made reference to "specific broader conceptual principles" (Scheerens, 2013, p. 1) and that out of those, 5 were based on Creemers' comprehensive model and the dynamic model, while the rest referred to different established theories.

5 PROBLEM SETTING AND RESEARCH QUESTIONS

5.1 Problem Setting

An examination of the results from TIMSS 2015 reveals that in fourth grade, all GCC countries perform in the lowest quartile of the achievement scale for mathematics. Additionally, the region exhibits quite a large achievement gap (of about a standard deviation) between the lowest achieving country (Kuwait) and the highest achieving countries (Bahrain and the United Arab Emirates). A similar situation can be seen for the science achievement on a primary level. Here again, all GCC countries are positioned at the lower end of the ranking scale – with Bahrain being the highest performing country of the region and Kuwait again at the lower end, resulting in a variance of even more than one standard deviation in the region. More information about the mathematics and science achievement distribution in the last two cycles of TIMSS can be found in Table 2-6.

TIMSS results show, that in terms of achievement, even relatively poor TIMSS participants like Chile (13,576 USD) or Turkey (10,742 USD) outperform all of the GCC countries in both subjects – albeit even given the fact that the GCC countries belong to the wealthiest countries in the world, ranging from rank 1 (Qatar with 127,660 USD) to rank 21 (Oman with 46,698 USD), measured by the Gross Domestic Product (GDP) per capita in 2016 (International Monetary Fund, 2017). More details can be found in Table 2-1.

Next to a low achievement level in mathematics and science, the region is still characterized by high disparities between different groups of the population, especially in terms of gender and nationality status. While gender gaps in in many other TIMSS countries are predominantly in favor of boys, gaps have been narrowed over the past 20 years, and in more than half of the countries the gender difference in grade four now is insignificant in the recent cycle of TIMSS 2015 as can be derived from Mullis, Martin et al. (2016, exhibit 1.10). GCC countries, however, are still among the countries with the highest gender gaps – interestingly, in favor of girls. For mathematics, the highest differences occur for Saudi Arabia, with 43 points, and the lowest for the United Arab Emirates and Qatar with an insignificant difference of 3 points. For grade four science, however, the Gulf countries show the highest gender differences of all 50 countries and seven benchmarking participants on the TIMSS scale, ranging from 79 points in Saudi Arabia to 14 points in the United Arab Emirates, as can be seen in exhibit 1.10 (Martin, Mullis, Foy et al., 2016, exhibit 1.10).

While in the last decades financial resources from oil and gas revenues were used within a short period of time to improve the quantitative dimensions of schooling (such as enrollment, student-teacher ratio, resources for learning, etc.), it seems that qualitative dimensions still lag behind in terms of achievement outcomes, but also in terms of equity. Compared with the previous cycle of TIMSS (TIMSS 2011), four out of the six GCC countries partially show huge improvements, reaching up to 40 points difference in mathematics and 54 points in science in Oman (see Table 2-6 for more details). On the other hand, the achievement dropped in Kuwait for science and even more in Saudi Arabia for both subjects – where the difference amounted to 27 score points in mathematics and 39 score points in science during the last four-year period.

Especially in primary education, competencies in basic skills such as mathematics, reading, or science, are important factors for further development of an educational system. Consequently, national governments in the region are concerned about the quality of their education, and the topic in general is a strong focus of international organizations such as the United Nations or the World Bank. The United Nations developed a set of educational targets, the Sustainable Development Goals (SDGs) that should be achieved by 2030. SDG 4 targets the field of education in this way: "Ensure inclusive and equitable quality education and promote lifelong learning opportunity for all" (United Nations, 2015, p. 17). The following targets stipulated by the UN for SDG 4 are partly touched on by this study:

- Target 4.1: By 2030 all girls and boys should complete equitable quality education leading to relevant and effective learning outcomes
- Target 4.5: Gender disparities should be eliminated and access given to persons with disabilities, indigenous people and children in vulnerable situations
- Target 4.A: Facilities should be build that provide safe, non-violent, inclusive, and effective learning environments
- Target 4.C: Supply of qualified teachers should be substantially increased

As explained in more detail in section 3.2, which deals with educational quality, student achievement can serve as an interpretation of effectiveness in terms of quality of education and can be measured by subject-specific tests in subjects such as in mathematics or science (Scheerens & Bosker, 1997). It is argued here that with the competition of the global market, and, consequently, with the evolution of the Arab GCC countries from a more traditional schooling system towards a more "world-wide modern kind of school system" (Adick, 1992, p. 244), the social functions of the educational systems in the Gulf and in the West converge more and more. The focus here will be set on the schools' function to "generate" qualified

members of the society to make the economy competitive, and to provide better chances for good work conditions and high salaries for the individual (Fend, 2006, p. 51) which is an important precondition for the region in terms of their current developments towards a knowledge society. The term knowledge society here is used as defined by the United Nations Development Programme (UNDP): "A knowledge-based society is one where knowledge diffusion, production and application become the organising principle in all aspects of human activity: culture, society, the economy, politics, and private life" (UNDP, 2003, p. 2). The convergence of schooling systems is also described by Kirk (2011, p. 41), who stated that "there is an ongoing and prevalent perception in the region that Western educational credentials are seen as the key to entry into the globalized knowledge economy, and lead to higher status and reward, both individually and for nations." Nevertheless, it should be kept in mind that due to the historic and cultural context of the region, other important goals of schooling – such as the legitimization of the respective system of government (Fend, 2006) – might still play a more prominent role than in other regions of the world (see corresponding paragraphs in section 3.2 for a discussion on the different objectives of schooling). According to the authors of the Arab Human Development Report 2003 (UNDP, 2003, p. 53), the function of education to legitimize political systems affects curriculum and instruction more in the areas of social sciences and humanities, subjects which they describe "generally indulge in both self-praise and blame of others, with the aim of instilling loyalty, obedience and support for the regime in power." Consequently, when it comes to mathematics and science learning, the investigation of the functioning of the educational systems of the Gulf from a somewhat Western educational effectiveness perspective should also remain valid. The Gulf region has been chosen for this research as, according to the literature review of the author, no comparative educational effectiveness research specifically targeting the Gulf region has been conducted so far. The author was interested in investigating the extent to which educational effectiveness research concepts, which were predominantly developed in the Western hemisphere, also would work in this culturally distinct region. Furthermore, the Gulf States appear to form a historically and culturally homogeneous region, and as such, are expected to face similar conditions and challenges in terms of their educational contexts. The countries of the GCC show several common characteristics, as all are deeply rooted in Arab culture and history. All have a monarchy as their form of government, and are mostly conservative and tribal in nature, with strong family and tribal ties. Additionally, for most of them, oil and gas are at the center stage of their politics.

Analyses of achievement differences between GCC members and major subgroups of their populations require both comparable data and a well-developed framework. During the past half-

century, researchers have studied variations in student achievement in various educational systems based on educational and non-educational factors, finally leading to a development of theories and models to explain these differences, and hence the effectiveness of schools or educational systems as a whole. However, such studies were usually conducted on a national level or below, and often only included a rather small and non-representative sample of schools; results, therefore, often did not allow for generalizations on a national level or even regional level. Moreover, results in the educational effectiveness area mainly stem from studies conducted in Western countries or in East Asia. The availability of comparable large-scale assessment data for all GCC countries allows for the exploration of educational effectiveness factors in this vastly different region. However, using cross-sectional large-scale assessment data for EER poses certain challenges and limitations, which are further discussed in section 8.2.

5.2 Rationale for the Study

While definitions of educational quality differ depending on the interests of the stakeholders involved (see section 3.2) and the objectives an educational system is supposed to fulfill, there is a certain agreement among researchers and policymakers alike that in modern school systems, educational quality can be regarded as the discrepancy between a desired outcome versus certain status or input condition, and that this difference can be measured on certain evaluation criteria. Hence, to assess the educational quality of their systems, and to monitor the adequacy of education, policymakers in the Gulf region, like in Western countries, also mainly rely on the educational outcomes of schooling. During the last couple of years, awareness of the importance of education in this region has grown; consequently, there is also a rising interest in monitoring the implementation of educational reforms by benchmarking the national educational systems against other educational system worldwide. As a result, all six GCC countries participated in the most recent cycle of TIMSS (TIMSS 2015), while Qatar and the United Arab Emirates also participated in recent cycles of PISA (OECD, 2016b). Thus, a more regional approach of analyzing educational effectiveness factors in the Arab Region becomes possible with the data at hand.

Highly standardized international large scale assessments provide a good opportunity to “dig deeper,” as discussed above; secondary analysis of the data may assist teachers, principals, and policymakers in identifying key factors that are important for learning in the region. Gained insight into educational practices and differences among the countries of the region, as well as among major population groups, can aid in achieving a better understanding of the learning

environment in these six countries, and hence might help to improve achievement in the region and bridge achievement gaps in the populations.

Thus far, secondary analyses of TIMSS data with a certain focus on EER have focused mainly on comparing similarities or differences between European countries (Bos & Kuiper, 1999), between Asian countries (Leung, 2002), or between the USA and Asian (outside the area of the Gulf Cooperation Council) or European countries (O'Dwyer, 2005), or Australia (Lamb & Fullarton, 2001). Other studies, such as those conducted by Kyriakides (2006) or by Martin and Mullis (2013), included all participating countries of a certain study cycle, but did not focus specifically on the conditions of the Gulf area. The current study investigates achievement differences in a historically and culturally rather homogeneous set of countries, which are characterized by a combination of specific characteristics – such as great wealth, while still exhibiting low achievement levels and high achievement disparities among certain subgroups of the populations. The research project is conducted from the perspective of educational effectiveness, using a research framework that includes most recent research findings while concurrently endeavoring to account for contextual realities in the region under consideration, and, as much as possible, bearing the limitations of the available large-scale assessment data in mind. In this sense, the current study aims to contribute in enhancing the consistency and validity of EER concepts and theories which were mainly developed and empirically validated in the Western Hemisphere.

5.3 Aims of the Study and Research Questions

The purpose of this study is to explore the achievement differences of primary school students in the GCC countries concerning mathematics and science from the perspective of an educational effectiveness framework. Achievement differences shall be investigated by means of secondary analyses of data from TIMSS 2015, the most recent cycle of the IEA international large-scale assessment. This investigation involves the following steps:

- To create a framework, and subsequently a model of educational effectiveness, suitable for the region and for the data at hand
- To identify factors likely to influence mathematics and science achievement, taken from the TIMSS 2015 background questionnaires on different levels of education (student, class, school), according to the framework developed

- Based on solid theoretical concepts, to obtain, via clear disentanglement between home background and school learning environments, a better understanding of malleable factors on course and school levels
- To provide interpretations for the variation in learning outcomes, based on the operation of the educational effectiveness-enhancing factors identified in the region

A better understanding of the operation of the effectiveness-enhancing factors in the region should help to design appropriate policy recommendations and interventions, and hence lead to improvements in the learning environment for students in the region. Further, applying educational effectiveness concepts in a region culturally very different from the Western Hemisphere can add empirical support for a generalization of educational effectiveness concepts.

Two main research questions can be derived from the discussion of the previous sections:

Research Question 1: To what extent does TIMSS 2015 reflect essential factors in terms of educational effectiveness research?

To answer the first main research question, a theoretical educational effectiveness framework must firstly be developed by analyzing existing frameworks, assuring that the special conditions of the region under consideration are incorporated. Concurrently, the framework needs to take limitations of the available large-scale assessment data into consideration. In a second step, the TIMSS questionnaire data need to be examined in terms of the developed framework. These steps allow answering the following sub question:

- How should an EER framework that takes into account recent findings of educational effectiveness, the special educational conditions in the Gulf area, and the restrictions imposed by using cross-sectional large-scale assessment data be constructed?
- Can TIMSS 2015 grade four student, teacher, and school questionnaire data be used to give empirical support for the developed educational effectiveness framework in the GCC countries, using mathematics and science achievement as outcome variables?

Research Question 2: According to the framework specified, which educational factors are most effective from the perspective of EER with regard to learning outcomes on primary level in the GCC countries?

Because of the assumed commonalities in the region, the research questions should be answered by using a regional approach, and investigate the extent to which the different factors identified

as important for the region can be regarded as either generic, or only specifically relevant for explaining the performance differences in a subset of countries.

Because an emphasis is placed on an examination of malleable factors related to the school learning environment of the students, it also is necessary (to the greatest extent possible) to disentangle home background and school-related factors. The construction of an indicator for the home background, which is suitable for the region and can be used as a controlling factor when investigating the main research question, is therefore necessary. The starting point here is recent research related to the development of indicators of SES in large-scale assessments (Brese & Mirazchiyski, 2013; Caro & Cortés, 2012; Ehmke & Siegle, 2005; Sirin, 2005). After developing a suitable background indicator for the region, the following sub-questions should be answered:

- How do the different educational effectiveness factors identified associate with students' mathematics and science achievement in the different GCC countries, when controlling for the home background?
- Do effectiveness factors operate in a similar way in the region for both subjects, and can a regional pattern be identified?
- To what extent do the educational effectiveness factors identified for the region explain differences between the GCC countries, after controlling for the student background?

An additional aspect that should contribute to the body of EER is the question of the extent to which the findings of the study can give certain empirical support for the generalizability of theoretical constructs related to educational effectiveness.

6 CONCEPTUAL FRAMEWORK OF THE STUDY

6.1 Introduction

In this chapter, formulation of the framework for the current research project is detailed. The framework to be used must fulfill several important conditions: it should be based on empirically-validated research but also consult recent research findings, and, if applicable, integrate them. The model created from the framework should be parsimonious, but at the same time allow for the differentiation of the most important elements and factors of the model. In addition, it should be possible to validate the final model, to the greatest extent possible, using data from cross-sectional large-scale assessments, the only available comparable data source for the region under consideration. While the framework per se is kept generic, a specific emphasis was placed on the special conditions in the GCC countries, and consequently factors assumed to be important in the region must be emphasized.

6.2 Developing the Framework

The conceptual framework for this research project is mainly based on the models described in chapter 4, namely Creemers' model (Creemers, 1994), the dynamic model (Creemers & Kyriakides, 2008), and partly on Scheerens' model (Scheerens, 1992). Additionally, it consults more recent research in the area of instructional effectiveness (Helmke, 2009; Klieme & Baumert, 2001; Nilsen, Gustafsson, & Blömeke, 2016; Seidel & Steen, 2005). The foundation for the new framework will be based on Creemers' integrated model of educational effectiveness. His model contains all essential features of an integrated effectiveness model, as discussed in section 4.2.2: it is based on input-process-output functions, although a strong focus is set on the process and output dimensions. It combines the organizational orientation with the teaching learning orientation of educational effectiveness, and it recognizes that influences on student achievement are multilevel. In contrast to models oriented more towards the organizational structure of educational systems (such as Scheerens' model), Creemers placed special emphasis on the classroom processes of teaching and learning, and considered higher levels of education rather as preconditions for effective learning. This is in agreement with more recent research findings that regard the factors at the teaching and learning level as the dominant effectiveness factors especially in primary schools (Hill & Rowe, 1996; Kyriakides & Creemers, 2008, p. 18). As stated in section 4.2.2.5, the main indicators of Creemers' model are well-researched and have been empirically validated in several studies. Additionally, the model is parsimonious,

and is suitable for application with the available cross-sectional data at hand. While the dynamic model is further developed, and also takes into account the dynamics of educational effectiveness, it is also more complex, and is not strictly suitable for the cross-sectional TIMSS data at hand, but rather would require longitudinal data and qualitative classroom observations. In consequence, the new framework will stem from Creemers' approach, which is more suitable for the data at hand, and adjust his model to reflect new research findings and also accommodate the special conditions in the Gulf State area under consideration.

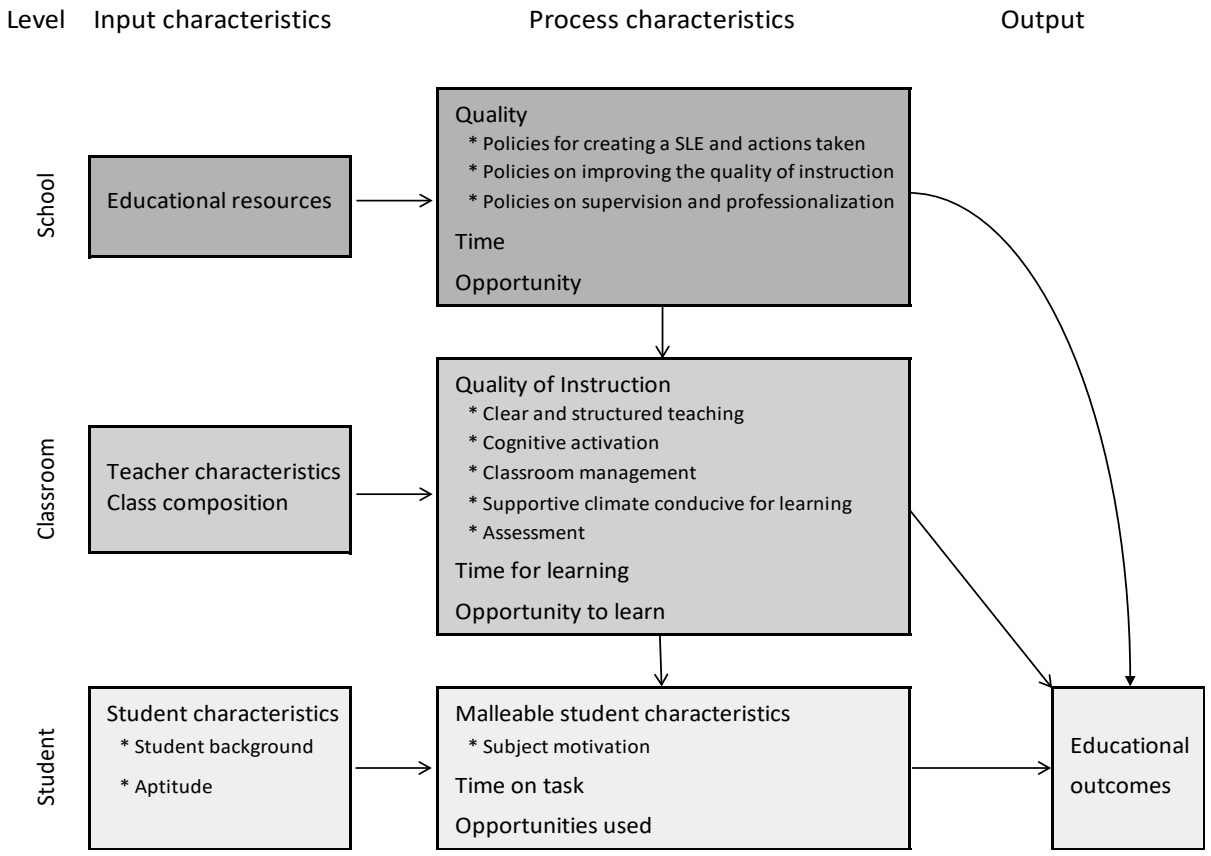


Figure 6-1: Proposed model of educational effectiveness – Summary

The new model differs from Creemers' approach in three important main aspects: firstly, the addition of an *input* dimension; secondly, the revision of sub-components related to the *quality* category on classroom level (and, to a lesser extent, also on school level); and thirdly, the inclusion of recent research findings related to the elements of *instructional quality* and *climate*. The *input* dimension was added to allow for the inclusion of important resource variables, teacher background variables, and student composition characteristics that are assumed to play an important role, given the limitations of the available data and the special conditions in the Gulf area. The construct describing the *quality of instruction* has been revised and partly relabled, and new approaches of teaching rooted in constructivism have been added to the model. In addition, climate variables have been given a stronger focus in the new model.

Creemers' and the dynamic model also regard policies as important. These target the main factors of *quality*, *time*, and *opportunity* on a context level; hence, the models include them in a separate educational level. While context-specific influences are recognized here, the current analyses will be restricted to school, class, and home levels, which are closest to the classroom level where teaching and learning are mainly supposed to take place. Besides, the necessary data for the context level, which in is TIMSS collected through a curriculum questionnaire and the TIMSS encyclopedia (see TIMSS & PIRLS International Study Center, Boston College, 2016a for more information), is only partly available and less comparable across countries. A graphical overview of the new model is given in Figure 6-1.

Neither Creemers' nor the dynamic model include resource or other input variables, as they are not assumed to have a direct effect on teaching and learning. However, as described in section 3.3.6.5, monetary and physical resources might (at least to a certain threshold) act as an important precondition for effective instruction, especially in developing countries. As differently-distributed resources might affect the *opportunity to learn* for certain groups of students also in the Gulf area, it was decided to additionally include in the current model the availability of important educational resources on school level. The dynamic model (Creemers & Kyriakides, 2008, p. 132) also regards provision with learning resources as an important aspect of educational effectiveness, albeit the classification here is different: the dynamic model integrates the resource aspects into the school policies for creating a school learning environment.

Moreover, there is research evidence suggesting that certain teacher background factors are related to student outcomes (see section 3.3.5.1); thus, teacher qualifications and other teacher characteristics are usually regarded as important input factors in the consulted organizational input-process-output models and models about teaching instruction (Baumert et al., 2010; Helmke, 2009; Nilsen et al., 2016; Scheerens, 1992; Seidel & Steen, 2005).

Creemers' model and the dynamic model, however, only focus on observable teacher behavior which directly influence student learning (Creemers & Kyriakides, 2008, p. 117); consequently, *input* characteristics such as the teacher background, which may have an indirect impact on student learning, are not considered in these models. While the author could not find any justification for the restriction to observable teacher behavior in these models, the rationale is likely to be that teacher background variables ultimately can be assumed either to influence or transform into a certain observable teacher behavior in the classroom, which then in turn influences student learning. While from a theoretic perspective the researcher agrees, it is hypothesized for the current research, that the available cross-sectional data based on teachers' self-ratings

and student ratings will not suffice to tap into the whole range of teacher behaviors related to the *quality of instruction* construct. Teacher background variables, such as teacher qualifications and teacher characteristics, are therefore seen as important additional input characteristics that should be included in effectiveness models if only quantitative questionnaire data is available. Moreover, most of the input variables are at least partly malleable, and in this way highly relevant for practitioners and policy-makers interested in improving educational systems. Teacher qualification especially can be seen as an important variable in the GCC countries, as the teaching force in this area even today consists to a large extent of expatriate male teachers with heterogeneous cultural and professional backgrounds.

Additionally, it was decided to include not only the students' individual background on student level, but also to include the student composition as an input factor on class level. As was demonstrated in section 3.4.6, the student composition, especially in terms of average achievement, students' SES, and related to the proportion of girls, was often found to be highly related to achievement beyond the individual student's background – and is assumed to strongly influence the learning environment in the class. Student composition, therefore, is regarded as an important factor for the current framework. Because of the large differences in educational conditions and in achievement levels between nationals and non-nationals in the Gulf region, composition in terms of nationality will be modelled as well. While the school composition in effectiveness models often is not directly regarded, or is classified as a context variable (as in the models of Scheerens, 1992, and Helmke, 2009), student intake will be treated as an important input characteristic for processes at the classroom level.

The second main adaptation to Creemers' model relates to the *quality of instruction* construct on classroom level. In more recent research, and similarly to the dynamic model (Creemers & Kyriakides, 2008), this construct has been further elaborated by integrating more modern constructivist approaches of teaching. A more detailed description of the constructivist approaches of teaching incorporated into the current framework can be found in section 3.3.5.2. While the current framework tries to keep the model as parsimonious as possible, it simultaneously attempts to classify the most important related factors in a meaningful way, and to distinguish between their most important dimensions. For the *quality of instruction* construct, the model draws on a categorization developed by Klieme and Baumert (2001). Based on evaluation of German data from the TIMSS Video study and the subsequent first PISA cycle in the year 2000, Klieme and Baumert (2001, p. 51) defined "three global dimensions of classroom process quality" which they termed *classroom management* [Unterrichts- und Klassenführung],

supportive climate [Schülerorientierung], and *cognitive activation* [Kognitive Aktivierung]. Independently, Kane and Cantrell (2012) identified quite similar dimensions based on their class-room-observation studies carried out in the USA. However, the definition of Klieme and Baumert (2001) for *supportive climate* also contains factors related to a *clear and structured instruction*, an important dimension which is usually handled separately in the literature, and is also kept as a separate factor in the dynamic model. This research project will consider elements of a *clear and structured instruction* to be a fourth dimension of *instructional quality*. This dimension will also consider important elements of the direct teaching approach that are emphasized in research and the integrated effectiveness models, namely *questioning techniques* and *practice* (refer to section 3.3.5.2 for more details). A similar approach was used by Blömeke et al. (2016) in their study on teacher quality and instructional quality based on TIMSS 2011 data. In this way, the final construct also allows for linkages with the eight main instructional factors of teaching described in the dynamic model: *Orientation* and *Teaching/Modeling* of the dynamic model will be generally summarized here under *cognitive activation*, while the three more "traditional" teaching approaches of the dynamic model – that is, *structuring*, *questioning*, and *application* – mainly correspond to the *clear and structured instruction* dimension in the proposed framework. In the Gulf area, a strong emphasis on rote learning and traditional teaching approaches (BouJaoude & Dagher, 2009, p. 3; Ridge, 2014, p. 39) still persists today, which also calls for keeping the traditional teaching approaches separate. This procedure also allows for the observation of differences between *clear and structured instruction* and the development of higher-order thinking skills and problem-solving included in *cognitive activation*. In addition to the core dimensions of instructional quality, the model will also include the factor *classroom assessment* as a fifth dimension. As could be demonstrated in section 3.3.5.2, monitoring student progress and formative assessments wherein the results are used to give constructive feedback to students constitute essential factors for effective instruction. This factor, which is also contained in Creemers' model (termed *evaluation*), as well as in the dynamic model, was always considered an important dimension in its own right; this perspective will be adopted here as well.

Additionally, in recent research, the importance of creating an *environment conducive for learning* has earned higher recognition as being important for educational effectiveness, especially on class-room level. Consequently, this factor is more comprehensively included in the dynamic model compared to Creemers' approach, where mainly *high expectations* would qualify for this dimension (see also section 3.3.6.2 for more details about climate factors). As the classroom climate depends on different preconditions and interactions between different actors in the

school environment, neither a clear definition nor a localization of this factor is straight forward in educational effectiveness models; hence, in this regard, authors follow different approaches. Helmke (2009, p. 73) treated the school and classroom climates as context variables, while Creemers' and the dynamic model (who don't use a separate context dimension) integrated them into the process quality dimension. In this context, the current framework will follow the approach of the dynamic model and focus on classroom level, specifically on the teacher's contribution to establish a productive learning environment (Creemers & Kyriakides, 2008, p. 113). Additionally, the element *policies for creating a school learning environment*, taken from the dynamic model, is included on school level to reflect this element of the school climate more prominently and comprehensively than was the case for Creemers' model – as it "is seen as the most important predictor of school effectiveness, since learning is the key function of a school" (Creemers & Kyriakides, 2008, pp. 131–132). Indeed, in nearly all studies related to educational effectiveness, this factor emerged as an important predictor for student outcomes.

In addition to the main adaptations described above, the following further changes to Creemers' model have been implemented: on school level, certain sub-items have been slightly renamed to be more concise, mainly by following the convention of the dynamic model. So as to address recent aspects of public discourse, and also in line with the dynamic model, *policies related to support students with extra learning needs* have been added.

At the classroom level, the items related to *grouping procedures* were removed. In Creemers' model, grouping was regarded as an important factor related to the quality of instruction. While grouping may influence the opportunity to learn (see section 3.3.3), research results do not give empirical support for grouping procedures to work as a general factor, and the benefit of producing different learning opportunities to different student groups via grouping/ tracking is considered among scholars to be controversial. Grouping procedures as a separate factor, therefore, will not be kept in the framework, albeit it is acknowledged that teachers also might use grouping procedures to achieve an effective instruction, for example to balance the amount of *opportunity to learn* of different student groups to enhance equity.

At the student level, the elements of Creemers' model were kept, but the framework now distinguishes between more stable elements (*aptitude* and *social background*), which were categorized as input factors, and a (partly) malleable element (the *subject-motivation*) which is influenced by school processes, background, and educational outcomes, and is listed under the *process* category. *Motivation* in the Gulf area is an important factor as high differences between the sexes and also between foreign and national students can be discerned. It is acknowledged

here that further factors, such as *student thinking styles* and other personality traits, might be of importance. As for the current research, the model will rather focus on malleable factors on school level, keeping Creemers' core elements in the interest of parsimony.

Finally, the inclusion of Creemers' (1994) formal criteria was reconsidered. His full model makes some tentative statements about the joint impact of the effectiveness factors by introducing the formal principles of *consistency*, *cohesion*, *constancy*, and *control*. The underlying assumption here is that educational effectiveness can only be assured if the different contributing factors work in line with each other in a *consistent* approach, and over a longer period of time (*constancy*). The school staff also needs to act according to agreed-upon school policies, which creates *cohesion* among them (Kyriakides & Creemers, 2008, p. 45). Moreover, outcomes, teacher behavior, but also the school climate and the educational policies themselves need to be evaluated; if necessary, corrective measures need to be applied, which calls for *control*. While the formal principles are acknowledged as essential for an effective instruction, they are difficult to see and measure directly, especially if only cross-sectional questionnaire data is at hand. *Cohesion*, and to a lesser extent *consistency,* were incorporated into the details for the rules and agreement section on school level. From a theoretical perspective, the author agrees with Creemers and Kyriakides (2008) that *control* is a separate evaluation element that should be connected with the school and class level via feedback loops. The author also agrees that a theoretical review over time would address the *constancy* principle. However, as these principles cannot be measured with the data at hand, they will not be included here in the study-specific framework.

Even though elements are depicted in Figure 6-1 as being clearly distinguishable from each other, they interact and are interrelated in ways that Creemers' partly tried to describe using the different formal criteria. While it is acknowledged that there are, by far, more connections between the different elements and levels than depicted; and, for example, some influences from the lower educational levels to the higher ones can also be postulated; effort was made to parsimoniously use arrows by focusing only on the main assumed interrelations.

As for the current research, students' motivation is considered to be a *predictor* for achievement, and thus the arrow is only pointing from motivation to outcomes. Nevertheless, it is recognized that there is theoretical and empirical evidence for an interrelation between both variables, and that for other research objectives, motivation also could be regarded as an outcome variable.

Table 6-1: Details of the factors for the proposed model of educational effectiveness

Level	Factors		Details of factors
SCHOOL	Input	Educational resources	Equipment and material for mathematics and science instruction: computers and software, library resources, laboratories and science equipments for experiments
	Quality	School learning environment	Policies and actions related to student behavior (orderly and safe school atmosphere), Values in favor of learning (school culture inducing effectiveness)
			High expectations of teachers and students, Emphasis on academic outcomes, Shared vision, cohesion and collaboration among staff
		Quality of instruction	Rules and agreements about classroom instruction, Professional development of staff Policies on supervision, Monitoring and evaluation system
	Time		Management of teaching time, Rules and regulations related to absenteeism of teachers and students, Homework regulations, Regulations about lesson schedule and time table
	Opportunity		Policies and regulations related to the content of the curriculum, the teaching aims, and the curricular material being used Rules and regulations on how to implement the curriculum Policies related to extra-curricular activities such as field trips Policies related to the support of students with extra learning needs
CLASS	Input	Teacher characteristics and qualifications	Pedagogical content knowledge, Teacher education, Job experience, Professional development, Major area of study, Gender
		Student composition	Prior achievement, Gender, Socio-economic status
	Quality of instruction	Clear & structured teaching	Structured lessons, Clear explanations, Reinforcing of major points, Summarizing the content, Questioning and feedback, Ample practice
		Cognitive activation	Provision of objectives for tasks & lessons, Engaging environment linked with daily life, Cooperative learning, Teaching of higher-order thinking skills & problem-solving, Helping students develop own strategies
		Classroom management	Organization of classroom environment to maximize engagement, Clear rules
		Supportive climate	High expectations, Emphasis on academic outcomes, Relationship between teachers and students, Attitudes towards teaching
		Assessment	Formative assessment to identify students' needs (and evaluate own practice)
	Time for learning		Instructional time assigned by the teacher, Homework
	Opportunity		Curriculum content taught
STUDENT	Student charact.	Student background	Socio-economic status, Ethnicity, Language, Gender, Parental involvement
		Aptitude	Prior achievement/ knowledge
		Subject motivation	Achievement motivation, attitudes towards learning, values
	Time on task		Time spent on homework, Private tutoring Extra-curricular activities related to mathematics/ science
	Opportunities used		Homework, Tutoring, Absenteeism, Attention

Another major focus of recent effectiveness research is the focus on the dynamics of teaching and learning over time, and depending on the current situation of a school. The dynamic model also recognizes that differential effectiveness might occur, meaning that different effectiveness factors might work differently for different groups of students. The current base model presented here was developed for the cross-sectional data at hand, and will focus on the more generic aspects of educational effectiveness. However, the model can likewise be used to investigate important subgroups of the populations which often differ significantly in their achievement levels – such as according to SES, gender, or ethnic composition. Details and subcomponents for the different factors proposed in the model depicted in Figure 6-1 can be found in Table 6-1.

7 IEA LARGE-SCALE ASSESSMENT TIMSS 2015

7.1 Introduction

Since the 1960s, the International Association for the Evaluation of Educational Achievement (IEA), a non-governmental research organization, has conducted more than 30 international comparative assessments in different subjects, among them mathematics, science, languages, civic education, and computer literacy. Included in the objectives of these studies is to identify factors likely to be related to student learning, and to thus to help policymakers develop evidence-based policies to improve education.

One of the IEA's core studies is the Trends in International Mathematics and Science Study (TIMSS), a large-scale international comparative assessment which is conducted every four years since 1995 and focuses on mathematics and science achievement (Mullis & Martin, 2016). The TIMSS assessment is regularly administered on primary level at the fourth grade and on secondary level at the eighth grade (with the exception of TIMSS 1999, in which only eighth graders were assessed). Additionally, in TIMSS 1995 and under the acronym of TIMSS-Advanced in 2008 and 2015, students in their final year of secondary schooling were also assessed. During the first cycle of the study (TIMSS 1995), adjacent grades (grade three and grade seven) were also included in the assessment. With increased participation of developing and lower-achieving countries, an additional, less difficult version of the TIMSS fourth grade mathematics assessment was introduced – namely, TIMSS Numeracy. Both versions of the TIMSS assessment can be linked on the same scale, thus allowing for the comparison of mathematics achievement of students who took different versions of the test. The TIMSS Numeracy assessment was administered to a subset of students in two out of the six Gulf Cooperation Council States: Bahrain and Kuwait. Additional details regarding the TIMSS and the TIMSS Numeracy scaling can be found by consulting Foy and Yin (2016).

Overall, 57 countries and 7 benchmarking entities (regional jurisdictions of countries), with over 580.000 students, participated in TIMSS 2015, marking the sixth administration of the TIMSS assessment (Mullis, Martin et al., 2016). 50 countries and 7 benchmark entities administered the assessment in grade four, while 39 countries and 7 benchmark entities participated in grade eight. All six Gulf Cooperation Council countries participated in both grades of TIMSS 2015. An overview of their participation in the different TIMSS cycles can be found in Table 2-5.

TIMSS provides countries with various insights about their students' achievement in mathematics and science, and about their educational system in general. The cyclic nature of TIMSS' administration allows for the measurement of trends in educational achievement in both subjects. Additionally, comparison across countries, especially when performed on countries with similar educational contexts, might help to explain achievement differences and thus assist in the identification of effective educational practices. Finally, the four-year administration cycle allows for the measurement of achieved advancement between the fourth and eighth grade cohorts. Complementing the mathematics and science achievement tests are questionnaires administered to students selected for participation in TIMSS, as well as their teachers and school principals. Since 2015, a home questionnaire has also been administered to the parents of the assessed TIMSS students at the fourth grade level. These questionnaires are intended to capture contextual school, classroom, and home information to paint a more complete picture of mathematics and science learning in the participating countries.

The rest of this chapter provides a general overview of the TIMSS assessment, especially on the assessment design, the instruments, and on measures designed to ensure data quality and comparability.

7.2 Design and Framework of the TIMSS assessment

The TIMSS assessment uses a curriculum model, comprising three aspects as the major organizing concepts: the *intended* curriculum, the *implemente*d curriculum, and the *attained* curriculum (Mullis & Martin, 2016, p. 4). The *intende*d curriculum represents the mathematics and science topics students are expected to learn, as defined in the curricula of the participating countries. For this purpose, the assessment is evaluated by experts and matched to the national curricula. How well the curriculum of each country matches the final assessment can be identified in the so-called test curriculum matching analysis (TCMA; TIMSS & PIRLS International Study Center, Boston College, 2016b). A summary of the results for the GCC countries can be found in section 2.3. The *implemented* curriculum refers to the organization of the educational system in facilitating learning, and considers what is actually taught, how it is taught, and also looks at the characteristics of the teachers. The *attained* curriculum finally observes what students have learned and also examines their attitudes towards learning (Mullis & Martin, 2016, p. 4).

The TIMSS assessment framework is organized around content domains, which specify the subject matter to be assessed; and cognitive domains, which specify the thinking processes to

be assessed (Mullis & Martin, 2016). The final assessment framework for TIMSS 2015 covers three content domains for mathematics in fourth grade (i.e. number, geometric shapes and measures, data display) and four domains in eight grade (number, algebra, geometry, data and chance). Similarly, the science assessment covers three domains in fourth grade (life science, physical science, earth science) and four domains in eighth Grade (biology, chemistry, physics, earth science). For both subjects, three cognitive domains are assessed: knowing, applying, and reasoning – with the percentages of the higher-order thinking skills being lower for the primary level.

The *TIMSS 2015 Assessment Frameworks* (Mullis & Martin, 2016) describes the item development process and the assessment design. Test items are developed by the National Research Coordinators (NRCs) of participating countries in collaboration with subject matter experts. All items are piloted, field-tested, and thoroughly reviewed by panels and experts before being included in the final assessment. In order to cover all assessment domains in both subjects appropriately, for the 2015 cycle, 350 items altogether had to be developed for grade four, and 450 items for grade eight, amounting to around 8,5 hours and 10,5 hours of testing time, respectively (Mullis & Martin, 2016, p. 89). Items were grouped into blocks of around 10 to 12 items, which in turn were used to compile 14 different booklets using a spiral rotating design. Each student was assigned one test booklet, consisting of two testing sessions of 36 minutes each for fourth graders and 45 minutes each for eighth graders. Each of the booklets contained mathematics and science blocks and half of the material included consisted of items from TIMSS 2011 in order to ensure the link between both assessments. At least half of the items were presented in multiple-choice format, while the remaining items were constructed-response items which in a later stage had to be manually scored.

7.3 Background Instruments

Next to the mathematics and science assessment instruments, a set of background questionnaires was administered to the selected students at fourth and eighth grade, their teachers, principals, and additionally to curriculum specialists of each participating country. Starting in 2015, a home questionnaire was also completed by the parents of the fourth grade students participating in TIMSS. The questionnaires collect policy-relevant information about the country's home and school contexts for teaching and learning.

Information about national and community contexts is gathered via a curriculum questionnaire, and through a description of the country's educational system in the TIMSS encyclopedia

(TIMSS & PIRLS International Study Center, Boston College, 2016a). These documents answer questions about the organization and structure of the educational system, the curricula, teacher education, and the monitoring of the curriculum implementation in different countries (TIMSS & PIRLS International Study Center, Boston College, 2016a, p. 62).

Each principal of sampled schools was administered a school questionnaire, which asked about the school context for learning. Topics to be covered related to school characteristics, school environment, school resources and instructional time, school climate, the role of the principal, and students' readiness to learn (TIMSS & PIRLS International Study Center, Boston College, 2016a, p. 97).

Each mathematics and science teacher of the selected students received a teacher questionnaire, which collected information about the classroom contexts for learning. Important topics comprised questions related to teacher preparation, the content taught, classroom instructional resources and time, instructional engagement, and assessment (TIMSS & PIRLS International Study Center, Boston College, 2016a, p. 97).

Every student who participated in the assessment also received a student background questionnaire to be completed, in addition to his or her test booklet. The questionnaire was designed to gather information about general student characteristics and attitudes towards learning. The questions addressed student readiness to learn, motivational aspects, students' self-concept, and general student characteristics (TIMSS & PIRLS International Study Center, Boston College, 2016a, p. 96).

Finally, all parents or caregivers of participating students in fourth grade also received a questionnaire asking about the home context for learning, specifically about home resources for learning, early learning experiences, parental attitudes towards learning, as well as parental education and occupation (TIMSS & PIRLS International Study Center, Boston College, 2016a, pp. 96–97).

More information about the TIMSS 2015 questionnaires and the questions selected for the current research project can be found in APPENDIX A.

7.4 The TIMSS Sample

TIMSS is a curriculum-based assessment, and measures the achievement of students in their fourth and eighth year of formal schooling. The TIMSS sampling design was described by LaRoche, Joncas, and Foy (2016, chapter 3) as follows: the study uses a two-stage random

sample design, where a sample of schools is drawn in the first stage, and intact classes of students are randomly selected within schools. To obtain a nationally representative sample of the target population, the first step includes development of a national sampling plan, in a collaborative effort between National Research Coordinators (NCRs) and international sampling experts. This included assurances that exclusions on population group level, school level, and student level are kept to a minimum, and that all students in the target grades had a non-zero chance of being selected. Stratification procedures were used to improve the efficiency of the sample, and to ensure that specific groups of the population were represented proportionally in the sample. Within each stratum (with stratum in this context denoting a group of schools that share common characteristics), schools were ordered by their measure of size (MOS), which indicates the number of students in the target grade. A randomly initiated systematic sampling procedure assured that schools were selected proportionally to their size (PPS). In a second sampling step, classes were randomly selected within schools. While in most countries only one classroom was selected per school, some countries opted to choose two or more classrooms per school. In five GCC countries, a mixture of schools with only one class and schools with two classes were selected, whereas in Saudi Arabia only one class per school was consistently selected.

Table 7-1: TIMSS 2015 sample sizes in the GCC countries (own calculations)

Country	Schools	Courses		Students	
		Mathematics	Science	Mathematics	Science
Bahrain	182	336	340	8575	4429
Kuwait	166	289	281	7296	3703
Oman	300	352	353	9105	9105
Qatar	211	240	227	5194	5194
Saudi Arabia	189	189	188	4337	4337
United Arab Emirates	558	812	787	21177	21177
Gulf Average	**268**	**370**	**363**	**9281**	**7991**

In most countries, about 150 schools and a student sample of around 4000 students were needed to obtain a good representation of the student population and to reach the TIMSS precision requirements in terms of standard errors, which for the country's mean achievement should not exceed .035 standard deviations (LaRoche et al., 2016, 3.9). Replacement schools from the same stratum, and hence with similar characteristics, were selected if the originally sampled school refused to participate. No replacements were drawn for classes or students that did not participate in the study, since this was assumed to introduce bias to the sample. A minimum requirement of 75% for combined school, classroom, and student participation, as well as rigorous standards regarding acceptable levels of non-response, minimized the potential for non-

response bias (LaRoche et al., 2016, 3.10). The sample sizes in terms of schools and students are given in Table 7-1. The differences between the number of students having participated in the mathematics test and those in the science test in Bahrain and Kuwait is due to the participation of both countries (with a subset of students) in the TIMSS Numeracy assessment, which only administered the mathematics portion of the TIMSS assessment. Sampling information then was used to calculate weights at the school, class, and student level. Weights are the inverse of the sampling probability, and have to be taken into account when analyzing the data in order to assure correct representation of the different population subgroups, as well as to adjust in cases of non-response on the different levels.

7.5 The TIMSS Achievement Scores

The TIMSS assessment seeks to cover mathematics and science literacy domains as broadly as possible, and at the same time to measure trends across different cycles of the assessment. In order to address both objectives simultaneously, TIMSS administration calls for application of a rotated booklet design, wherein each individual student is only administered a subset of the available item pool. The TIMSS approach to deriving proficiency estimates from students' answer patterns to the test items (the scaling process) relies on item response theory (IRT). In contrast to classical test theory (CTT), IRT allows the comparison of answer patterns between different students, even when different samples of students answered different blocks of items.

CTT assumes that items of a test can measure, albeit associated with unsystematic measurement errors, a latent trait (the student's ability) – and that addition of the correctly answered results, in relation to the number of test items (the percentage correct), would be an appropriate measure of this ability. However, this approach is always dependent on the specific test administered, as well as on on the student population answering the test, and therefore does not allow for the comparison of student proficiencies obtained from different parts of a test. IRT, on the other hand, uses a probabilistic model which links the ability of a person to the difficulty of the test item, and possibly other item parameters. The probability that a student responds correctly to a given item is dependent on both his or her ability on the item difficulty (and other model parameters). As described in chapter 12 of the "Methods and Procedures in TIMSS 2015" (Martin, Mullis, & Hooper, 2016), TIMSS uses a logistic model, wherein the item difficulty is represented in terms of the log of odds of a person with a certain ability to achieve a certain response to the item. In consequence, person and item parameters can be represented together on a unidimensional scale – and item parameters will remain independent from the examinee tested, if model assumptions hold. Expressed differently, IRT provides item-independent person ability

measures, and person-free item difficulty measures. In this way, the achievement of students that are administered different test booklets can be compared to each other. In TIMSS IRT models, up to 3 item parameters are considered: in addition to the item difficulty, the item discrimination and a guessing parameters are also estimated. The item difficulty is defined as the point on the ability scale where the probability of obtaining a correct response is 50%. The discrimination parameter indicates how well an item differentiates the latent trait tested between examinees with different abilities. The guessing parameter is estimated for multiple-choice questions, and reflects the possibility of selecting the correct response by guessing alone. However, it should be noted IRT is also based on certain theoretical assumptions that must be met for correct model applications. In particular, IRT assumes that the underlying latent traits (in this context: mathematics and science literacy) are unidimensional, and that the conditional independence criteria are met. The latter point means that the probability for a correct response should depend only on the ability of the examinee, and is unaffected by other student characteristics, data collection conditions, and other items presented in the test (Martin, Mullis, & Hooper, 2016, 12.3).

In TIMSS 2015, depending on the item type and scoring procedures, the following IRT models were used: a three-parameter (3PL) model was applied for scaling the multiple choice items, a two-parameter model (2-PL) was utilized for constructed-response items that were scored as either correct or incorrect, and a generalized two-parameter partial credit model was used for extended responses with three different score levels (Martin, Mullis, & Hooper, 2016, 12.1-12.2). To obtain student proficiency scores, the data from all TIMSS 2015 countries were scaled together with the data from previous cycles, thereby allowing for the construction of a common scale and the comparison of achievement results for all cycles on a common metric. The obtained logit scores usually had a value range of -5 to +5. For easier interpretation they were converted to a mean score of 500 and a standard deviation of 100.

As stated above, TIMSS only administered a subset of items from the item pool to each student, allowing for broad coverage of the content and cognitive domains. While this approach facilitates more efficient population estimates, the design is less optimal for estimating the abilities of individual students, and requires taking into account the measurement error introduced with the matrix-sampling design. To address this issue, TIMSS adopts plausible value methodology, wherein five plausible values are drawn from an ability distribution of each student. To estimate the ability distributions more precisely for populations and subgroups, background questionnaire data were also included in the model in addition to the assessment responses, in a process called *conditioning*. From the obtained student ability distribution, five random values were

ultimately drawn. These values are called *plausible values,* and the variance between them indicates the magnitude of the measurement error stemming from the fact that students were only administered a subset of the available item pool. More information about the TIMSS scaling procedures can be found in chapter 12 of the "Methods and Procedures in TIMSS 2015" (Martin, Mullis, & Hooper, 2016).

7.6 TIMSS Data Quality Considerations

Once an assessment is used in secondary analyses, assurance that the data is of sufficiently high quality is necessary. There are three main criteria that are used to define the quality of a psychometric test: namely, *objectivity*, *reliability*, and *validity*.

7.6.1 Objectivity

An assessment is *objective* if the test measures do not depend on outside influences. This means that test administration and evaluation, and interpretation of the test, do not rely on the person administering, evaluating, or interpreting the test (Bühner, 2011, p. 58). TIMSS stipulated several quality assurance steps to ensure that the *objectivity* criteria were met. All persons entrusted with the test administration received detailed and clear instructions for operations and procedures via detailed survey operation manuals and additional training workshops. Furthermore, a quality control monitor program observed the testing sessions in different schools and countries in order to ensure comparability. *Objectivity,* related to the evaluation of constructed responses, was obtained by providing all scorers with a detailed scoring guide, training sessions, and training material. See Johansone (2016) for more detailed information regarding the survey operation procedures and related quality control steps.

7.6.2 Reliability

Reliability indicates the precision of a test, independent from its adequacy. In quantitative research, *reliability* "is essentially a synonym for dependability, consistency, and replicability over time, over instruments and over groups of respondents" (Cohen, Manion, & Morrison, 2007, p. 146). This implies that a reliable assessment is expected to deliver similar results in a similar context, with a similar group of students. There are different kinds of *reliability* that can be assessed for a test: the *internal consistency* of a test (measured for example by investigating in the split-half reliability); the *test-retest reliability* (measured as a correlation between a test administered at different points of time); and the *parallel-forms reliability* (which compares two tests which measure exactly the same construct). Additionally, the *inter-rater reliability* is

the degree of measurable agreement between two or more raters or scorers. In TIMSS, about half of the items are trend items from the previous assessment, ensuring a reliable measurement over time. Moreover, TIMSS uses a multitude of items (350 in grade four and 450 in grade eight) to measure the mathematics and science domains quite reliably. However, as each student is only administered a fraction of the item pool in order to prevent overburdening, IRT scaling needs to be applied to establish links between the different test forms. To improve reliability, TIMSS therefore applies conditioning – which includes additional information from the background questionnaires for the scaling process.

In addition, a double-scoring process was included to monitor the *inter-rater reliability* or *scoring reliability* of the constructed-response items. For this purpose, in each country, approximately 200 randomly-selected responses per constructed-response item were scored twice by two independent scorers. The *trend scoring reliability* and the *cross-country scoring* reliability were also documented (Johansone, 2016, 6.12).

7.6.3 Validity

In this context, *Validity* describes the extent to which a test measures what it aims to measure (Bühner, 2011, p. 61). Usually, three types of validity are differentiated: *content validity*, *criterion-related validity*, and *construct validity*. *Content validity* indicates how well the item measures the construct to be covered. A test needs to be created in such a way that it comprises a representative item pool, out of a "universe" of all items possible, to measure the intended construct. *Content validity* in TIMSS is achieved by enormous efforts put towards the item development procedures. Country representatives, international subject matter specialists, and panels review newly developed items in several cycles, align them with the assessment framework, and classify them according to the domains specified in the assessment framework. Items are then piloted and field-tested, and again reviewed after each administration. Great care is taken to ensure that items administered in different cultural contexts and to different groups of the population work equally well. *Criterion-related validity* describes the association between test achievement and other criteria that are expected to correlate with the test. It can be said that TIMSS partially endeavors to achieve *criterion-based validity* by cross-checking test results with certain background indicators, that, according to literature and other tests, are associated with math and science achievement – for example, students' SES. Questions related to the same topic asked in different questionnaires also allow for certain triangulation, adding to the *criterion-based validity*. *Construct validity* indicates how theoretically meaningful the test is, and is often based, in a more narrow sense, on a comparison of the relation of test constructs against

the same constructs from different tests (*convergent validity*) or by differentiation between discriminant tests (*discriminant validity*; Bühner, 2011, p. 64). Further information about the TIMSS item development process can be found in Mullis, Cotter, Fishbein, and Centurino (2016).

An additional aspect to be considered in a multi-national assessment is the equivalency of the testing material in different target languages. In TIMSS, as a first step, the assessment material and the questionnaires were translated into the language(s) of instruction by skilled translators, and then verified within each participating country. As described by Ebbs and Korsnakova (2016), a thorough process of translation verification was then performed on an international level, in collaboration with the participating countries, ensuring both that the meaning of the international item did not change when being translated to the target language, and that the difficulty-level of the language used was equal. This is also important in the GCC countries as all of them administered the test in English in certain private schools with English curricula, in addition to Arabic, which is the main language of instruction in the area.

8 RESEARCH DESIGN AND METHODS

8.1 Introduction

The objective of the current research project is to explain differences in mathematics and science achievement in the GCC countries through quantitative secondary analyses of the TIMSS 2015 data. As the researcher here acts as an observer of social reality, and the analyses transfer methodological procedures of the natural sciences to the area of social sciences, the current research can be seen as based in the *positivistic* paradigm (Cohen et al., 2007, p. 10). However, this approach may face certain limitations due to the "complexity of human nature and the elusive and intangible quality of social phenomena" (Cohen et al., 2007, p. 11) especially in the classroom and school environment. The fact that the analyses are based on previously collected and examined data, and that the current research project applies it to a different research framework than that for which the data was originally collected, introduces additional complexity. The subsequent section (8.2) discusses the use of secondary cross-sectional assessment data from the TIMSS assessment in EER. Section 8.3 describes the data analyses and data reduction procedures performed to obtain the final regional set of variables to be used for multilevel modeling. Section 8.4 specifically focuses on the creation of an indicator of the students' background, and section 8.5 outlines the multilevel analyses steps.

8.2 Using TIMSS for Educational Effectiveness Research

8.2.1 Secondary analysis of data

The current study is based on a secondary analysis of the TIMSS 2015 data. In secondary data analyses, data that was collected for a certain purpose is used to study a different problem (Herrnson, 1995, p. 452). Here, data from the multi-purpose large-scale assessment of TIMSS is used to obtain a more in-depth understanding of the educational effectiveness factors at play in the Gulf Cooperation Council countries. The use of data for secondary analyses comes with some advantages and disadvantages that should be acknowledged. Advantages naturally include economic and time efficiency aspects, as the primary data collection and processing steps already have been performed; another advantage may be the availability of more comprehensive data, of a higher quality than can be collected by the researcher him- or herself (Boslaugh, 2007). The TIMSS instrument development, sampling, data collection, data cleaning, and documentation has been demonstrated to show a high degree of validity and reliability, as well-

designed processes were applied in the development and administration of IEA studies (Gregory & Martin, 2001). When the primary data already has a high degree of validity and reliability, this, of course, will apply to the use of data in secondary analysis. That the sample was collected in such a way so as to be nationally representative, while only displaying small standard errors for populations and subpopulations, is of special importance here. Considering the combination of internationally comparable test instruments and standardized administration procedures (see section 7.6 for more information about quality considerations related to TIMSS), the data should allow for valid country comparisons in the Gulf area.

Nevertheless, secondary analyses also come with some disadvantages that should be taken into account. One major disadvantage is that the data were not collected to answer the specific research question in mind and consequently, particular information that would be needed might be missing. TIMSS is a multi-purpose comparative assessment; thus, the choice of variables that can be used to examine questions in the area of EER is restricted. However, it is argued here that as TIMSS is concerned with collecting policy-relevant data about the context for learning mathematics and science (Mullis & Martin, 2016, p. 4) and investigating students' opportunity to learn, the study should offer a broad array of data that can be used for EER. A second major disadvantage, according to Boslaugh (2007, p. 5), is the fact that analysts generally do not participate in the planning and execution of a study, and therefore cannot really pass judgment regarding certain issues that might negatively affect the data quality. In TIMSS however, extensive documentation on all steps undertaken for planning and executing the study exists (see Martin, Mullis, & Hooper, 2016 for more details); moreover, the researcher was also involved in project management and execution. Nevertheless, when performing secondary analyses, the author agrees with Cho (2010, p. 144) that it is of utmost importance to develop a strong theoretical framework up-front and examine the available data with in-depth analyses and interpretation within the specified framework.

8.2.2 Using large-scale assessment data for educational effectiveness research

As argued earlier, highly standardized international large-scale assessments such as TIMSS provide unique opportunities to dig deeper into the international dimension of educational effectiveness, and allow for the expansion of our knowledge outside education systems of the Western Hemisphere. Until now, large-scale assessment data have only scarcely been used for analyses related to educational effectiveness, especially in the Gulf area. The reason might be that such assessments in general require rather complex analysis techniques due to their study design and sampling approaches. Additionally, they are not specifically designed to identify

educational effectiveness factors (as discussed above), and their data is of a cross-sectional nature. All this makes them an easy target for criticism when being used for EER.

According to Teddlie and Reynolds (2000, p. 242) large-scale cross-sectional studies such as TIMSS, PIRLS, or PISA face, due to certain limitations of the study design, two major problems when used to address issues of educational effectiveness. First, all societies must be compared in their performance on the same skills. This raises questions, for example about the *cross-cultural validity* of the measures used (meaning that constructs and indicators between different cultures must be comparable), translation issues, etc. Besides, educational causes of differences must be isolated from other possible causes of country differences. As all of these studies use a cross-sectional design, meaning that (at least on individual level) they only obtain measurements at one point in time, there is therefore a need to disentangle the various non-educational background influences from the school environment factors under consideration. Certain authors argue that *value-added* effects (here understood as the effects comprising the school and teacher contributions to student learning) can only be measured using longitudinal data (for example, Lauder, Jamieson, & Wikely, 2003, p. 63). While these criticisms have merit, the fact that methodological improvements during the past decades have allowed for more efficient separation of the effects of home environment, teacher, and school on student outcomes and attitudes, which should allow at least some indication about the effectiveness factors that play a role in the school environments, should be considered. Support for this assumption is given by Lenkeit (2012), who compared a longitudinal growth model based on three time points with two cross-sectional status models (a contextual attainment model and a prior attainment model), using data from the longitudinal achievement study ELEMENT that was administered to fourth to sixth graders in the city of Berlin, Germany. She found that the effectiveness measure yielded by the growth model was accompanied by high uncertainty, while the two status models led to results that were more reliable. She came to the conclusion that her findings "legitimate the adjustment of achievement scores in cross-sectional studies to obtain measures of effectiveness" (Lenkeit, 2012, p. 54).

Besides, improved analysis techniques now allow researchers to take the multilevel structure of the data into account, and to control for non-educational influences on each level separately. However, to clearly separate out school educational factors, it is important to have a good and valid measure of the non-educational background influences; research in this field is therefore ongoing (see for example: Brese & Mirazchiyski, 2013; Caro, Sandoval-Hernández, & Lüdtke, 2014; May, 2006; Stubbe, 2003).

Criticisms concerning the use of large-scale assessment for effectiveness research in this study should be addressed to the greatest extent possible by:

- Comparing only countries which are characterized by a similar cultural background. The study will apply a regional approach and focus only on data from countries forming part of the Gulf Cooperation Council (GCC).
- Choosing mathematics achievement as an outcome measure for the general analyses steps, where cross-cultural definitions about the correct answers are expected to be more in agreement than in other subjects such as science (Teddlie & Reynolds, 2000, p. 242). It can also be assumed that mathematics is the area wherein students usually acquire most of their knowledge at school, as supposed to reading or language acquisition, which might be more influenced by their home environments (Mandeville & Anderson, 1987, p. 213). On the other hand, a single outcome variable would give a very limited view of the school's effectiveness (Teddlie & Reynolds, 2000, p. 116). In this study, therefore, science achievement as an additional outcome measure is used and results are compared between both cognitive subjects.
- Applying multilevel modeling approaches, as the educational process takes place across different levels.
- Attempting to separate the influences of the home context and the school environment as far as possible. For this purpose, the measurement of non-educational background factors should be based on existing theoretical concepts, and simultaneously should take into account the special conditions of the target culture (see section 8.4 for more details). The model applied here is classified in OECD's publication on the "Best practices to assess the value-added of schools" (OECD, 2008, p. 12) as belonging to the group of *contextualized attainment models* (CAM). The authors distinguish these models from real *value-added* modelling, which would require the availability of longitudinal data.
- Focusing the analyses on the early grades of schooling (primary level grade four), where the impact of schooling appears to be more pronounced than at the secondary level (Teddlie & Reynolds, 2000, p. 185). The researcher expects that at this level, a lower number of teachers are usually teaching the students over longer periods of time, and that the curriculum is usually less complex.

It is hence argued here that international data from large-scale assessments, under certain circumstances, can and should be used to expand the knowledge concerning "an area so far relatively undeveloped…" (Teddlie & Reynolds, 2000, p. 256). Nevertheless, it is important to mention that results must be interpreted with caution due to certain limitations, related to the availability of suited indicators, the cross-sectional structure of the data, and so forth.

8.3 Data Analysis

The data analyses were conducted in several steps, which are described in more detail in the following sections. After conducting some preliminary analyses related to disparities in terms of gender and nationality status (8.3.1), variables from the TIMSS background questionnaires were matched to the different factors of the proposed educational effectiveness framework (8.3.2), and school, teacher, and student data were matched (8.3.3). Subsequently, the obtained datasets were further explored (8.3.4) and procedures to handle missing data were applied (8.3.5). Afterwards, data reduction procedures such as principal component analyses (8.3.6.1) in combination with reliability analyses (8.3.6.2) were used to reduce the number of variables, and to identify the underlying constructs. Finally, correlation analyses in combination with theoretical considerations were used to select factors to be included in the subsequent multilevel analyses (8.3.6.3).

8.3.1 Preliminary analyses related to disparities in terms of gender and nationality status

In order to gain insight into important subgroup differences in terms of gender and nationality, some preliminary analyses were calculated using the International Database (IDB) Analyzer, a plug-in for SPSS and SAS developed by the IEA. The IDB Analyzer applies jackknife repeated replication (JRR) procedures and combines results obtained with each of the plausible values to obtain an appropriate standard error that takes the complex sample and test design of the TIMSS data into account (Foy, 2017, p. 12). The IDB Analyzer also assures the use of the correct sampling weights to account for unequal sampling probabilities and non-response adjustment.

8.3.2 Identifying variables related to the proposed framework

All TIMSS 2015 background questionnaire items were compared to the factors of the framework proposed in chapter 6 for each of the different levels, and categorized accordingly if a correspondence was found. In a second step, the categorization was then cross-checked against

research projects from scholars who applied similar approaches (specifically Cho, 2010; de Jong et al., 2004; Driessen & Sleegers, 2000; Kyriakides et al., 2000; Kyriakides, 2005, 2006; Nilsen et al., 2016; Reezigt, Guldemond, & Creemers, 1999). For some of the constructs, information was available from different background questionnaires and even partly from different educational levels. The following cases can be distinguished here: if exactly the same information was available from the student questionnaire and the home questionnaire, then the home questionnaire data was given preference, as it was assumed to be more reliable than those of the fourth grade students. This affected the *number of books at home* as well as the information about the *nationality status* of the father. However, as the home background data suffered from high percentages of missing values, student information was used to replace missing codes by valid answers wherever possible.

In some cases, the student and the teacher questionnaire covered aspects of a certain construct but provided complementary information. This was the case for the factor *quality of instruction*. Scholars differ in opinion regarding the issue of which data source should be given preference. While on one hand concerns have been raised about the likeliness of social bias in self-administered teacher questionnaires, on the other hand a lack of competence and stability often is seen as a threat to validity in assessments administered to younger students (Nilsen et al., 2016). Results from Scherer and Gustafsson (2015), who analyzed student assessments of classroom instructions in TIMSS and PIRLS 2011 data, in turn suggested that aggregated student responses on classroom level might be both valid and reliable. Another important point to be made for the current study is that the construct *quality of instruction* was underrepresented in both questionnaires. It was therefore decided to aggregate student-related questions for one dimension of *instructional effectiveness* on class level, and to additionally use the information from the teacher questionnaire for the remaining dimensions.

If there were overlapping concepts between teacher and school questionnaire, in general the teacher information was given preference. As teaching and learning takes place in the classrooms, and the teachers were in direct contact with the tested students, they should be in a better position to judge. For both subjects, altogether more than 550 variables from the school, teacher, student, and home questionnaires were reviewed; about 170 variables could be regarded as sufficiently matching the theoretical framework specified from a theoretical perspective of EER, and therefore selected for further processing. More information about the available questionnaire variables and the variables selected can be found in APPENDIX A.

8.3.3 Matching school, teacher, and student data

As a preparatory step for later multilevel analyses, the different datasets had to be merged, and assurance that all student level elements were linked to exactly one single element on the next higher level (usually the teacher or class level) was needed. Analysis of the student-teacher linkages, however, revealed that in Oman, Qatar, and the United Arab Emirates, up to around 5% of the students had two or more teachers linked to them. A similar pattern for these countries also could be found for science teacher linkages, and in Kuwait close to 17% of the sampled students were linked to more than one science teacher (see Table 8-1 for more information). While in primary education the vast majority of students are still linked to only one teacher per subject, and sometimes even to one teacher across subjects, remedial or advanced courses could be a reason for additional teachers found in the class.

Table 8-1: Percentages of students linked to more than one teacher

Country	Students linked to more than 1 mathematics teacher (%)	Students linked to more than 1 science teacher (%)
Bahrain	0	0
Kuwait	0	17
Oman	2	2
Qatar	4	2
Saudi Arabia	0	0
United Arab Emirates	5	5

When merging student and teacher data together, the resulting data set contains one entry for each student-teacher linkage combination. This means that the student information in the combined dataset is multiplied by the number of teachers teaching a student. This is not a problem for many student level analyses wherein the teacher weight is simply split by the number of teachers linked to the student. To perform multilevel analyses, on the other hand, each student only will need to be linked to one single element of the next higher analysis level. With the data at hand it is not possible to determine which teachers are the ones with the highest influence on a student, which would have made it easier to simply drop data from any additional "less-important" teachers. In order to keep all relevant data, therefore, an approach similar to the one described by Schulz-Heidorf (2016, p. 125) was adopted: instead of the Class Identification code (IDCLASS) that could have been used as an identifier for linking both levels if there was always only one single teacher per class, the Course ID (IDTEALIN), which is a combination of the teacher identification code (IDTEACH) and the link number (IDLINK) identifying a specific course taught by that teacher, was used. Each Teacher-Link or Course ID (IDTEALIN)

is unique within one and the same country and each student can be clearly assigned to a specific math or science course. In the case of two teachers teaching different math or science courses within one and the same class, this procedure now creates two *courses* with their student information duplicated, thus allowing for the inclusion of both in the final analyses. Student weights were adjusted accordingly, so that all created courses amount to the same weight as the sum of the weights from the original class. In this way, all available data could be kept – although, as noted, in the vast majority of classes only one teacher was teaching a class or course, and in all those cases the use of the Class ID and the use of the Teacher-Link ID would be equivalent.

In five out of the six GCC countries, either one or two classes per school participated in TIMSS 2015, while in Saudi Arabia consistently only one class was sampled for participation in the assessment. With this design, the variance on school level and the variance between classes cannot be clearly disentangled. In most educational effectiveness frameworks, the class level is seen as the most important level for studying school effects, as is it where teaching and learning take place. Correspondingly, for this research, the course level was selected as the main level of analysis (level 2 in the multilevel modeling steps) and school variables, in consequence, were disaggregated and merged to each course.

8.3.4 Preparing and exploring the data sets

Items that were identified as matching the factors of the proposed framework were selected and recoded to suit subsequent analyses. Variables were usually coded in such a way that higher values of the variable would, according to the literature review summarized in chapter 3, be expected to be associated with higher student achievement. In cases were an item was negatively phrased, the codes were reversed correspondingly. Only the *number of absences* was left with the original orientation, as a higher value here is assumed to be associated with a lower achievement.

Afterwards, descriptive statistics analyses were performed to gain a better understanding of the data. Basic statistics such as the mean and standard deviation, minimum and maximum, as well as percentiles were performed on all variables that were selected for inclusion in further analyses. Categorical variables were checked for valid ranges and distribution among categories, and the few numerical variables were checked for possible outliers. While a few relatively high values were found for the number of computers in schools (up to 1112 in Kuwait) and for the instructional time for math and science (up to 600 min per week in Oman, Qatar, and the United Arab Emirates), those values still were judged as being plausible. Hence, the cut-off points for

out-of-range values specified in the data sets by the TIMSS & PIRLS International Study Center, and applied during the international data cleaning procedures, were trusted. A special focus was set on identifying and handling missing data, as described in the following section.

8.3.5 Missing data

Missingness in survey data can be related to a variety of different causes. For example, simple mistakes during data entry, coding, or saving might be responsible for a subset of missing values. Missing values also can occur in cases where the respondent cannot or does not want to answer a certain question. For example, it might be too difficult for a fourth grader to answer a question about the highest education level of his or her parents. Additionally, missingness can be related to specific response patterns of the interviewees, leading to contradictory answers or complete denial of responses to certain questions or whole parts of the questionnaire. Responses related to certain personal background characteristics, such as personal income or profession, might be particularly prone to denial, an issue that begs the question of the extent to which such missing values occur randomly – or whether the level of missingness is associated with certain background characteristics of the respondent, such as ethnicity, age, or SES.

According to Schafer and Graham (2002, p. 151), who based their system on the initial typology developed by Rubin (1976), researchers generally distinguish between three different types of missing data: *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). *Missing completely at random* indicates that the probability for the occurrence of a missing value is not dependent on any other variable in the data set, missing or observed. *Missing at random* allows for the probabilities of missingness to be related with the observed data. For cases of *missing not at random*, the occurrence of missing values depends on other related variables or the missingness of those variables.

Different procedures are applied to handle missing data in analyses. Traditionally, missing data has been excluded from analyses, either by deleting the whole record if any single value is missing (*listwise deletion*), or by deleting only the specific missing values from analyses (*pairwise deletion*), with the latter procedure resulting in different sample sizes for each parameter. In general, deletion of missing values will result in a loss of sample size and statistical power. As mentioned above, certain characteristics of the respondents also might influence their willingness to respond; therefore, removing all the missing cases can lead to bias in the results. Another common practice is the *mean substitution*, in which each missing value is replaced by the average of the valid observed values, which might however affect the distribution of the

data by reducing the variances and lowering the correlations (Schafer & Graham, 2002, p. 149). While the above-mentioned procedure might be acceptable in cases with small rates of missing data, more elaborate procedures are needed once the missing rate becomes higher.

The missingness in the TIMSS 2015 grade four data for the GCC countries varies from country to country, and also differs quite largely from one background question to the next. While response rates were quite high in school, teacher, and student questionnaire data, lower rates were obtained from home questionnaire variables. The average missing rate in the home questionnaires is about 16% in four out of the six countries, reaching 24% in Qatar and even 29% in Kuwait. Missing rates are highest for the occupational status data, with a range of 29% (United Arab Emirates) to 40% (Kuwait) for mothers' and fathers' highest occupation level.

While missingness in surveys such as TIMSS cannot be assumed to be completely at random, determining the type of missingness usually is out of the analyst's control. For the purpose of the current study, missing at random will be assumed as recent research has demonstrated that an erroneous assumption of missing at random, for example by disregarding causes for missingness, often has only a minor impact on calculations and standard errors (Collins, Schafer, & Kam, 2001).

For the analyses described in this study, *multiple imputation* procedures will be applied. Multiple imputation replaces each missing value with a set of plausible values, allowing researchers to obtain an estimate of the uncertainty about the correct value to be imputed (Yuan, 2000, p. 1). To impute plausible values, the SAS™ procedure Proc MI (SAS Institute Inc., 2015) was used. Similar to the calculation of TIMSS mathematics and science achievement scores, five imputed background data sets were generated, and the results later combined, for the multilevel analyses. This process results in valid statistical inferences which properly reflect the uncertainty associated with the missing imputation process. To impute the missing values, the Markov chain Monte Carlo (MCMC) method was used, wherein pseudorandom draws from multidimensional probability distributions via Markov chains are taken. For more details about the procedure, refer to Yuan (2000). Altogether three imputation models, each including all analysis and the corresponding outcome variables, were calculated for each country: a student-level, course-level, and school-level model. The student-level model imputed plausible values for the student and the parent questionnaire data. This process resulted in five datasets with no missing data, but slightly different values for each of the imputed variables. After imputation, the SAS™ procedure PROC MIANALYZE was used to obtain summary statistics about the imputed datasets, which subsequently were roughly compared with the unimputed means and standard

deviations. Statistics for all compared variables between the imputed and the unimputed datasets were nearly identical.

Traditionally, the rounding off of imputed data, so as to conform to the nature of actual data, was recommended. This implied that imputed values follow the ranges of the non-missing data, and take on only discrete values for dichotomic or polytomous variables. However, simulation studies conducted by Ake (2005) and Allison (2005) found that the rounding off of responses might lead to estimation bias for calculating proportions and for regression parameter estimates. As the imputed values calculated for this study are mainly used for correlation and multiple regression analyses, it was decided to follow Allison's (2005) recommendation – which prefers unbiased estimates over a more "plausible" data structure by not rounding imputed values.

8.3.6 Data reduction procedures

This section describes the steps undertaken to reduce the number of variables to a smaller number of underlying factors that could be used in the subsequent multilevel analyses. In order to construct valid scales, sets of items were first inspected by principal component analysis (PCA), which belongs, in a wider sense, to the methods of factor analyses. Once a meaningful and statistically sound factor solution was obtained, the internal consistency of the items contributing to the final scales were examined through reliability analyses. Construct-validity of the final scales, as well as of single items that were retained for further analyses, were examined by way of correlation analyses, which investigated the relationship between the retained scale or variable and mathematics and science achievement. Principal component and reliability analysis were conducted by the statistical package SPSS™ (IBM Corp., 2011) and applied in parallel to all five imputed datasets. For this analysis step, a regional approach was followed, meaning that all data from the six GCC countries were pooled together.

8.3.6.1 Principal component analysis

Factor analysis in general allows for the combination of variables with common characteristics. Two general methods can be distinguished: *exploratory factor analysis* and *confirmatory factor analysis*. While the former, especially in its form as *principal component analysis,* allows for the identification of underlying patterns in previously unknown groupings of variables, *confirmatory factor analysis* is used to verify a hypothesized relation and grouping of certain identified factors (Cohen et al., 2007, p. 560). As the objective for the current study was to specify the underlying constructs in the region, *principal component analysis* was applied to extract the main factors in the form of a regional approach. This means that the analyses were conducted

on the pooled data of the six GCC countries by weighting each country equally. As no teacher questionnaire data was available for up to around 6% of the teachers in the six countries, separate group-level sampling weights were calculated to assure that each country has exactly the same weight of 500. For general variables, the adjusted group-level weights were based on the TIMSS general teacher weight (TCHWGT), while mathematics-related variables were based on the mathematics teacher weights (MATWGT), and science-related variables on the science teacher weights (SCIWGT). The analyses were applied to each of the five imputed data sets separately, based on the variables identified as being related to the conceptual framework in chapter 6. As only participating students with properly adjusted weights were included in the final international TIMSS database, level 1 student- (and parent-) level principal component analyses could be weighted using the available TIMSS student-level senate weight (SENWGT).

For the principal component analyses, only questions on an interval scale were included. The following steps were applied:

- The suitability of each set of variables for principal component analysis was checked by verifying the sample sizes and the item correlation matrix. For this purpose, the Kaiser-Meyer-Olkin (KMO) criterion was checked. The KMO ranges from 0 to 1 and the coefficient should obtain an absolute minimum of 0.5 (Bühner, 2011, p. 347). In addition, the result of Bartlett's test of sphericity was considered. This test checks the null hypothesis that the sample is originating from a population where the considered variables are uncorrelated. Bartlett's test should significantly reject the null hypothesis with $p < .05$ (Backhaus, 2011, p. 341).
- As the objective of this step was the data reduction towards underlying constructs, factor extraction was performed by applying principal component analysis (PCA). PCA can be used to reduce a large set of possibly correlated variables to a smaller number of uncorrelated factors (the principal components). Hereby, the first factor (or linear combination of variables) is extracted, such that the maximum shared variance of the original data set can be explained. In subsequent steps, other factors are then successively extracted, each time trying to explain the maximum portion of the remaining variance. The extraction results in a set of uncorrelated factors. To determine the maximum numbers of factors to be extracted, in general the *Kaiser criterion* was applied, meaning that all factors with an *eigenvalue* larger than 1 (which would correspond to the variance of one standardized variable) are extracted. The eigenvalue is a measure for the contribution of explained variance from one single factor regarding the variance of the whole variable set. Additionally, the graphical representation of the *Scree*

test was evaluated, and only factors above the elbow (or break) in the plot were retained (see Backhaus, 2011, p. 359).

- In order to allow for a better interpretation of the obtained factor results, a factor rotation was performed. For this study, *Varimax rotation*, an orthogonal rotation procedure, was applied. *Varimax rotation* maximizes the variance between factors and thus helps to more clearly identify the groups of variables that are closely correlated, and distinguish them from other variables (Cohen et al., 2007, p. 566).
- Factor loadings, which represent the correlation between original variables and obtained factors, were examined. Factor loadings can assume values between -1.0 and 1.0, with higher absolute values indicating stronger relationships. For the purpose of this study, loadings above 0.3 were considered as acceptable (Bühner, 2011, p. 350). In a few cases, items with double-loadings, which means that they load on more than one factor, were removed from the model. Additionally, communalities of each item after factor extraction were checked. The explained variance of each items by the factor solution should attain at least 10% (Bühner, 2011, p. 358).

8.3.6.2 Reliability Analyses

Once suitable and interpretable constructs were obtained through the PCA step, the internal reliability of the constructed scales were assessed by means of reliability analyses. The study adopted the *Cronbach's alpha* (α) as a measure of internal consistency of the scales created. Cronbach's alpha is used for multi-item scales, and checks inter-item correlations by measuring the correlation of each item with the sum of all other items (Cohen et al., 2007, p. 507). The coefficient ranges from 0 to 1; coefficients above 0.67 or even 0.8 are mostly regarded as acceptable (Cohen et al., 2007, p. 506). However, as the current study follows an exploratory research design and uses rather unreliable background questionnaire data, a lower coefficient of 0.5 for scales that are well justified from a theoretical perspective will still be considered for inclusion into further analyses. This approach is in line with other researchers, such as Bos (2002) and Cho (2010). In general, an alpha value between 0.7 and 0.8 will be judged as "acceptable", between 0.8 and 0.9 as "good", and above, 0.9 as "excellent". As an additional means of checking the scale homogeneity, each single item was checked to ascertain whether the whole scale would obtain a higher reliability if the item were dropped. In the case that removal of an item would enhance the scale reliability, this item was dropped from the scale.

Once factor and reliability analyses confirmed the consistency of the created scales, factor scores were saved and retained for further analyses.

8.3.6.3 Correlation Analyses

As a final step of the data reduction process, bivariate correlation analysis were performed in order to determine the association between background scales and mathematics and science achievement. First, correlations between the obtained scales, as well as single variables that were retained for further analyses, were verified to ensure that no *multicollinearity* occur in the data. Multicollinearity may occur when predictors are highly correlated with other predictors in the model. In multiple regressions (as will be applied for the multilevel analysis), this can interfere with determining the precise effect of each predictor in the final model. In cases of high correlations between variables, it therefore is suggested to either remove one of the concerned predictors or to combine the highly correlating variables into a new one (Cohen et al., 2007). For the current analyses, inter-item correlations between indicator variables higher than 0.8 (Field, 2004, p. 132) were further investigated.

For this study, the *Pearson product moment coefficient* was calculated with help of the IEA IDB Analyzer (see 8.3.1). The coefficients range from -1 to 1, indicating direction and strength of the relationship. Usually, correlations below 0.35 are classified as "low", between 0.35 and 0.65 as "medium", and above 0.65 as "high". However, this study will apply a minimum value of 0.2 for the correlation coefficient. This low cut-off point was chosen as in exploratory studies, when considering the high sample size, it might be worthy to also explore low correlations in exploratory relationship research (Cohen et al., 2007, p. 536). A correlation could also be described as the common variance that is obtained by squaring the correlation results (Cohen et al., 2007, p. 536). This means that a correlation of 0.2 then would explain 4% of the shared variance. Those percentages of explained variances that are relatively low, nevertheless, might still be important from a policy perspective (Teddlie & Reynolds, 2000, p. 98). All level 1 correlations were calculated based on the country-specific student sample sizes listed in Table 7-1, while level 2 correlations were calculated between the course averages of the predictor variables and the corresponding course averages of the outcome scores.

8.4 Creating an Index of Economic Social and Cultural Status (ESCS)

This section describes the construction of an index to describe the social background of the students in the GCC countries. The index will be part of the student variables used to control for differences in the student background, when evaluating relations with course and school

environment variables on student's academic achievement. Controlling for the student background should allow for disentangling the home background from the school environment effects, hence allowing to better capture the school effects "net of" the influences from the social background, as suggested for example by Buchmann (2002, p. 151).

In modern national and international assessments, a variety of variables are used to capture the social family background and, depending on the use of the indicator, different techniques are used for its construction. Sirin (2005, p. 418), conducting a meta-analysis in this area, found that in social studies the social background was often defined as relating to the concept of the socio-economic status (SES). SES, according to Mueller and Parcel (1981, p. 14), describes an individual's (or a family's) position in a hierarchically-organized society to access or exert control over wealth and power – often with parental income, parental education, and parental occupation as core indicators. Recent research is oriented by the theoretical foundations of the underlying processes, often drawing from the theory of capitals by Bourdieu which also will be the main focus for the current study (see also section 3.3.4.3 for more details on theories to explain disparities in the social background of students). Bourdieu not only focused on the importance of economic capital, but also elaborated that the cultural and social capital of a family are important to finally explain (and maintain) social disparities, which in turn influences the school learning experiences of the children. More recent studies, therefore, extend the economical component by a cultural component to more comprehensively and accurately measure social background conditions. However, variables related to *social capital* are scarce in international assessments, probably because the measurement of social capital is far more complex than for other forms of capital, as it would need comprehensive information about the network of the family or person. Accordingly, the TIMSS 2015 data does not contain relevant variables well suited to measure the social capital. As concluded from the discussions detailed in section 3.3.4.3, it could be assumed that especially in the GCC countries, among the national populations, social capital might be of high importance – because it is the social network and the proximity to the ruling elite that defines access to material goods, power, and also education.

To capture the economic and cultural capital of a family, a variety of variables are available in the TIMSS questionnaires, as detailed below.

Economical capital can be measured by questions related to the absolute or relative economic wealth of a family, or certain home possessions, which express economic wealth. Suitable country-specific items of GCC countries asked in TIMSS 2015 might comprise for example: *swimming pool* (United Arab Emirates, Bahrain), *luxury car* (United Arab Emirates), *private house*

maid (Qatar), or *private garden* (Saudi Arabia). However, as exhibited above, items selected by the GCC countries vary across the region, indicating diminishing suitability for the current project, which is focused on regional comparability.

As questions directly related to income and economic wealth usually generate high missing rates, the occupational status of the parents is often used as a proxy for the economic situation of a family. A certain occupation also usually requires a particular education level, which makes occupation therefore also partly an indicator for the institutionalized cultural capital. However, as the current study does not intend to separate the effects between both forms of capital, using the occupational level of the parents here is deemed as suitable.

As a starting point for the comparison of different occupations of parents, in terms of wealth characteristics or prestige, a standardized manner of collecting occupational information is needed. For this purpose, the International Labour Office (2012) developed the *International Classification of Occupations (ISCO)*, allowing for the hierarchical categorization of jobs into clearly defined groups according to its tasks and duties. Based on this standardized job classification, different models are available from which to derive information about the prestige of a certain occupation. For international comparisons, such as in PISA, the *International Socio-Economic Index of Occupational Status (ISEI)* is mainly used (Ehmke & Siegle, 2005; Marks, Cresswell, & Ainley, 2006). Based on the occupational data of 16 countries, the ISEI was developed by Ganzeboom, Graaf, and Treiman (1992) and can be interpreted as measuring "the attributes of occupations that convert a person's main resource (education) into a person's main reward (income)" (Ganzeboom et al., 1992, pp. 8–9). Occupations coded with the ISCO classification can be transferred to a corresponding value between 16 and 90 on the ISEI scale. While the occupational data in TIMSS were not collected according to the ISCO classification system, they can still be represented on the ISEI scale, following a matching procedure developed by Caro and Cortés (2012) for PIRLS 2006. As the parental occupation classification in the TIMSS 2015 questionnaires still matches the classification used in PIRLS 2006 exactly, the TIMSS 2015 parental occupations can also be translated to the ISEI scores using the same matching procedure. The obtained ISEI scores for each of the original TIMSS occupational categories are shown in Table 8-2.

.

Table 8-2: Match between TIMSS occupation categories (Variables ASBH23A/B) and ISEI scores following the procedure of Caro and Cortés (2012)

TIMSS 2015 original Occupational Categories		ISEI Score
ASBH23A/B	Label	
1	Has never worked outside the home for pay	22
2	Small business owner (< 25 employees)	57
3	Clerk	49
4	Service or sales worker	45
5	Skilled agricultural or fishery worker	31
6	Craft or trade worker	37
7	Plant or machine operator	33
8	General laborers	24
9	Corporate manager or senior official	67
10	Professional	73
11	Technician or associate professional	52

Notes. ASBH23A = Father's occupation level/ ASBH23B = Mother's occupation level
ISEI = Economic Index of Occupational Status

The main indicator for *cultural capital* in international comparative assessments is usually the educational level of the parents, which more precisely is an indicator of the *institutionalized* cultural capital. To ensure a standardized and comparable collection of the parental level of education among different countries, in TIMSS, as well as in other comparative assessments, the *International Standard Classification of Education* (*ISCED*; UNESCO, 2012b) is used. TIMSS data is collected according to the most recent ISCED 2011 standard. Similar to Ehmke and Siegle (2005), the ISCED levels were converted to an approximation of the number of school years in order to obtain the respective education level, allowing for a better comparison among the countries and especially between the different levels of education. The data for the conversion was obtained from the UNESCO Institute for Statistics [UIS] (2017). Table 8-3 shows the final matching values used after evaluating the UNESCO data.

Table 8-3: Match between ISCED, TIMSS educational categories, and years of schooling

ISCED 2011 levels	TIMSS education levels	Years of schooling					
		Bahrain	Kuwait	Oman	Qatar	Saudi Arabia	United Arab Emirates
	Did not go to school	0.0	0.0	0.0	0.0	0.0	0.0
1	Some primary or lower secondary	6.0	5.0	6.0	6.0	6.0	5.0
2	Lower secondary	9.0	9.0	9.0	9.0	9.0	9.0
3	Upper secondary	12.0	12.0	12.0	12.0	12.0	12.0
4	Post-secondary, non-tertiary	14.0	14.0	13.0	13.0	13.0	13.0
5	Short cycle tertiary	14.0	14.5	14.5	14.0	14.5	14.0
6	Bachelor or equivalent	16.0	16.0	16.0	16.0	16.0	16.0
7	Masters/Doctor	18.0	18.0	18.0	18.0	18.0	18.0
8		21.0	21.0	21.0	21.0	21.0	21.0

Note. Years of schooling for the different ISCED levels approximated by statistics from UIS (2017).
ISCED = International Standard Classification of Education (UNESCO, 2012b)

As an additional measure for cultural capital, indicators related to the *objectified form* of cultural capital are also often used. While the objects themselves are just material goods that can easily be exchanged to a form of economic capital, cultural goods such as *books* or *musical instruments* can only be adequately used by those who also possess the corresponding necessary *incorporated cultural* capital (Bourdieu, 1983, p. 190). Additionally, participation in cultural events may complete the indicators of cultural capital, but related questions are not available in the TIMSS 2015 questionnaires, and comparability between countries is doubtful.

TIMSS 2015, however, contains information about the *number of books at home*. It should be noted here that especially in Arab countries, the *number of books* might be of limited value as an indicator for cultural capital. This assumption can be drawn due to a long tradition in the region of orally transmitting information, and a late introduction of printing technology (Robinson, 1993). Correspondingly, the associations between number of books at home and student achievement can be expected to be somewhat lower when compared to Western countries, especially for the Arab national population. Nevertheless, the number of books may work better for the large non-national populations in the GCC countries, and also may have increased value for the new group of national *businessmen*. Therefore, the variable still should be included here.

Hence, for the current study, the following variables will be included for the index creation: the highest occupation level of the parents transferred into HISEI scores, the highest education level of the parents converted into years of schooling, and the number of books at home.

Modeling of the main components related to economic and cultural capital follow the conception of Ehmke and Siegle (2005), who used different variables to combine the economic and cultural aspects into a common index, namely the Economic, Social, and Cultural Status (ESCS)

index. Analyzing PISA data, they could show that the ESCS index covers the student background more comprehensively than single concepts of economic or cultural capital used in earlier studies, and they could also explain more variance in the student achievement when compared to using single variables. Moreover, the current study applies complex multilevel models with a number of different variables, which also calls for a parsimonious conception of the student background portion. Ehmke and Siegle (2005) used as a basis for their index the highest occupational status of either parents measured on the ISEI scale (the so-called HISEI), the highest parental education level converted into years of schooling (HISCED), and a measure of different home possessions. Here, ISEI and years of schooling are calculated separately for both parents in order to yield a greater reliability, with more variables included, and to achieve a better balance between concepts and variables, as argued by Caro and Cortés (2012, p. 25). As the country-specific home possessions defined by the GCC countries turned out to be quite different across the region, and combinations of common home possession items included in the TIMSS questionnaires only showed very low correlations with student achievement, the number of books at home was the only item included in the creation of the ESCS index, instead of a larger set of home possessions (similar to the approach of Schulz-Heidorf, 2016, p. 140).

Table 8-4 and Table 8-5 show the correlation between the variables used for the background model and mathematics and science achievement, respectively. The tables are based on the student level sample sizes presented in Table 7-1. Results show that of all variables used to create the ESCS Index, the Z-standardized educational levels of the parents (ZSJSBH20A = years of schooling of the father, ZSJSBH20B = years of schooling of the mother) have the highest correlation with student achievement in all countries. They are followed, with the exception of the maternal occupational status in Bahrain, by both of the Z-standardized paternal ISEI scores (variables ZJSBH23A and ZJSBH23B). The number of books at home seems to be less important, especially in Kuwait and Saudi Arabia. The created ESCS index (variable F_ESCS) exhibits higher correlation with achievement, in comparison to each of its components, in all countries.

As discussed in section 3.3.4.3, in the Gulf region, nationality and gender are also important determinants for an individual's position in society; therefore, both variables will also be used in the subsequent multilevel models to better capture the student's social background. Nationality is based on the question "Was your father born in country" – as in all GCC countries the birthplace of the father is the main determinant for the nationality of his children (see APPENDIX B for more details).

Table 8-4: Correlation between SES variables and mathematics achievement

Mathematics							
Variable	**Description**	**BHR**	**KWT**	**OMN**	**QAT**	**SAU**	**ARE**
JBOOKS	# of books at home	**0.16**	**0.08**	**0.14**	**0.13**	**0.05**	**0.16**
ZJSBH20A	Education father (years of schooling)	**0.27**	**0.23**	**0.21**	**0.35**	**0.13**	**0.41**
ZJSBH20B	Education mother (years of schooling)	**0.24**	**0.21**	**0.21**	**0.31**	**0.13**	**0.39**
ZJSBH23A	Occupational status father (converted to ISEI)	**0.18**	**0.17**	**0.18**	**0.23**	**0.11**	**0.27**
ZJSBH23B	Occupational status mother (converted to ISEI)	**0.15**	**0.11**	**0.15**	**0.21**	**0.09**	**0.22**
F_ESCS	*Economic, social and cultural status (index)*	**0.29**	**0.26**	**0.25**	**0.38**	**0.15**	**0.44**

Notes. Significant correlations (0.05 level [2-tailed]) are marked in bold
BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates

Table 8-5: Correlation between SES variables and science achievement

Science							
Variable	**Description**	**BHR**	**KWT**	**OMN**	**QAT**	**SAU**	**ARE**
JBOOKS	# of books at home	**0.14**	**0.07**	**0.14**	**0.13**	**0.08**	**0.13**
ZJSBH20A	Education father (years of schooling)	**0.24**	**0.22**	**0.22**	**0.34**	**0.17**	**0.43**
ZJSBH20B	Education mother (years of schooling)	**0.22**	**0.23**	**0.22**	**0.29**	**0.17**	**0.41**
ZJSBH23A	Occupational status father (converted to ISEI)	**0.16**	**0.14**	**0.18**	**0.22**	**0.14**	**0.27**
ZJSBH23B	Occupational status mother (converted to ISEI)	**0.13**	**0.13**	**0.16**	**0.19**	**0.11**	**0.23**
F_ESCS	*Economic, social and cultural status (index)*	**0.26**	**0.26**	**0.26**	**0.36**	**0.20**	**0.45**

Notes. Significant correlations (0.05 level [2-tailed]) are marked in bold
BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates

8.5 Multilevel Analysis

EER, as performed in this study, examines the relationships between individuals and the social context in which they learn. Individuals are influenced by the groups to which they belong; in turn, characteristics of the groups influence the individual. Educational systems form a hierarchical system of students nested within classes, classes within schools, and schools within regions or countries. Relations of student, teacher, and school characteristics can be described at different levels, as described in the general framework chapter. TIMSS data was collected according to the hierarchical structure of the educational systems, as also described in the framework. Multilevel modeling allows researchers to take effects on the different levels into account, and therefore is the recommended analysis technique for this kind of data. Kyriakides and Charalambous (2005), in comparing single-level regression analyses with multilevel analyses based on TIMSS 1999 data, concluded that the errors associated with not taking the hierarchical structure of the TIMSS data into account are substantial, and may not be ignored. They strongly recommended the use of multilevel analyses with hierarchically structured IEA data, as multilevel modeling allows for the partitioning of variance into different levels, thus generating a better understanding of the phenomenon under study.

For the purpose of this study, multilevel regression models will be applied, which are essentially multilevel versions of the standard multiple regression models. The basic concept is that a hierarchically organized data set is analyzed with one single outcome variable on the lowest level (in this case, mathematics and science achievement), and variables on all educational levels serve as explanatory variables. The term *multilevel regression model* is known in the research literature under a variety of terms, such as *random coefficient model*, *hierarchical linear model*, or *variance component model* (Hox, Moerbeek, & van de Schoot, 2017, p. 8).

8.5.1 General characteristics

TIMSS data is collected via a hierarchical structure, in which lower-level units are sampled in subsequent stages within the higher-level units. This procedure results in school-level data sets that contain the responses from the principals, and teacher files that contain the teacher- and class-related information. The teachers are linked to their schools via a hierarchical identification system; on the lowest level, the student and home data sets contain contextual information, as well as mathematics and science assessment results on the student level. The hierarchical identification system allows linkages of students to their teachers and their schools. For the current analyses, two levels will be considered: the student level (level 1), and the course level (level 2). In the vast majority of cases, in which students are taught by only one mathematics and one science teacher (which also could be the same person), the course level is equivalent to the class level, selected as a second stage in the sampling process. However, as discussed in the section on preparing the data (8.3.3), in some cases students were taught by different teachers. In such cases, student information was multiplied, corresponding to the number of teachers by which they were taught. This process ensures a consistent hierarchical structure, wherein groups of students are linked to a specific *course* a teacher is teaching, which is a precondition for the multilevel modeling.

An advantage of multilevel modeling is that variables can be defined at any level of the hierarchy, and the variance components on different levels are dealt with simultaneously, without the need to disaggregate or aggregate variables to a certain level of interest. Aggregation and disaggregation of data might lead to a statistical problem, the so-called *aggregation* or *disaggregation bias*. Aggregation bias occurs if data points from lower-level units are combined with fewer higher-level units. In such cases, much information is lost, and the analysis results lose power. Additionally, aggregation bias might lead to misinterpretation of the data by tapping into the so-called *ecological fallacy*. This term describes the possible mistake of applying inferences made from results found for the group back to the sub-group or individual level, while

the effects on both levels might be substantially different in hierarchical systems. Disaggregating data to lower levels, on the other hand, results in an overestimation of the sample size by standard analyses methods, resulting in "many 'significant' results that are totally spurious" (Hox et al., 2017, p. 3). Neglecting the hierarchical structure of the data is also no solution in such cases, as the ordinary regression results would give an "uninterpretable mixture of effects from within and between group effects" (Cronbach, Deken, & Webb, 1976, p. 236).

An additional aspect to be considered is that in nested data, individual observations are rarely independent from each other. For instance, students from the same schools tend to be more similar than students picked randomly across the country. They share, in general, a similar background and are influenced by the same school factors. As a result, correlations measured between students from the same group will be higher compared to variables measured from students from different schools. Applying standard statistical procedures would result in an underestimation of standard errors, and again might result in spurious "significant" results (Hox et al., 2017, p. 4).

An additional advantage of multilevel modeling is the possibility of investigating the interaction between factors within each level, but also between levels (cross-level interaction). Consequently, multilevel analysis provides a picture of the decomposition of achievement variance in the whole educational system, and at the same time, the factors affecting it.

8.5.2 Building the models

All components retained from the previous analysis steps during the variable selection procedures on regional level were used as a common "frame" which, in the subsequent multilevel analyses, were further examined. This means that the starting point for the multilevel analyses is a common set of variables, identified as possible effectiveness-enhancing factors in the region, according to the proposed framework. Nevertheless, as the preliminary results show, the identified factors are expected to work differently among countries in the region. The multilevel analyses, therefore, will fit separate models for each country based on the previously-selected regional components.

As only one class per school was selected for participation in the assessment in about 30% of the schools in the region, variances on course level on the one hand, and school level on the other, cannot be clearly disentangled. It was therefore decided to apply a two-level approach, with level 1 describing the student level and level 2 describing the course level – including the disaggregated school variables. The two-level models allow for the distinguishing of variance

in student mathematics and science achievement, accounted for at the student level, from variance components accounted for at the course/school level. The main objective was to quantify the relationship of school-, teacher-, and course-level factors – identified according to the proposed model – with student achievement. Controlling for the home background here allows for more clear identification of the effect of the influences of the school environment "net of" outside school influences, thus providing important information for policymakers to detect malleable school instructional and environmental factors in their educational systems, as a basis for actions and regulations to improve the quality of education. Detailed information on the set of analysis variables used for the subsequently described models can be found Table 9-32 for mathematics and in Table 9-33 for science, while detailed information regarding the centering of predictor variables, the weighting, and the variance estimation procedures can be found at the end of the chapter.

The model building procedures are summarized below.

Step 1: building a null model

The *null model*, also called an *unconditional model* or *intercept-only model,* does not contain any explanatory variables and is used to estimate the total variance and the variance components between courses (level 2) and within courses (level 1). The equation for the null model is given below:

$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$

Y_{ij} = the dependent variable (TIMSS mathematics, respectively science achievement scores) with $_i$ denoting the individual student and $_j$ denoting the courses

γ_{00} = intercept (or regression coefficient), the expected value of the dependent variable when all explanatory variables have a value of zero

u_{0j} = residual error at the course/school level (level 2)

e_{ij} = residual error at the student level (level 1)

The difference to a single-level regression model is given by attaching the subscript to the regression coefficient, thus indicating that each course (level 2 element) has a different intercept coefficient and different slope coefficients (Hox et al., 2017, p. 13).

The residual error at the student level (e_{ij}) was assumed to follow a normal distribution, with a mean of zero and a variance of σ^2. Similarly, the random group level effect (u_{0j}) was assumed to be normally distributed, with a mean of zero and a variance of τ_{00}. Thus the total variance in mathematics and science achievement, respectively, is the sum of the within- and between-courses/school variance: Var (Y_{ij}) = $\sigma^2 + \tau_{00}$

The proportion of group-level variance is referred to as the *intra-class correlation* ρ:

$$\rho = \frac{\tau_{00}}{\sigma^2 + \tau_{00}}$$

The variance components on level 1 (σ^2) and on level 2 (τ_{00}) represented the total available variance. The focus of the current analyses was to explain variance, especially on group level, by the addition of student- and course-level predictors in the subsequent models. The null model also serves as a benchmark for comparison with the subsequent, more complex models, in terms of the model fit, as measured by the deviance (Hox et al., 2017, p. 19).

Step 2: building the level 1 model

To build the level 1 country models, all level 1 variables selected in the previous analyses steps were simultaneously added to the null model. The equation for the level 1 (or student-level model) can then be given as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{30}X_{3ij} + \gamma_{40}X_{4ij} + \gamma_{50}X_{5ij} + \gamma_{60}X_{6ij} + u_{0j} + e_{ij}$$

with:

$X_{1..6ij}$ = student level background variables (1: *ESCS*, 2: *early numeracy*, 3: *nationality status*, 4: *student likes learning* (subject motivation), 5: *absence from school*, and 6: *help with homework*), and

$\gamma_{10..60}$ = regression coefficients for the student level variables.

Hence, the variables considered on level 1 included the *index of economic, social, and cultural status (ESCS)*, the factor created from the *early numeracy tasks* as a proxy for student's aptitude, the *nationality status* of the student (coded as 0=national/ 1= non-national), students' *subject motivation*, the *number of absences per month* as an indicator for the model factor *time*, and the *parental support for homework* as an indicator for the model factor *opportunity*. Students' *gender* – albeit recognized as an important student background indicator in the region – was not entered on level 1, in order to allow for structurally equal models across the countries.

In Saudi Arabia, all classes were gender-segregated (as were many classes in other countries), and with no gender variability in those classes, a gender effect cannot be estimated. Gender, however, was regarded as a composition variable in subsequent models.

Step 3: building a home background control model by entering aggregated home background indicators on level 2

In order to quantify the percentage of between-course variance attributable to the student's home background, home background variables were entered on both levels of the model.

In addition to the average ESCS index and the course average of a factor created from the *early numeracy tasks*, the share of *girls per class* (ranging from 0 = no girls to 1 = 100% girls), and an indicator about the average *nationality status* (ranging from 0 = no non-nationals to 1 = 100% non-nationals) were also entered on level 2. As was discussed in chapter 2, the latter two variables are assumed to play important roles as additional indicators for the student background in parts of the Gulf region.

The models in this and subsequent steps are created as *variance component models,* in which their residual variance is divided into components corresponding to each level of the hierarchy. Variance component models assume random regression intercepts and fixed regression slopes (Hox et al., 2017, p. 46), and thus are also called *random-intercept models* (see for example Raudenbush & Bryk, 2002, p. 102).

The complete student background model can then be formulated as:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{30}X_{3ij} + \gamma_{40}X_{4ij} + \gamma_{50}X_{5ij} + \gamma_{60}X_{6ij} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{3j} + \gamma_{04}Z_{4j} u_{0j} + e_{ij}$$

with:

$X_{1..6ij}$ = student level background variables (1: *ESCS*, 2: *early numeracy*, 3: *nationality status*, 4: *student likes learning* (subject motivation), 5: *absence from school*, and 6: *help with homework*),

$\gamma_{10..60}$ = regression coefficients for the student level variables,

$Z_{1..4j}$ = aggregated course/school-level student background variables with 1: *ESCS* (avg.), 2: *early numeracy* (avg.), 3: *Nationality status* (avg.), and 4: *gender* (avg.), and

$\gamma_{01..06}$ = regression coefficients for the aggregated (level 2) student variables.

Step 4: building the level 2 explanatory model

The purpose of the level 2 explanatory model was to investigate the association between course- and school-level variables of the model with achievement, but without controlling for home background. Hence, neither the level 1 student background variables nor their level 2 aggregates (the student composition variables) were included in this model. The residual variance components were used to calculate the share of variance that can be explained on level 2 by the course/school-level explanatory variables.

This model can be formulated as:

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z_{1j} \ldots \gamma_{0q}Z_{qj} + u_{0j} + e_{ij}$$

With $Z_1 - Zq$ denoting all explanatory course and school level variables on level 2.

Step 5: building the full model

The full model was constructed by entering the background model variables from Step 3, and the course/school-level predictors from Step 4, jointly into a common model for each of the countries. The full models were then used to quantify the association of school context factors with student mathematics and science achievement, while controlling for home background. These models again are calculated as *variance component models* (or *random-intercept models*) with random regression intercepts and fixed regression slopes. Hox et al. (2017, p. 46) suggest to start with those kind of models, as they can usually be estimated with higher precision compared to models with random parts.

The complete model then can be formulated as:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{01}Z_{1j} + \ldots + \gamma_{p0}Xx_{ij} + \gamma_{0p}Z_{pj} + u_{0j} + e_{ij}$$

with:

X_{pij} = student level explanatory variables (from Step 2)

Z_{qj} = course/school-level explanatory variables (from Step 4 and including the aggregated student background indicators from Step 3).

Detailed information about the predictors included in all the different models can be obtained from Table 9-32 for the mathematics multilevel analyses, and from Table 9-33 for the science.

As the applied imputation method resulted in five imputed course-level data sets (which also included the disaggregated school predictors), and also in five imputed student-level data sets, all files had to be merged before multilevel analyses could be performed. See also section 8.3.5 for more information on missing data imputation. All analyses were performed five times, always pairing one of the five imputed background variable sets with one of the five plausible values used as outcome variable. Final estimations were obtained by averaging the results from the five calculations, and appropriate standard errors were calculating according to the formula described by Little and Rubin (1989, p. 305). Pairing each of the imputed datasets with one plausible value used to be the common approach when analyses for the current research project were started – and as such was continued throughout the project. However, with today's higher computer power, future analyses using imputed datasets may obtain a slightly more accurate calculation when all plausible values are paired with all imputed datasets.

Centering

The appropriate centering of variables is an important issue in multilevel research, as it can make the interpretation of results more meaningful and may reduce collinearity between predictive variables and interactive variables containing these predictors (O'Connell & McCoach, 2008). Centering is usually achieved by subtracting the overall mean (*grand-mean centering*) or group means (*group-mean centering*) from each individual predictor value. While grand-mean centering produces a model which is mathematically equivalent to the raw-score model, group-mean centering of the predictor variables removes all information related to between-group differences, resulting in different parameter estimates.

Grand-mean centering is generally recommended for most multilevel analyses of school effects (O'Connell & McCoach, 2008, p. 95), but group-mean centering is preferred if the researcher is particularly interested in investigating how group compositions affect student performance (Paccagnella, 2006, p. 70) .

The primary focus of the analyses here is set on the influences of course- (including school-) level variables on individual student achievement. Following the guidelines of Enders and Tofighi (2007, p. 136), grand-mean centered level 1 variables consequently needed to be entered into the multilevel models. Only the dichotomous *nationality indicator* was entered uncentered to the models, in order to allow for easier interpretations of the results. Using grand-mean centering for school effects research is also supported by O'Connell and McCoach (2008,

p. 97), who suggested using grand-mean centering for analyses with a primary interest on identification and interpretation of school effects on student achievement – which is the primary interest of the current project.

However, in order to also compare the findings of the current study to analyses from other authors using a similar research design but with different choices in terms of centering, calculation of the full models were also performed by centering all level 1 variables around the group-mean. Results of the explained level 2 variances can be found in APPENDIX D (Table D-1 for mathematics and Table D-2 for science).

Calculating the explained variance

To obtain a measure of the effect size for the different models, for each set of models, the proportion of explained variance was calculated following the guidelines from Raudenbush and Bryk (2002), who recommended developing the complete level 1 model first (see Step 2), and only then to proceed with entering level 2 predictors. The rationale is that an introduction of level 1 predictors, in addition to reducing the level 1 residual variance, also may change level 2 variance components. This means that a reduction in level 2 variance is only interpretable for the same level 1 model (Raudenbush & Bryk, 2002, p. 144). The level 1 model is used to explain the share of level 1 variance, while level 2 variance is explained by adding additional level 2 predictor variables (see Step 3 and Step 5) to the level 1 base model and then comparing the level 2 variance components against the level 1 model (Step 1), according to the following formula:

$$Percent\ \tau_{00\ (explained)} = \frac{\tau_{00\ (level\ 1\ model)} - \tau_{00\ (model\ including\ level\ 2\ predictors)}}{\tau_{00\ (level\ 1\ model)}}$$

Only the model containing solely level 2 predictor variables (Step 4) was compared directly to the null model (Step 1) to calculate the possible reduction in level 2 variance.

Sampling weights

All multilevel analyses were weighted according to the math and science teacher weights provided with the TIMSS international database, respectively. For the current analyses, no appropriate level 2 weight components were available, as neither teachers nor courses were specifically selected in a separate step of the sampling process. Instead, intact classes were selected, and subsequently all students within classes were (usually) selected for participation. Accordingly, teachers (or, more specifically, the related course data) rather represent *attributes* of the

students selected for the test. Therefore, the most appropriate approach for the current project was regarded to be the use of only level 1 weights. Thus, the current analyses in this perspective followed a similar approach as the school effectiveness analyses conducted by Martin and Mullis (2013), who stated that using the overall student sampling weights specified at the student level would make it unnecessary to provide sampling weights at the school level (Martin & Mullis, 2013, Technical Appendix B). However, in contrast to Martin and Mullis' analyses, students in the current analyses were linked to courses taught by the TIMSS teachers; more importantly, student entries were duplicated for those cases in which students were taught by different teachers (see section 8.3.3 for more details on this procedure). Student entries appearing more than once, therefore, needed to be weighted down properly (meaning that the student weight needed to be divided by the number of teachers linked to him or her), which resulted in the use of the corresponding teacher weights already available in the TIMSS database.

Software used

There are several software programs available for estimating multilevel regression models, including HLM, Mlwin, MPlus, and SAS. For the current analyses, the SAS 9.4 Procedure PROC GLIMMIX (SAS Institute Inc., 2015, PROC GLIMMIX) was used. The parameters are estimated by maximum likelihood, wherein the marginal distribution is numerically approximated by the adaptive Gaussian quadrature. PROC GLIMMIX is a relatively new SAS procedure introduced first as an add-on to SAS 9.1. Further description of the use of PROC GLIMMIX with complex survey data is given by Zhu (2014).

9 RESULTS OF VARIABLE SELECTION AND FACTOR/ RELIABILITY ANALYSES

9.1 Results of the Preliminary Analyses Related to Gender and Nationality

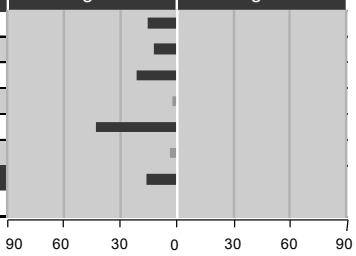
The following results will give an overview of the distribution of two important background variables in the region: namely, gender and nationality status, as well as their respective associations with achievement. These variables are presented here as they show quite high disparities in most countries of the region, while at the same time are important determinants regarding an individual's position in the hierarchy of the GCC States (see "The societies in the GCC countries" in section 3.3.4.3).

Gender differences

Overall gender differences can be retrieved directly from the international mathematics and science reports, and don't need to be calculated separately. Nonetheless, results are listed here, as they should be regarded in more depth for additional analyses. Although gender differences between TIMSS 2011 and 2015 were considerably reduced in most of the countries, especially in secondary education, the GCC countries are still among the countries with the highest gender disparities – in favor of girls – in both grades of TIMSS. Details regarding gender differences in TIMSS grade four can be found in Table 9-1 and Table 9-2. The results of PIRLS 2011 (Mullis, Martin, Foy, & Drucker, 2012) and PISA 2015 (OECD, 2016a) also show a similar pattern.

Table 9-1: TIMSS 2015 average mathematics achievement by gender

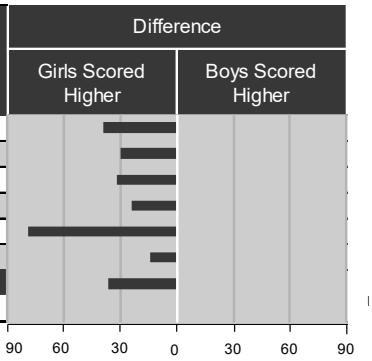
Country	Girls		Boys		Difference (Absolute Value)	Difference	
	Percent of Students	Average Math Score	Percent of Students	Average Math Score		Girls Scored Higher	Boys Scored Higher
Bahrain	50 (0.7)	459 (1.7)	50 (0.7)	443 (2.3)	15 (2.5)		
Kuwait	51 (2.0)	359 (5.4)	49 (2.0)	347 (5.6)	12 (6.2)		
Oman	50 (0.7)	436 (3.0)	50 (0.7)	415 (2.8)	22 (2.9)		
Qatar	51 (2.5)	440 (4.1)	49 (2.5)	438 (4.9)	3 (5.9)		
Saudi Arabia	49 (1.0)	405 (4.4)	51 (1.0)	363 (6.5)	43 (7.7)		
United Arab Emirates	48 (2.2)	453 (3.9)	52 (2.2)	450 (3.4)	3 (5.4)		
Gulf Average	**50 (1.7)**	**426 (3.9)**	**50 (1.7)**	**409 (4.5)**	**16 (5.4)**		
Int. Average	**49 (0.2)**	**505 (0.5)**	**51 (0.2)**	**505 (0.5)**			



Notes. Data summarized from Mullis, Martin et al. (2016)
() Standard errors appear in parenthesis
Bars in dark color indicate statistically significant differences

Table 9-2: TIMSS 2015 average science achievement by gender

Country	Girls		Boys		Difference (Absolute Value)	Difference	
	Percent of Students	Average Science Score	Percent of Students	Average Science Score		Girls Scored Higher	Boys Scored Higher
Bahrain	50 (0.8)	478 (3.0)	50 (0.8)	439 (3.5)	39 (4.0)		
Kuwait	51 (2.1)	352 (7.6)	49 (2.1)	322 (7.6)	30 (9.1)		
Oman	50 (0.7)	447 (3.4)	50 (0.7)	415 (3.6)	32 (3.1)		
Qatar	51 (2.5)	448 (4.7)	49 (2.5)	424 (6.0)	24 (7.2)		
Saudi Arabia	49 (1.0)	431 (5.3)	51 (1.0)	352 (7.6)	79 (9.0)		
United Arab Emirates	48 (2.2)	459 (4.4)	52 (2.2)	444 (4.0)	14 (6.4)		
Gulf Average	**50 (1.7)**	**436 (5.0)**	**50 (1.7)**	**399 (5.7)**	**36 (6.8)**		
Int. Average	**49 (0.1)**	**508 (0.5)**	**51 (0.1)**	**504 (0.6)**			



Notes. Data summarized from Martin, Mullis, Foy et al. (2016)
() Standard errors appear in parenthesis
Bars in dark color indicate statistically significant differences

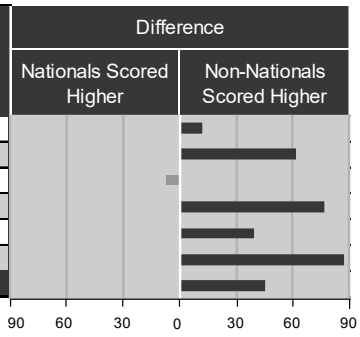
While Qatar and in the United Arab Emirates show no significant gender disparities in mathematics, in Saudi Arabia the differences reach 43 score points in favor of girls. In science, all GCC countries show significant gender differences in favor of girls; in general, the magnitude is larger, and even reaches up to 79 score points in Saudi Arabia. Thus, the average gender gap for science in the region amounts to more than twice that of mathematics. Outcomes from international assessments are in line with results from national examinations in the GCC countries, where girls regularly outperform boys across all grades of schooling (Alkhateeb, 2001; Egbert, 2012; Ministry of Education Oman & World Bank, 2012). Furthermore, the gender gap seems to exhibit quite early in schooling. The Omani Ministry of Education reported that already in grade 1, when boys and girls are still co-educated, Omani teachers consistently assign higher ratings to girls in all evaluated subjects except sports (Ministry of Education Oman & World Bank, 2012, p. 26).

Differences in terms of nationality status

While literature in this regard seems to be scarce, the different achievement level outcomes in TIMSS mathematics and science for nationals and non-nationals, as depicted in Table 9-3 and Table 9-4, suggest a relation with the very different living conditions for both sub-populations, which were discussed in section 2.1. The nationality status for the current study was defined by the father's place of birth, as in the GCC countries only children born to a national father are automatically citizens of that country. In some GCC countries, children born to a stateless father and a mother with the nationality of the respective country are also usually citizens of that country (Albarazi, 2017).

Table 9-3: TIMSS 2015 average mathematics achievement by nationality status

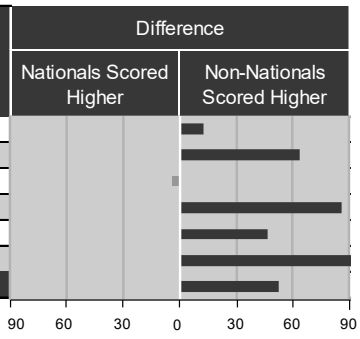
Country	Nationals		Non-Nationals		Difference (Absolute Value)	Difference	
	Percent of Students	Average Math Score	Percent of Students	Average Math Score		Nationals Scored Higher	Non-Nationals Scored Higher
Bahrain	66 (0.9)	448 (1.3)	34 (0.9)	460 (4.3)	12 (4.6)		
Kuwait	71 (1.8)	336 (3.8)	29 (1.8)	398 (8.6)	62 (7.5)		
Oman	83 (1.0)	427 (2.7)	17 (1.0)	420 (5.2)	7 (5.6)		
Qatar	42 (1.5)	396 (4.0)	58 (1.5)	473 (3.7)	77 (4.6)		
Saudi Arabia	87 (1.1)	379 (4.3)	13 (1.1)	419 (6.3)	40 (7.1)		
United Arab Emirates	37 (1.0)	398 (3.0)	63 (1.0)	486 (2.6)	87 (3.5)		
Gulf Average	**65 (1.2)**	**397 (3.2)**	**35 (1.3)**	**442 (5.5)**	**47 (5.7)**		



Notes. Own calculations based on the information whether the father was born in country.
() Standard errors appear in parenthesis
Bars in dark color indicate statistically significant differences

Table 9-4: TIMSS 2015 average science achievement by nationality status

Country	Nationals		Non-Nationals		Difference (Absolute Value)	Difference	
	Percent of Students	Average Science Score	Percent of Students	Average Science Score		Nationals Scored Higher	Non-Nationals Scored Higher
Bahrain	66 (0.9)	455 (2.9)	34 (0.9)	468 (5.3)	13 (6.1)		
Kuwait	72 (1.8)	321 (5.5)	28 (1.8)	385 (11.1)	64 (9.4)		
Oman	83 (1.0)	432 (3.4)	17 (1.0)	429 (5.6)	4 (6.2)		
Qatar	42 (1.5)	389 (4.5)	58 (1.5)	474 (4.1)	86 (4.9)		
Saudi Arabia	87 (1.1)	385 (5.2)	13 (1.1)	432 (6.5)	47 (7.1)		
United Arab Emirates	37 (1.0)	384 (3.6)	63 (1.0)	494 (2.8)	110 (3.9)		
Gulf Average	**65 (1.3)**	**394 (4.3)**	**35 (1.3)**	**447 (6.5)**	**54 (6.5)**		



Notes. Own calculations based on the information whether the father was born in country.
() Standard errors appear in parenthesis
Bars in dark color indicate statistically significant differences

However, non-nationals in the region are rarely given citizenship (please refer to APPENDIX B for more details). The data shows that mathematics and science achievement is significantly higher for non-national students in all countries except Oman. For mathematics, significant differences range from 12 score points in Bahrain to 87 score points in the United Arab Emirates. The absolute differences in most countries are a bit higher yet for science, reaching up to

110 score points in the United Arab Emirates. The effects of nationality differences on the countries' overall mathematics or science achievement heavily depend on the percentage of non-national students. While for example the country average in Bahrain and the United Arab Emirates for mathematics is about the same (451 vs. 452 score points), we can see that the national populations in both countries differ by about score 50 points, in favor of Bahrain. The United Arab Emirates, on the other hand, has a far higher share of higher-achieving non-nationals, which outweighs the weaker performance of the nationals.

Gender by Nationality

Additionally, the question of whether gender differences are of equal magnitude for nationals and for non-nationals was explored. The results can be retrieved from Table 9-5 and Table 9-6.

Analyses distinguishing national and non-national populations, in terms of their gender gaps, showed a somewhat more differentiated picture: while for mathematics, gender differences for non-nationals are only significant for Saudi Arabia (27 score points), they are significant for all national populations except the United Arab Emirates. In science, where gender differences overall are higher, girls perform significantly better in all national populations, and in three of the immigrant populations. While beyond the scope of the current study, further research should be conducted to explore the interaction effects between nationality and gender variables in the region (for example, by means of an analysis of variance), given the importance of the variables and the large disparities in the region indicated by the data listed above.

Table 9-5: TIMSS 2015 average mathematics achievement by nationality status and gender

Country	Nationals		Difference (Absolute Value)	Non-Nationals		Difference (Absolute Value)
	Boys	Girls		Boys	Girls	
Bahrain	437 (1.9)	458 (1.8)	21 ▲	458 (6.5)	461 (3.5)	3
Kuwait	330 (5.4)	342 (4.4)	12 ▲	392 (9.3)	404 (9.0)	12
Oman	415 (3.0)	440 (3.3)	25 ▲	418 (6.0)	422 (6.1)	4
Qatar	388 (6.4)	404 (5.0)	16 ▲	477 (5.3)	469 (4.7)	9
Saudi Arabia	358 (6.6)	402 (5.0)	43 ▲	404 (9.3)	431 (7.4)	27 ▲
United Arab Emirates	392 (3.4)	404 (5.3)	12	485 (3.7)	486 (3.0)	1
Gulf Average	**387 (4.8)**	**408 (4.3)**		**439 (7.0)**	**446 (6.0)**	

Notes. Own calculations
() Standard errors appear in parenthesis
▲ Difference statistically significant (always in favor of girls)

Table 9-6: TIMSS 2015 average science achievement by nationality status and gender

Country	Nationals		Difference (Absolute Value)	Non-Nationals		Difference (Absolute Value)
	Boys	Girls		Boys	Girls	
Bahrain	433 (4.0)	478 (3.3)	44 ▲	455 (7.3)	480 (5.3)	25 ▲
Kuwait	303 (7.5)	338 (6.9)	35 ▲	378 (12.3)	393 (12.2)	15
Oman	414 (3.8)	451 (3.9)	37 ▲	426 (7.2)	431 (5.9)	5
Qatar	367 (6.8)	409 (5.5)	43 ▲	470 (6.3)	478 (6.0)	8
Saudi Arabia	346 (7.9)	427 (5.8)	81 ▲	401 (10.1)	456 (7.1)	55 ▲
United Arab Emirates	371 (4.1)	398 (5.6)	27 ▲	489 (4.0)	500 (3.3)	10 ▲
Gulf Average	**372 (5.9)**	**417 (5.3)**		**437 (8.3)**	**457 (7.2)**	

Notes. Own calculations
() Standard errors appear in parenthesis
▲ Difference statistically significant (always in favor of girls)

Achievement by school types

Table 9-7 and Table 9-8 list the average mathematics and science achievement of non-nationals and nationals in non-nationals-only schools, mixed schools, and national-only schools, respectively. The data shows that most of the nationals, and also the non-nationals, attend mixed schools. This is an interesting finding, as literature often reports on the difficulties faced by non-nationals in accessing public school systems in the GCC countries (see for example Ardent, 2015). Significant differences are marked in bold: for mathematics, analyses show that nationals perform significantly better when being enrolled in national-only schools for Bahrain and Kuwait. A similar pattern can be seen in Bahrain and Qatar in science. In the United Arab Emirates, non-nationals perform significantly better in non-nationals-only schools.

Table 9-7: TIMSS 2015 average mathematics achievement by type of school (mixed versus segregated in terms of immigrant status)

Country	Non-Nationals-Only Schools				Mixed Schools								Nationals-Only Schools			
	Non-Nationals				Nationals				Non-Nationals				Nationals			
	Percent		Score		Percent		Score		Percent		Score		Percent		Score	
Bahrain	4	(0.1)	454	(4.5)	91	(0.2)	**446**	(1.4)	96	(0.1)	460	(4.5)	9	(0.2)	**462**	(4.0)
Kuwait	4	(4.6)	368	(1.2)	91	(2.5)	**338**	(4.2)	96	(4.6)	399	(9.5)	9	(2.5)	**313**	(7.1)
Oman	17	(9.7)	415	(18.6)	74	(2.7)	427	(3.4)	83	(9.7)	421	(5.3)	26	(2.7)	427	(6.1)
Qatar	6	(1.7)	503	(18.9)	96	(2.3)	395	(4.2)	94	(1.7)	471	(3.8)	4	(2.3)	415	(26.5)
Saudi Arabia					68	(3.6)	378	(5.3)	100	(0.0)	419	(6.3)	32	(3.6)	382	(8.5)
United Arab Emirates	18	(1.6)	**520**	(6.5)	95	(1.6)	398	(3.0)	82	(1.6)	**478**	(2.8)	5	(1.6)	402	(17.6)
Gulf Average	**10**	**(4.5)**	**452**	**(11.3)**	**86**	**(2.4)**	**397**	**(3.8)**	**92**	**(4.5)**	**441**	**(5.8)**	**14**	**(2.4)**	**400**	**(14.1)**

Notes. Own calculations
() Standard errors appear in parenthesis
Significant differences are marked in bold

Table 9-8: TIMSS 2015 average science achievement by type of school (mixed versus segregated in terms of immigrant status)

Country	Non-Nationals-Only Schools				Mixed Schools								Nationals-Only Schools			
	Non-Nationals				Nationals				Non-Nationals				Nationals			
	Percent		Score		Percent		Score		Percent		Score		Percent		Score	
Bahrain	4	(0.1)	471	(8.3)	91	(0.2)	**454**	(3.0)	96	(0.1)	468	(5.5)	9	(0.2)	**470**	(6.7)
Kuwait	4	(4.7)	376	(13.2)	91	(2.5)	322	(5.8)	96	(4.7)	385	(11.8)	9	(2.5)	313	(14.2)
Oman	17	(9.7)	439	(18.6)	74	(2.7)	431	(4.2)	83	(9.7)	427	(5.9)	26	(2.7)	435	(7.6)
Qatar	6	(1.7)	509	(20.1)	96	(2.3)	**388**	(4.7)	94	(1.7)	472	(4.3)	4	(2.3)	**412**	(7.6)
Saudi Arabia					68	(3.6)	389	(6.2)	100	(0.0)	432	(6.5)	32	(3.6)	378	(10.3)
United Arab Emirates	18	(1.6)	**538**	(5.7)	95	(1.6)	384	(3.3)	82	(1.6)	**485**	(3.0)	5	(1.6)	390	(25.8)
Gulf Average	**10**	**(4.5)**	**466**	**(13.1)**	**86**	**(2.4)**	**395**	**(4.7)**	**92**	**(4.5)**	**445**	**(6.8)**	**14**	**(2.4)**	**400**	**(13.7)**

Notes. Own calculations
() Standard errors appear in parenthesis
Significant differences are marked in bold

Teachers' Gender

Table 9-9 and Table 9-10 list the results for teacher gender in relation to single-sex and mixed-gender schools. We can see that children in single-sex schools are usually taught by teachers of their gender. In mixed schools, children are predominantly taught by female teachers. The data shows a heterogeneous pattern in terms of achievement associated with teacher gender, but percentages of male teachers in girls' schools and vice versa are often quite low. For mathematics, only one significant relation with outcomes was found: in Saudi Arabia, where male teachers outperform female teachers by 48 score points in boys' schools. For science, Kuwaiti female teachers significantly outperform male teachers in boys' schools and in mixed schools.

Table 9-9: TIMSS 2015 average mathematics achievement by teachers' gender and school type (mixed versus single-sex)

Country	Boys Schools				Girls Schools				Mixed Schools			
	Male Teachers		Female Teachers		MaleTeachers		Female Teachers		MaleTeachers		Female Teachers	
	Percent	Score	Percent	Score	Percent	Score	Percent	Score	Percent	Score	Percent	Score
Bahrain	50 (1.2)	436 (3.7)	50 (1.2)	432 (2.2)			100 (0.0)	458 (1.6)	17 (2.9)	479 (15.4)	83 (2.9)	464 (6.5)
Kuwait	16 (4.5)	319 (18.6)	84 (4.5)	333 (6.1)			100 (0.0)	348 (6.8)	17 (5.6)	435 (24.0)	83 (5.6)	392 (14.7)
Oman	37 (25.2)	346 (39.2)	63 (25.2)	410 (23.1)			100 (0.0)	424 (17.2)	2 (0.7)	406 (24.7)	98 (0.7)	427 (2.6)
Qatar	16 (4.1)	386 (19.7)	84 (4.1)	403 (8.7)	1 (1.0)	414 (3.5)	99 (1.0)	413 (5.4)	24 (4.0)	492 (8.4)	76 (4.0)	494 (6.5)
Saudi Arabia	95 (2.4)	**365** (6.8)	5 (2.4)	**317** (13.0)	3 (1.6)	415 (7.1)	97 (1.6)	405 (4.5)				
United Arab Emirates	20 (3.5)	415 (20.6)	80 (3.5)	424 (7.1)	4 (2.1)	444 (37.8)	96 (2.1)	425 (6.3)	13 (1.9)	499 (12.5)	87 (1.9)	489 (3.4)
Gulf Average	**39 (10.7)**	**378 (21.4)**	**61 (10.7)**	**386 (12.0)**	**3 (1.2)**	**424 (15.8)**	**99 (1.2)**	**412 (8.5)**	**15 (3.2)**	**462 (16.6)**	**85 (3.2)**	**453 (7.3)**

Notes.
() Standard errors appear in parenthesis
Statistically significant differences between male and female teachers are marked in bold

Table 9-10: TIMSS 2015 average science achievement by teachers' gender and school type (mixed versus single-sex)

Country	Boys Schools				Girls Schools				Mixed Schools			
	Male Teachers		Female Teachers		MaleTeachers		Female Teachers		MaleTeachers		Female Teachers	
	Percent	Score	Percent	Score	Percent	Score	Percent	Score	Percent	Score	Percent	Score
Bahrain	49 (1.3)	437 (5.7)	51 (1.3)	432 (4.0)	2 (1.2)	410 (40.1)	98 (1.2)	481 (3.3)	18 (3.2)	485 (18.1)	82 (3.2)	459 (8.4)
Kuwait	16 (4.6)	**264** (19.2)	84 (4.6)	**312** (10.5)			100 (0.0)	346 (9.1)	11 (3.6)	**428** (26.5)	89 (3.6)	**378** (15.0)
Oman	14 (15.2)	338 (6.7)	86 (15.2)	359 (33.7)			100 (0.0)	417 (14.6)	2 (0.6)	399 (25.6)	98 (0.6)	433 (3.2)
Qatar	13 (3.4)	388 (15.4)	87 (3.4)	383 (11.0)	1 (0.8)	399 (3.1)	99 (0.8)	423 (5.9)	20 (4.6)	476 (13.1)	80 (4.6)	491 (6.4)
Saudi Arabia	95 (1.6)	350 (7.6)	5 (1.6)	383 (31.6)	3 (0.8)	399 (26.9)	97 (0.8)	432 (5.5)				
United Arab Emirates	17 (3.3)	383 (24.2)	83 (3.3)	417 (9.0)	2 (0.9)	372 (54.9)	98 (0.9)	425 (7.8)	9 (1.9)	519 (12.1)	91 (1.9)	498 (4.1)
Gulf Average	**34 (6.8)**	**360 (14.9)**	**66 (6.8)**	**381 (20.2)**	**2 (0.8)**	**395 (29.9)**	**99 (0.8)**	**421 (8.5)**	**12 (2.8)**	**461 (18.3)**	**88 (2.8)**	**452 (7.8)**

Notes.
() Standard errors appear in parenthesis
Statistically significant differences between male and female teachers are marked in bold

9.2 Variable Selection and Categorization

As a first step, all TIMSS 2015 background questionnaire items were compared to the framework specified in chapter 6, and all questions related to the main factors of the model (*input*, *quality*, *time*, and *opportunity*) for each of the different levels were selected for further analyses if a correspondence was found. Table 9-11 shows an overview of the TIMSS 2015 questionnaire options that were categorized according to the framework specified. The "Question" column lists the corresponding TIMSS background questionnaire location, with "SCQ-" referring to the school questionnaire, "TQ-" to the teacher questionnaire, "SQ-" to the student questionnaire, and "HQ-" to the home learning survey questionnaire. General questionnaire items in the teacher and student questionnaires are indicated by the letter "G", while items related to mathematics learning are marked by the letter "M", and science-specific items by the letter "S". An overview on all questionnaire items and detailed results of the categorization process can be found in APPENDIX A. APPENDIX B gives further details on the recoding procedures for variables and indicators mentioned in the "Comments" column of the table below. Altogether, 9 questions from the school questionnaire, 24 questions from the teacher questionnaire, 10 questions from the student questionnaire, and 7 questions from the home questionnaire were initially regarded as suitable and kept for further analyses. For the indicator of immigrant status (see APPENDIX B), eventually only data from the fathers' birth location was used, resulting in 9 questions finally being used from the student questionnaire and another 6 from the home questionnaire. The learning activities in mathematics and science (questions MS2 & MS5) were only used as course-level aggregates for the factor *quality of teaching*. For the sake of comparability between the mathematics and science models, two questions which were only available for science, namely SCQ-12 ("availability of a science laboratory") and some additional options in question TQ-S3 related to the concept of *cognitive activation* were not included in the current study. On all three educational levels covered by the TIMSS background questionnaires, indicators for the main factors of the specified framework could be identified, except for *quality of instruction* on school level.

Table 9-11: Questions and options selected from the TIMSS 2015 questionnaires

Level	Factor	Factor - Details	Question	Description	Comments (see App. B for more details)
School	Input	Resources	SCQ-11	Number of computers in school	
			SCQ-13/A	Availability of school library and # of books	
			SCQ-14	Shortage of resources (M/S)	
	Quality	Environment (SLE)	SCQ-15A-E,K-M	Emphasis on academic succes	
			SCQ-16D-J	School discipline and safety	
	Time		SCQ-8A/B/C	Instructional time	
			SCQ-16A/B & 17A/B	Problems with absenteeism	
	Opportunity		SCQ-10A/B	Policies related to tracking	
Course	Input	Teacher background	TQ-G1	Teaching experience (years)	
			TQ-G2	Gender of teacher	
			TQ-G4	Teacher's highest education level	Recoded to years of schooling
			TQ-G5A/B	Teacher majored in edu. and subject (M/S)	
			TQ-M2/S2	Confidence in teaching (M/S)	
			TQ-M10/S9	Time spent on professional development (M/S)	
			TQ-M11/S10	Preparedness to teach subject (M/S)	
	Quality of Instruction	Structured teaching	SQ-MS2 & MS5 (A/B/E/F/I)	Clear and structured teaching (M/S)	(student perception - course average)
		Activation	TQ-G14	Cognitive activation	
		Management	TQ-G15D	Limitation of teaching (disruptive students)	
		Climate	TQ-G6	Emphasis on academic success	
			TQ-7D-H	Orderly learning environment	
		Assessment	TQ-M7C/S6C	Verification of homework assignment (M/S)	Recoded to indicator
			TQ-M8A/S7A	Monitoring progress	
	Time		TQ-M1/S01B	Teaching time spent on subject	Recoded to hours per week
			TQ-M7A/B & S6A/B	Amount of homework assigned (M/S)	Indicator
	Opportunity		TQ-M6/S5	Number of topics covered (M/S)	
Student	Student Characteristics	Student Background	SQ-G1	Gender	
			HQ13/SQ-G4	Books at home	
			SQ-G5	Home possessions	
			HQ-20A/B	Highest level of parental education	Converted into years of schooling
			HQ-23A/B	Highest occupational level	Converted into HISEI
			HQ17A/B & SQG6A/B	Parents born in country	Used to create indicator of nationality status
		Aptitude	HQ-8A-C	Early numeracy activities (number sense)	
		Subject motivation	SQ-MS1/MS4	Student likes learning (M/S)	
	Time		SQ-G8	Student's absence from school	Recoded into number of absences per month
	Opportunity		HQ-9BB	Parental help with homework	Recoded to times per week

Note. SCQ – School questionnaire/ TQ –Teacher questionnaire / SQ – Student questionnaire /HQ – Home questionnaire

9.3 Results from the Principal Component and Reliability Analyses

After the categorization of variables related to the main factors of the framework, several statistical analyses were conducted to address the first part of the first research question, which addressed the extent that TIMSS 2015 questionnaires can be used to reflect effectiveness-enhancing factors in the GCC countries, based on the framework defined. Data files had to initially be prepared for analyses by merging different file types, recoding all original variables according to the expected positive association with student achievement, and implementing multiple imputation procedures for the missing cases (see section 8.3 for more details on the data preparation procedures). Once the data was prepared, for most Likert scale questions, principal component analyses (as described in section 8.3.6.1) were applied in order to reduce the number of variables to a smaller number of underlying factors main factors. For three indicators, a different approach was taken by creating a sum score of the different item options. This approach was performed in order to weight each option equally, and concerns the following questions: *teacher's preparedness to teach* (questions TQ-M11/TQ-S10), *topics already taught* (questions TQM-6/TQS-5), and finally, for an indicator related to the *verification of homework* (questions TQ-M7/S6). Details about indicator creation can be found in APPENDIX B, where additional variables that were converted to different units for easier interpretation of the multilevel analyses are also described.

The KMO criterion and Bartlett's test of sphericity was examined for each of the factors. All of the factors showed KMO values above the minimum of 0.5 and significant results for Bartlett's test of sphericity. In a few cases, single items were removed due to low communalities (below 0.3), or due to double-loadings on different factors. Once a stable and meaningful factor solution was obtained, the internal consistency among the items of each constructed scale was measured using Cronbach's alpha (see 8.3.6.2 for more details). Except for the index on *homework verification*, a Cronbach's alpha of at least 0.7 was maintained; due to the exploratory nature of the analyses, however, the *homework verification*, which exhibited a Cronbach's alpha of 0.56 for mathematics and 0.64 for science was also kept. In few cases, single items had to be removed from the scale to enhance scale reliability. Additionally, the eigenvalue of each factor is listed. As the principal component analyses were conducted on standardized variables, deviding the eigenvalue of a factor by the number of variables will result in the proportion of variance explained by that factor ("Principal Component Analysis | SPSS Annotated Output").

The following section shows the final factor solution and describes differences from the initial solution. The factor and reliability analyses were calculated separately for each of the five imputed datasets. Results presented below show the average values for the five imputed datasets.

9.3.1 Student level

On student level, data from three questions of the student questionnaire, consisting of multiple options in Likert scale format, and four questions from the parent questionnaire were analyzed by means of factor and reliability analyses to create four student background factors altogether, which were kept for the subsequent analyses steps.

For each factor, a corresponding table is displayed which shows the factor name, the value for the KMO criterion, the eigenvalue (EIGEN) for the factor, and Cronbach's Alpha (ALPHA) in the table header. For each variable included in the factor, the following information is displayed: the variable name, a short description obtained from the TIMSS codebook, and the factor loading. Variable names starting with the letter "J" generally indicate that missing values for this variable were imputed. Original variable names begin with the letter "A" to denote TIMSS grade four (= population A).

Economic and social cultural status (ESCS)

Table 9-12: Economic and social cultural status

Variable	F_ESCS KMO: 0.65/ EIGEN: 2.37/ ALPHA: 0.70 Label	Factor loading
Zscore(JSBH20A)	GEN\LVL OF EDUCATION\FATHER	0.81
Zscore(JSBH20B)	GEN\LVL OF EDUCATION\MOTHER	0.82
Zscore(JSBH23A)	GEN\WHAT KIND OF MAIN JOB\FATHER	0.69
Zscore(JSBH23B)	GEN\WHAT KIND OF MAIN JOB\MOTHER	0.67
Zscore(JBOOKS)	GEN\AMOUNT OF BOOKS AT HOME	0.36

The creation of the indicator for the economic and social cultural status is described in more detail in section 8.4. It contains the Z-scores of the parental occupation (JSBH23A & B) converted into ISEI scores, the highest parental education level (JSBH20A & B) converted into years of schooling, and, as an indicator for home possessions, the number of books at home (JBOOKS).

Table 9-12 shows that all factor loadings are well above 0.3, and a value of 0.70 for Cronbach's alpha (a coefficient used to measure the internal consistency of the items by their inter-item correlation) is acceptable.

Early numeracy skills

Table 9-13: Early numeracy skills

Variable	F_EARLYNUM KMO: 0.74/ EIGEN: 2.56/ ALPHA: 0.91 Label	Factor loading
JSBH08A	GEN\SKILLS\COUNT BY HIM-/HERSELF	0.90
JSBH08B	GEN\SKILLS\RECOG WRITTEN NUMERAL	0.94
JSBH08C	GEN\SKILLS\WRITE NUMBERS	0.93

The three Likert scale items related to students' general numeracy skills before attending primary schools, shown in Table 9-13, were combined to form a proxy of the student's *aptitude*. All items show high factor loadings, and the internal consistency of the construct (0.74) is acceptable.

Subject motivation for mathematics

Table 9-14: Subject motivation mathematics

Variable	F_MOTIV_M KMO: 0.93/ EIGEN: 4.49/ ALPHA: 0.91 Label	Factor loading
JSBM01A	MAT\AGREE\ENJOY LEARNING MATHEMATICS	0.79
JSBM01D	MAT\AGREE\LEARN INTERESTING THINGS	0.76
JSBM01E	MAT\AGREE\LIKE MATHEMATICS	0.87
JSBM01F	MAT\AGREE\SCHOOLWORK INVOLVES NUMBERS	0.74
JSBM01G	MAT\AGREE\LIKE MATH PROBLEMS	0.82
JSBM01H	MAT\AGREE\LOOK FORWARD TO MATH LESSONS	0.81
JSBM01I	MAT\AGREE\MATH FAVORITE SUBJECT	0.82

Altogether seven out of the nine items available from question SQ-MS1, which measured students' interest in mathematics, were used to create an indicator of the students' subject motivation related to mathematics. Options b) and c) were negatively worded, and both loaded during the initial analysis step to a separate factor. In line with Scherer and Nilsen (2016), who also found negatively-worded motivation items to load on a "substantially different" construct than positively-worded items, those items were dropped to avoid method bias and construct-irrelevant multidimensionality (Scherer & Nilsen, 2016, p. 61). Results are shown in Table 9-14. The internal consistency of the remaining items (0.91) can be regarded as excellent.

Subject motivation for science

Table 9-15: Subject motivation science

Variable	F_MOTIV_S KMO: 0.91/ EIGEN: 4.27/ ALPHA: 0.89 Label	Factor loading
JSBS04A	SCI\AGREE\ENJOY LEARNING SCIENCE	0.74
JSBS04D	SCI\AGREE\WISH HAVE NOT TO STUDY SCIENCE	0.78
JSBS04E	SCI\AGREE\SCIENCE IS BORING	0.86
JSBS04F	SCI\AGREE\LOOK FORWARD TO LEARN	0.81
JSBS04G	SCI\AGREE\SCIENCE TEACHES ME	0.74
JSBS04H	SCI\AGREE\SCIENCE EXPERIMENTS	0.73
JSBS04I	SCI\AGREE\FAVORITE SUBJECT	0.80

Table 9-15 displays the corresponding results for science. While the wording for most of the items from question SQ-MS4 correspond to their mathematics equivalents, options f) and g) are somewhat different. Similar to the corresponding mathematics question, options b) and c)

are negatively worded, and were ultimately removed for the final factor on science subject motivation. The internal consistency for the science motivation construct (0.89) is also close to excellent.

9.3.2 Course level

On course level, data from nine Likert scale questions of the teacher questionnaire related to the proposed framework were analyzed by means of factor and reliability analyses. Additionally, items from two questions on the student questionnaire (MS2 for mathematics and MS5 for science) were aggregated on course level, and used as an indicator for clear and structured instruction. Seven factors in total were kept for subsequent mathematics and science analyses.

Confidence in teaching methods for mathematics

Table 9-16: Confidence in teaching – mathematics

Variable	F_CONFIDENCE_M KMO: 0.92/ EIGEN: 4.23/ ALPHA: 0.87 Label	Factor loading
JTBM02A	MAT\CONFIDENT\INSPIRE STUDENTS	0.64
JTBM02B	MAT\CONFIDENT\VARIETY PROBLEM SOLVING STRATEGIES	0.73
JTBM02C	MAT\CONFIDENT\CHALLENGING TASKS	0.69
JTBM02D	MAT\CONFIDENT\ENGAGE STUDENTS INTEREST	0.76
JTBM02E	MAT\CONFIDENT\APPRECIATE MATH	0.77
JTBM02G	MAT\CONFIDENT\IMPROVE UNDERSTANDING	0.66
JTBM02H	MAT\CONFIDENT\MAKE MATH RELEVANT	0.77
JTBM02I	MAT\CONFIDENT\DEVELOP HIGHER THINKING	0.78

For creating the *confidence in teaching* methods scale displayed in Table 9-16, eight of the nine options from question TQ-M2 of the teacher questionnaire were combined. Option f) *Assessing student comprehension* was not regarded as fitting conceptually to the scale, and therefore was excluded. The overall reliability of the scale (0.87) can be considered as good.

Confidence in teaching methods for science

Table 9-17: Confidence in teaching – science

Variable	F_CONFIDENCE_S KMO: 0.93/ EIGEN: 4.83/ ALPHA: 0.90 Label	Factor loading
JTBS02A	SCI\CONFIDENT\INSPIRE STUDENTS	0.68
JTBS02B	SCI\CONFIDENT\EXPLAIN CONCEPTS	0.66
JTBS02C	SCI\CONFIDENT\CHALLENGING TASKS	0.71
JTBS02D	SCI\CONFIDENT\ENGAGE STUDENTS INTEREST	0.79
JTBS02E	SCI\CONFIDENT\APPRECIATE SCIENCE	0.75
JTBS02G	SCI\CONFIDENT\IMPROVE UNDERSTANDING	0.72
JTBS02H	SCI\CONFIDENT\MAKE SCIENCE RELEVANT	0.76
JTBS02I	SCI\CONFIDENT\DEVELOP HIGHER THINKING	0.79
JTBS02J	SCI\CONFIDENT\TEACH USING INQUIRY	0.72

Similar to the scale for mathematics, option f) was not regarded as fitting well with the concept to be measured, and therefore was not included in the confidence scale for science. Table 9-17 shows a Cronbach's alpha of 0.90 for the final scale, which can be rated as excellent.

Emphasis on academic success (teacher level)

Table 9-18: Emphasis on academic success

Variable	F_TCH_EAS KMO: 0.86/ EIGEN: 3.34/ ALPHA: 0.84 Label	Factor loading
JTBG06A	GEN\CHARACTERIZE\TCHS UNDERSTANDING	0.77
JTBG06B	GEN\CHARACTERIZE\TCHS DEGREE OF SUCCESS	0.80
JTBG06C	GEN\CHARACTERIZE\TCHS EXPECTATIONS	0.68
JTBG06D	GEN\CHARACTERIZE\TCHS WORKING TOGETHER	0.77
JTBG06E	GEN\CHARACTERIZE\TCHS ABILITY TO INSPIRE	0.77
JTBG06O	GEN\CHARACTERIZE\COLLABORATION TO PLAN	0.67

Emphasis on academic success on classroom level can be regarded as an important indicator for the learning environment of the learning group, and was shown in many studies to be positively related with student achievement. The scale created for the current study, as displayed Table 9-18, includes items indicating a positive school climate between the main actors of the school (students, teachers, and school management). As the parents usually do not belong to the main actors for effective instruction, and research about the relation between parental involvement and student achievement is somewhat inconclusive (see section 3.3.6.4 for more details), items concerning parental commitment, involvement, and the like were not included in this scale. The overall scale reliability can be judged as good.

Clear and structured instruction in mathematics

Table 9-19: Clear and structured instruction – mathematics

Variable	F_CLEARST_M KMO: 0.85/ EIGEN: 3.18/ ALPHA: 0.91 Label	Factor loading
JSBM02B_clsX	MAT\AGREE\TEACHER IS EASY TO UNDERSTAND	0.86
JSBM02E_clsX	MAT\AGREE\CLEAR ANSWERS	0.92
JSBM02F_clsX	MAT\AGREE\TEACHER EXPLAINS GOOD	0.92
JSBM02I_clsX	MAT\AGREE\HOW TO DO BETTER	0.87

As no well-suited items could be identified to measure this important subdimension of instructional quality on teacher level, the author decided to use four items from the student questionnaire, and to aggregate this information on course level, to avoid construct underrepresentation. The results are shown in Table 9-19. Advantages and disadvantages of student assessment of instructional quality are briefly discussed in section 8.3.2. As option a) *I know what my teacher expects me to do* reduced the overall scale reliability, it was ultimately removed. While theoretically fitting to the construct, the option is possibly a bit more vague in its conception, and therefore might not measure the concept as well as the other options. The final factor scale shows an excellent internal consistency of 0.91.

Clear and structured instruction in science

Table 9-20: Clear and structured instruction – science

Variable	F_CLEARST_S KMO: 0.86/ EIGEN: 3.25/ ALPHA: 0.93 Label	Factor loading
JSBS05B_clsX	SCI\AGREE\TEACHER EASY TO UNDERSTAND	0.88
JSBS05E_clsX	SCI\AGREE\CLEAR ANSWERS	0.92
JSBS05F_clsX	SCI\AGREE\TEACHER EXPLAINS GOOD	0.92
JSBS05I_clsX	SCI\AGREE\HOW TO DO BETTER	0.89

Table 9-20: Clear and structured instruction – science

Table 9-20 shows the same four items from the student questionnaire, but here related to science lessons (question MS5) which again were aggregated on course level. Likewise as done for mathematics, removing option a) improved the scale reliability for science to 0.93, which can be judged as excellent.

Cognitive activation

Table 9-21: Cognitive activation

Variable	F_COGNTACTIV KMO: 0.86/ EIGEN: 3.23/ ALPHA: 0.79 Label	Factor loading
JTBG14A	GEN\HOW OFTEN\DAILY LIVES	0.56
JTBG14B	GEN\HOW OFTEN\EXPLAIN ANSWERS	0.65
JTBG14C	GEN\HOW OFTEN\BRING INTERESTING MATERIAL	0.61
JTBG14D	GEN\HOW OFTEN\BEYOND INSTRUCTION	0.58
JTBG14E	GEN\HOW OFTEN\CLASSROOM DISCUSSION	0.68
JTBG14F	GEN\HOW OFTEN\LINK KNOWLEDGE	0.63
JTBG14G	GEN\HOW OFTEN\PROBLEM SOLVING PRCDS	0.68
JTBG14H	GEN\HOW OFTEN\EXPRESS IDEAS	0.67

The TIMSS 2015 teacher background questionnaire collected a number of items (most of them in question TQ-14) that can be seen as relating to more constructivist theories of student-centered instruction approaches. In the current study, such teaching approaches are subsumed under the term *cognitive activation* (see section 3.3.5.2. for more information). Table 9-21 lists the results obtained from the eight question items of question 14, which demonstrate an internal consistency of close to good (0.79) while factor loadings in general are somewhat lower compared to other scales.

Class learning environment

Table 9-22: Class environment

Variable	F_ENVIRONM KMO: 0.85/ EIGEN: 3.55/ ALPHA: 0.90 Label	Factor loading
JTBG07D	GEN\THINKING ABT CURR SCH\STUD BEHAVE	0.86
JTBG07E	GEN\THINKING ABT CURR SCH\STUD RESPECT	0.85
JTBG07F	GEN\THINKING ABT CURR SCH\RESPECT PROPERTY	0.87
JTBG07G	GEN\THINKING ABT CURR SCH\SCH CLEAR RULES	0.81
JTBG07H	GEN\THINKING ABT CURR SCH\RULES ENFORCED	0.82

The classroom climate or environment also proved to be an important precondition for effective teaching (see section 3.3.5.3 for more details). Five of the eight items in question TQ-G7, which relates to school environment aspects, were used to build a proxy for the classroom environment, as shown in Table 9-22. In a first analysis step, option b), which relates to the *schools' security policies* was included, but was ultimately removed due to double-loadings with a second factor. It should be noted that the question TQ-G7 asked teachers to consider the situation

in the school – and not only the situation in the teacher's actual course. However, it is hypothesized here that the teacher will be influenced by the experiences from his or her own course(s), and his or her answers are therefore sufficiently valid when used as a proxy for the climate on course level. The overall reliability of this scale (0.90) is excellent.

9.3.3 School level

On school level, three Likert scale questions from the principal questionnaire were analyzed by means of factor and reliability analyses to create altogether five school-level mathematics and science factors that were kept for the subsequent analyses steps.

Shortage in mathematics learning resources

Table 9-23: Shortage in mathematics resources

Variable	F_SHORTAGE_M KMO: 0.79/ EIGEN: 3.14/ ALPHA: 0.85 Label	Factor loading
JCBG14BA	GEN\SHORTAGE\MAT\TEACH SPEC MATH	0.78
JCBG14BB	GEN\SHORTAGE\MAT\COMPUTER SOFTWARE	0.83
JCBG14BC	GEN\SHORTAGE\MAT\LIBRARY RESOURCES	0.81
JCBG14BD	GEN\SHORTAGE\MAT\CALCULATORS	0.70
JCBG14BE	GEN\SHORTAGE\MAT\CONCRETE OBJECTS	0.83

Table 9-23 shows the five options from part B of question SCQ-14, which asked about the extent to which mathematics instruction is affected by certain shortages. All five options were selected to create a proxy indicating the availability of learning resources. The availability of sufficient and appropriate learning resources might have greater importance as a precondition for effective instruction, especially for developing countries (see also section 3.3.6.5). The scale reliability with a Cronbach's Alpha of 0.85 can be considered as good.

Shortage in science learning resources

Table 9-24: Shortage in science resources

Variable	F_SHORTAGE_S KMO: 0.77/ EIGEN: 3.03/ ALPHA: 0.89 Label	Factor loading
JCBG14CA	GEN\SHORTAGE\SCI\TEACH SPEC SCIENCE	0.84
JCBG14CB	GEN\SHORTAGE\SCI\COMPUTER SOFTWARE	0.87
JCBG14CC	GEN\SHORTAGE\SCI\LIBRARY RESOURCES	0.87
JCBG14CD	GEN\SHORTAGE\SCI\SCIENCE EQUIPMENT	0.90

Part C of question SCQ-14 asked about shortages in resources related to science instruction. The four available questionnaire options, displayed in Table 9-24, were combined to create a proxy for the availability of science resources. The internal consistency of the scale (0.89) is close to excellent.

Emphasis on academic success (school level)

Table 9-25: Emphasis on academic success (school level)

Variable	F_SC_EAS KMO: 0.90/ EIGEN: 4.80/ ALPHA: 0.90 Label	Factor loading
JCBG15A	GEN\SCH CHARACTER\TCH UNDERSTANDING	0.76
JCBG15B	GEN\SCH CHARACTER\TCH SUCCESS	0.80
JCBG15C	GEN\SCH CHARACTER\TCH EXPECTATIONS	0.79
JCBG15D	GEN\SCH CHARACTER\TCH WORKING TOGETHER	0.80
JCBG15E	GEN\SCH CHARACTER\TCH ABILITY TO INSPIRE	0.80
JCBG15K	GEN\SCH CHARACTER\STD DESIRE TO DO WELL	0.75
JCBG15L	GEN\SCH CHARACTER\STD REACH GOALS	0.77
JCBG15M	GEN\SCH CHARACTER\STD RESPECT	0.71

Similar to the course level, emphasis on academic access can also be regarded as an important indicator for the learning environment on school level. The selected items from question SCQ-15 are shown in Table 9-25. The internal consistency of the scale (0.90) is excellent.

School discipline and safety

Table 9-26: School discipline and safety

Variable	F_SC_SOS KMO: 0.94/ EIGEN: 6.20/ ALPHA: 0.96 Label	Factor loading
JCBG16C	GEN\DEGREE PROBS\CLASSROOM DISTURBANCE	0.81
JCBG16D	GEN\DEGREE PROBS\CHEATING	0.89
JCBG16E	GEN\DEGREE PROBS\PROFANITY	0.89
JCBG16F	GEN\DEGREE PROBS\VANDALISM	0.92
JCBG16G	GEN\DEGREE PROBS\THEFT	0.90
JCBG16H	GEN\DEGREE PROBS\INTIMIDATION AMONG STUD	0.88
JCBG16I	GEN\DEGREE PROBS\PHYSICAL FIGHT	0.87
JCBG16J	GEN\DEGREE PROBS\INTIMIDATION OF TEACHER	0.87

In addition to establishing an atmosphere with a special emphasis on academic success, it also could be shown that an orderly atmosphere and a positive disciplinary climate are other important components of the school learning environment – and, as such, are important preconditions for effective teaching (see section 3.3.6.2 for more details). For the purpose of the current study, options c) to j) of question SCQ-16, which asked about the extent of several problems related to discipline and safety, were combined, resulting in a school discipline and safety scale. The item loadings are displayed in Table 9-26. The obtained scale shows an Cronbach's alpha value of 0.96, and thus an excellent internal consistency.

Absenteeism

Table 9-27: Absenteeism

Variable	F_ABSENCE KMO: 0.73/ EIGEN: 2.82/ ALPHA: 0.86 Label	Factor loading
JCBG16A	GEN\DEGREE PROBS\ARRIVING LATE AT SCHOOL	0.81
JCBG16B	GEN\DEGREE PROBS\ABSENTEEISM	0.85
JCBG17A	GEN\DEGREE PROBS TEACH\ARRIVING LATE	0.86
JCBG17B	GEN\DEGREE PROBS TEACH\ABSENTEEISM	0.83

Effective school policies related to the management of teaching time and absenteeism should result in a low degree of problems related to absenteeism and late arrival. The extent of perceived problems with absenteeism was explored in SCQ-16 & 17 a) – b). For the purposes of the current analyses, they were used as a proxy to indicate the extent to which related policies were available or effectively being established. Table 9-27 shows the created scale, which was related to the factor *time* on school level and shows a good internal consistency.

The factor analyses resulted in a total of four factors on student level, seven factors on course level, and five factors on school level that were kept for subsequent analyses in association with mathematics and science achievement. In addition to the scales presented in this chapter, construction of scales to measure the quality of instruction in terms of assessment from questions TQM-08/07 for mathematics and respectively TQS-07/06TQS-06 for science was attempted, but the internal scale reliability was too low (below 0.5). For cases where different items/factors or indices were available for one and the same model factor, the question of how far these could be further combined was likewise explored. Further combinations, however, usually resulted in different factors and low internal scale consistencies; hence, the different identified indicators were kept separate for the subsequent step: the correlation analysis.

9.4 Results of the Correlation Analyses

Once items were combined to scales by means of principal component analyses, and the internal consistency of the created scale or index was examined, correlation analyses were performed as a final step of the preliminary analyses. Bivariate correlations between all components kept so far, and mathematics and science achievement, respectively, were calculated. Correlations were calculated with each of the five imputed datasets, and results were later combined. For all teacher- and course-level components, the course averages were correlated against the course mean achievement in mathematics or science. The intention of this step was to create a parsimonious specification of the components to be used for the final multilevel analyses. For this purpose, the components obtained thus far were revisited once more in light of the defined theoretical framework, and the validity of the constructs was assessed by measuring their correlation with student achievement. Furthermore, inter-correlations between the components were checked in order to investigate possible issues with multicollinearity, and the author attempted to minimize redundancy between constructs measuring the same model factor without losing important relations hypothesized in the research literature. The correlation analyses allowed for the identification of the component with the strongest relation to achievement in the region between different – although from a theoretical standpoint, equally well-suited – indicators, in order to measure a certain model factor. Components weakly correlated with achievement in all countries, and for both subjects (correlation coefficient below 0.2), could also be identified and removed from further analyses to reduce model complexity. The analyses were based on a regional framework and variable selection, but the strength of correlations between components and achievement among countries often varied to a certain extent, implying that compromises were needed to maintain both a regional set of variables for comparison and the

parsimonious nature of the model. In general, the following rules for variable selection were applied: if a model factor (specified in the column "Factor – details" in the subsequent tables) of the proposed framework was represented by more than one component kept from the previous analyses steps, then only those components where the Pearson product moment coefficient between component and mathematics or science achievement in at least one country reached 0.2 (see section 8.3.6.3 for a justification of the cut-off point) were retained for multilevel modeling. If none of the selected variables or scales fulfilled this criterion, then only components shown to be relevant in the region, according to the reviewed literature, were kept.

Additionally, the inter-item correlations between all the components were examined, to explore the possible existence of multicollinearity by investigating coefficients above 0.8.

9.4.1 Student level

First, the regional results of the correlation analyses with mathematics and science achievement were examined at the student level. Table 9-28 and Table 9-29 show the correlation results of all variables selected for the multilevel analyses with mathematics and science achievement, respectively, for all six countries. The correlation analyses were calculated using the IEA IDB Analyzer (see also section 8.3.1 about the software), and were based on the student sample sizes listed in Table 7-1. All correlations between predictor variables and achievement were statistically significant, except for the *parental homework* in Oman for both subjects, and the *nationality status* in Oman for science. The data shows that for all variables, except the *parental help with homework*, the correlation with achievement, at least in one country and subject, exceeds 0.2 – which was defined as the cut-off criterion for inclusion into the multilevel analyses. However, as *parental help* was the only indicator for the *opportunity* model factor, and as in Qatar for science it showed a (albeit negative) correlation of 0.17, it was decided to keep that variable as well.

Table 9-28: Correlation of student level factors with mathematics achievement

Factor	Factor - Details	Description	Variable	BHR	KWT	OMN	QAT	SAU	ARE
Student Characteristics	Student Background	Economic, social and cultural status (ESCS)	F_ESCS	**0.29**	**0.26**	**0.25**	**0.38**	**0.15**	**0.44**
		Gender	JTSEX	**0.09**	**0.06**	**0.11**	0.01	**0.23**	0.02
		Nationality status	JNATIONAL	**0.27**	**0.23**	**0.21**	**0.35**	**0.13**	**0.41**
	Aptitude	Early numeracy activities	F_EARLYNUM	**0.13**	**0.19**	**0.16**	**0.22**	**0.12**	**0.16**
	Subject motivation	Student likes learning	F_MOTIV_M	**0.11**	**0.08**	**0.18**	**0.14**	**0.10**	**0.11**
Time		Student's absence from school	JSBG08	**-0.23**	**-0.17**	**-0.16**	**-0.24**	**-0.16**	**-0.26**
Opportunity		Parental help with homework	JSBH09BB	**-0.13**	**-0.07**	-0.03	**-0.16**	**-0.08**	**-0.17**

Notes. significant correlations (0.05 level (2-tailed)) are marked in bold
BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates

Table 9-29: Correlation of student level factors with science achievement

Factor	Factor - Details	Description	Variable	BHR	KWT	OMN	QAT	SAU	ARE
Student Characteristics	Student Background	Economic, social and cultural status (ESCS)	F_ESCS	**0.26**	**0.26**	**0.26**	**0.36**	**0.20**	**0.45**
		Gender	JTSEX	**0.19**	**0.12**	**0.13**	**0.11**	**0.34**	**0.06**
		Nationality status	JNATIONAL	**0.06**	**0.23**	-0.01	**0.38**	**0.13**	**0.44**
	Aptitude	Early numeracy activities	F_EARLYNUM	**0.12**	**0.16**	**0.16**	**0.23**	**0.16**	**0.17**
	Subject motivation	Student likes learning	F_MOTIV_S	**0.23**	**0.16**	**0.24**	**0.22**	**0.15**	**0.24**
Time		Student's absence from school	F_MOTIV_S	**-0.24**	**-0.17**	**-0.15**	**-0.28**	**-0.17**	**-0.26**
Opportunity		Parental help with homework	F_MOTIV_S	**-0.10**	**-0.06**	-0.02	**-0.14**	**-0.08**	**-0.16**

Notes. significant correlations (0.05 level (2-tailed)) are marked in bold
BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates

The correlation analyses already reveal some interesting findings. In all countries, except for Saudi Arabia and Qatar for science, the ESCS index shows the strongest correlation to both students' mathematics as well as science achievement – with the highest value in the United Arab Emirates for both subjects. In Saudi Arabia, student gender shows the highest correlation with achievement of all level 1 factors for both subjects, while in Qatar it is the student's nationality status which shows the highest correlation with science performance.

Additionally, the nationality status in Qatar and in the United Arab Emirates displays correlations of 0.35 and higher with mathematics and science achievement. While a negative association between *abseenteeism* and student outcome was expected due to the coding of the variable, interestingly, the extent of *parental help with homework* was also found to be negatively associated with student achievement.

On student level, the inter-item correlation analyses did not reveal any correlations between the different components above 0.8; hence, no measures to avoid multicollinearity between student level variables were needed.

9.4.2 Course and school level

The following section presents the results of the correlation analyses between all level 2 variables components that were retained from the previous analyses steps. Correlations with mathematics achievement are shown in Table 9-30, and the corresponding results for science achievement in Table 9-31. Level 2 correlations were calculated by correlating the course averages of the predictor variables with the course averages of the achievement scores, using the IEA IDB Analyzer. Corresponding sample sizes can be found in Table 7-1. Statistically significant associations are marked in bold.

For group-level variables, a check for which variables would not show a relation of at least 0.2, in any of the countries for any of the subjects, was administered. Such variables were then removed from further analysis steps – unless the literature review indicated a special importance of that specific component. Correlations partly differ across subjects, as can be seen when comparing the results of both tables. Altogether, the following nine variables were identified as having low positive correlations with achievement for both subjects in all countries (with some of them having unexpected higher negative correlations):

Table 9-30: Correlation of course and school level factors with mathematics achievement

Level	Factor	Factor - Details	Variable Description	Variable	BHR	KWT	OMN	QAT	SAU	ARE
School	Input	Resources	Number of computers in school	JCBG11	0.00	-0.04	0.00	0.04	0.09	**0.25**
			Availability of school library and # of books	JCBG13A	**0.25**	**0.35**	0.00	0.13	0.11	**0.46**
			Shortage of resources	F_SHORTAGE_M	**0.20**	0.07	0.03	**0.30**	-0.02	**0.29**
	Quality	Environment (SLE)	Emphasis on academic succes	F_SC_EAS	**0.23**	**0.34**	-0.01	**0.30**	**0.23**	**0.41**
			School discipline and safety	F_SC_SOS	**0.25**	0.17	-0.09	0.08	0.06	**0.30**
	Time		Instructional time	JCDG08HY	0.08	0.14	0.05	**-0.14**	-0.09	0.06
			Problems with absenteeism	F_ABSENCE	**0.19**	**0.22**	0.02	**0.14**	0.11	**0.33**
	Opportunity		Policies related to tracking	JCBG10A	**-0.18**	-0.05	**0.11**	-0.08	-0.04	-0.01
Course	Input	Teacher background	Teaching experience (years)	JTBG01	0.14	0.14	0.07	0.08	-0.11	0.02
			Gender of teacher	JTBG02	0.08	-0.05	0.09	-0.11	**0.30**	-0.01
			Teacher's highest education level	JTBG04	-0.08	0.07	**-0.26**	-0.01	0.01	0.04
			Teacher majored in edu. and subject	JTDM05	0.08	-0.04	0.05	0.07	0.00	-0.08
			Time spent on professional development	JTBM10	-0.02	0.04	-0.01	**-0.20**	0.02	**-0.15**
			Confidence in teaching	F_CONFIDENCE_M	**0.11**	0.04	0.09	-0.10	**0.25**	**0.16**
			Preparedness to teach subject	JTBM11Z	**0.12**	**0.16**	**0.11**	0.11	0.02	0.08
		Student composition	Average economic and sociocultural status	F_ESCS_clX	**0.55**	**0.63**	**0.17**	**0.76**	**0.22**	**0.76**
			Average early numeracy skills	F_EARLYNUM_clX	**0.14**	**0.55**	**0.14**	**0.41**	0.08	**0.37**
			Average gender composition	JTSEX_clX	**0.23**	**0.13**	0.09	0.04	**0.37**	0.04
			Average composition in terms of non-nationals	JNATIONAL_clX	-0.02	**0.53**	**-0.18**	**0.60**	**0.14**	**0.59**
	Quality of Instruction	Structured teaching	Clear and structured teaching	F_CLEARST_M	**0.25**	**0.17**	**0.34**	**0.46**	**0.38**	**0.41**
		Activation	Cognitive activation	F_COGNACTIV	**0.20**	0.10	0.02	0.01	**0.16**	**0.26**
		Management	Limitation of teaching (disruptive students)	JTBG15D	**0.14**	**0.21**	-0.03	**0.21**	**0.25**	**0.27**
		Climate	Emphasis on academic success	F_TCH_EAS	**0.19**	**0.15**	**0.21**	-0.04	**0.26**	**0.26**
			Orderly learning environment	F_ENVIRONM	**0.27**	0.09	0.09	**0.14**	**0.24**	**0.39**
		Assessment	Verification of homework assignment	JTBM07Z	0.05	0.01	0.10	0.03	**0.26**	**0.17**
			Monitoring progress	JTBM08A	0.08	0.05	0.02	-0.01	**0.14**	**0.10**
	Time		Teaching time spent on subject	JTBM01	0.06	**0.26**	0.01	-0.02	0.11	**0.13**
			Amount of homework assigned	JTDM07Z	-0.10	0.22	-0.07	-0.12	-0.02	**0.12**
	Opportunity		Number of topics taught	JTDM06Z	**0.13**	-0.03	**0.18**	0.10	0.14	0.01

Notes. significant correlations (0.05 level (2-tailed)) are marked in bold
BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates

Table 9-31: Correlation of course and school level factors with science achievement

Level	Factor	Factor - Details	Variable Description	Variable	BHR	KWT	OMN	QAT	SAU	ARE
School	Input	Resources	Number of computers in school	JCBG11	-0.02	-0.04	0.03	0.06	0.06	**0.28**
			Availability of school library and # of books	JCBG13A	**0.21**	**0.30**	-0.01	0.13	0.10	**0.48**
			Shortage of resources	F_SHORTAGE_S	**0.19**	0.00	0.01	**0.22**	0.01	**0.24**
	Quality	Environment (SLE)	Emphasis on academic succes	F_SC_EAS	**0.21**	**0.34**	-0.03	**0.34**	**0.27**	**0.42**
			School discipline and safety	F_SC_SOS	**0.31**	0.17	-0.09	0.09	0.08	**0.32**
	Time		Instructional time	JCDG08HY	0.05	0.07	0.01	-0.10	-0.05	0.04
			Problems with absenteeism	F_ABSENCE	**0.23**	0.15	0.02	**0.14**	0.10	**0.34**
	Opportunity		Policies related to tracking	JCBG10B	-0.08	-0.04	0.11	**-0.18**	-0.10	**-0.11**
Course	Input	Teacher background	Teaching experience (years)	JTBG01	0.05	**0.16**	0.06	0.03	**-0.19**	-0.07
			Gender of teacher	JTBG02	**0.12**	0.11	0.08	-0.05	**0.54**	0.05
			Teacher's highest education level	JTBG04	-0.08	0.20	**-0.15**	-0.01	0.04	0.10
			Teacher majored in edu. and subject	JTDS05	**0.14**	-0.05	0.04	0.14	-0.13	-0.03
			Time spent on professional development	JTBS09	-0.02	-0.05	-0.04	**-0.19**	-0.02	-0.09
			Confidence in teaching	F_CONFIDENCE_S	0.05	0.14	0.04	-0.07	0.13	**0.20**
			Preparedness to teach subject	JTBS10Z	**0.14**	-0.11	**0.11**	0.10	**0.14**	-0.09
		Student composition	Average economic and sociocultural status	F_ESCS_clX	**0.40**	**0.59**	**0.15**	**0.70**	**0.32**	**0.78**
			Average early numeracy skills	F_EARLYNUM_clX	**0.12**	**0.41**	**0.18**	**0.43**	**0.15**	**0.42**
			Average gender composition	JTSEX_clX	**0.34**	**0.23**	**0.15**	**0.18**	**0.55**	0.08
			Average composition in terms of non-nationals	JNATIONAL_clX	-0.01	**0.43**	**-0.16**	**0.57**	**0.18**	**0.65**
	Quality of Instruction	Structured teaching	Clear and structured teaching	F_CLEARST_S	**0.30**	**0.24**	**0.37**	**0.51**	**0.40**	**0.46**
		Activation	Cognitive activation	F_COGNACTIV	**0.13**	**0.18**	0.07	**0.16**	**0.26**	**0.30**
		Management	Limitation of teaching (disruptive students)	JTBG15D	**0.21**	**0.17**	0.02	**0.21**	0.12	**0.31**
		Climate	Emphasis on academic success	F_TCH_EAS	0.05	**0.20**	**0.14**	0.09	**0.34**	**0.30**
			Orderly learning environment	F_ENVIRONM	**0.24**	0.15	0.06	**0.22**	**0.20**	**0.40**
		Assessment	Verification of homework assignment	JTBS06Z	0.08	0.14	0.00	**-0.29**	**0.23**	-0.01
			Monitoring progress	JTBS07A	0.00	0.06	**0.14**	0.03	0.00	0.05
	Time		Teaching time spent on subject	JTBS01B	-0.04	**0.20**	0.11	**-0.17**	0.11	-0.04
			Amount of homework assigned	JTDS06Z	-0.01	0.09	0.00	**-0.17**	0.14	-0.03
	Opportunity		Number of topics taught	JTDS05Z	**0.17**	**-0.22**	0.03	-0.02	0.01	-0.03

Notes. significant correlations (0.05 level (2-tailed)) are marked in bold
BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates

Overall instructional time (school level)

The overall *instructional time* was derived from three questions related to the *number of days the school is open for instruction*, the *typical instructional time per day*, and the *number of days per week*. Sufficient instructional time is needed for teaching a certain course syllabus to the students; thus, an association between the total instructional time available and achievement outcomes could be assumed. As the amount of instructional time devoted to core subjects in the Gulf area used to be relatively low in international comparisons, in the last years, Gulf countries reacted and enacted reforms allotting more time for mathematics and science, such as the *daily school timing initiative* in Bahrain (Al-Awadhi, 2016, p. 8). Concerning the finding that correlations between overall *instructional time* on school level and achievement don't show the expected higher correlation results, one reason could be that the amount of official school days, and their length, might be prescribed on national level – and hence, not too many differences between schools occur. Additionally, the question in the TIMSS questionnaire only relates to the *typical* school day. School closures, or periods without formal instruction due to natural disasters, strikes, or longer school closures during national testing periods or festivities and so forth, are not explicitly covered in the questionnaire.

Tracking policy according to mathematics and science achievement, respectively

There are certain indications that tracking and streaming might influence learning opportunities of the students (see section 3.3.3), but the literature does not give clear empirical evidence for a straightforward association between tracking procedures and achievement for different student subgroups. The general "yes/no" question asked in the TIMSS questionnaire is likely not specific enough to delve deeper into this issue, and therefore doesn't allow for the explanation of possible influences of tracking policies on student achievement.

Teachers' experience in years

Findings in relation to the mere gains in student learning measured by teachers' *teaching experiences* in years seem to be somewhat mixed. While TIMSS trend analyses over all countries show that achievement was highest, especially for mathematics, for teachers with more than 20 years of experience in grade eight (Mullis, Martin, Foy, & Arora, 2012, p. 292), other authors could not confirm a relation or assumed a rather curvilinear effect. Associations between teacher background factors and student achievement are further discussed in section 3.3.5.1.

Teachers' highest education level

Interestingly, in some of the countries, the teachers' *highest education level* is negatively associated with achievement – for example, with a correlation coefficient of -0.26 for mathematics in Oman. On the contrary, analysis from Blömeke et al. (2016) indicated the teacher educational level as the strongest predictor for student achievement across the TIMSS 2011 countries. Because of the interesting and contradictory nature of relations to achievement within the GCC countries, and the relative importance of the variable in other research, it was decided to keep the variable for the multilevel analyses.

Teachers' specialization in math and education

There is some indication in the research literature that the specialization and formal education of the teachers is related to student outcomes. However, the index created by the TIMSS & PIRLS International Study Center, which stems from two questions related to teacher's *formal post-secondary education* and their *main area of specialization* (TQ-G5A/B) only showed low correlations with achievement in all GCC countries – and therefore was dropped from further analyses.

Time spent for professional development

Based on the comparatively low quality of education in the Gulf area, and on research findings similar to Blömeke et al. (2016), who found based on TIMSS 2011 data that professional development activities are especially important for the Arab countries, it could be assumed that a higher amount of *time spent for professional development* might be related to students' mathematics and science achievement. However, in addition to the amount of time spent for training, factors like the quality and content coverage of the courses also play an important role. In this context, correlations to student achievement using TIMSS 2015 data are partly negative, reaching -.20 in Qatar. Because of the importance attributed to professional development by other researchers, the variable was kept to be evaluated further in the multilevel analyses.

Preparedness to teach

An indicator summarizing several variables related to teachers' perceived *preparedness to teach*, as related to different content domains, was created to give an indication about subject matter mastery. This can be seen as a basic requirement for good teaching, as indicated for example by Monk (1994). However, related literature revealed less consistent and rather weak relations to student achievement, as further described in section 3.3.5.1. Given the weak correlations with achievement in the region, the indicator ultimately was dropped from further analysis steps.

Assessment of ongoing work in mathematics and science

The *emphasis on assessment* of student's ongoing work was selected as one component for the assessment dimension of the factor quality of instruction. There is quite some empirical research evidence for the importance of evaluating students' work and giving timely feedback that can be used for student's improvement. However, the question, as asked in the TIMSS 2015 questionnaire, didn't collect information regarding *how* the information is going to be used – i.e., whether it was *summative* (as a kind of final judgment) or *formative* (which would mean that the results are used to influence subsequent teaching and learning strategies). The latter construct was especially found to be more strongly related to student performance (see assessment and feedback strategies in section 3.3.5.2). Because of the low correlations, this variable was excluded from further analyses.

Number of topics covered for math/science

The curriculum content coverage is usually regarded as an important indicator for students' opportunity to learn; correspondingly, a question asking teachers about the curriculum coverage of the TIMSS topics is included in all cycles of TIMSS. Because of its theoretical importance, this variable will be kept for further analyses, in spite of low correlations indicated for the GCC countries. Interestingly, for Kuwait, the science content coverage is even significantly negatively associated with student achievement.

In a subsequent step, and based on the current research framework, the extent to which selected variables and indicators represent the same or a very similar construct was investigated. In such cases, the indicator with the strongest correlation to mathematics and science achievement was kept, and the remaining indicators for the same construct were excluded from further analyses. As the starting point here was a regional analysis, certain compromises had to be made. In order to maintain the same variable set for all countries, the average correlation was considered for the selection process; additionally, the extent to which countries differed in their associations between indicators and achievement was investigated. In some cases, constructs between course and school level were also parallel. In detail, the following components were regarded.

Educational resources on school level

In total, three explanatory variables (the *number of computers*, the *number of books in the school library*, and the principal's perspective on *how much the instruction is affected by specific shortages*) were available in the principal questionnaire as possible indicators for educational resources. While nearly no correlations between resource indicators and achievement could be found in Oman, for all other countries, except Qatar, the strongest indicator by far was the *number of books in the school library*. For Qatar, the strongest indicator was the principal's perspectives on *how much the instruction is affected by shortages*. Here it was decided to drop the weakest component, namely the *number of computers available* for fourth grade students, and to keep the other two.

Learning environment

The school questionnaire contained two questions related to the school learning environment, one related to the *emphasis on academic success* and another to *school discipline and safety*. While both are important determinants for the school climate, they nonetheless likely to be related, to a certain extent. To keep the model parsimonious, it was decided to keep *emphasis*

on academic success, which showed, by far, a stronger correlation to both mathematics as well as science achievement in five of the six countries.

The question about *emphasis on academic success* was also administered to the teachers, in addition to a question related to the *orderly learning environment*. As the latter had a stronger correlation to achievement on course level, it was kept for the multilevel analyses as an indicator for a supportive climate on course level.

9.4.3 Final components kept for multilevel analyses

The final model to be used for the subsequent multilevel analyses, after removing redundancies in constructs and certain components showing only weak association with students' achievement for both levels, is summarized for mathematics analyses in Table 9-32 and for science in Table 9-33. The information displayed includes the variable names for the mathematics and science analyses; a description of the variable or factor that was kept; and further information concerning the variable, such as the range of categories for categorical variables, or additional recoding steps that were undertaken either for easier interpretation of the multilevel results or to obtain better measurement of the variables. Both tables also include the means and standard deviations of each indicator by country. Please note that for the PCA factor extraction, a regional approach was chosen, wherein countries contribute equally. For this approach, *general* level 2 variables were weighted using a combined (overall) course/teacher weight, while mathematics variables were weighted using mathematics course weights and science variables using science course weights. The results are presented using the mathematics and science course weights, respectively, which reveal that regional averages for the *general* level 2 variables may slightly differ from zero. Variable names, in general, follow the TIMSS 2015 convention, with the exception of variables names beginning with 'J' – an indication that the variables were imputed. All variables starting with "F_" indicate factor variable scores obtained from the principal component analyses, as described Table 9-12. Further details about the recodings undertaken can be found in APPENDIX B.

The results of the inter-item correlation analyses to investigate possible cases of multicollinearity between different indicators showed coefficients above 0.8 only for the correlation between the constructs *school discipline and safety* and *instructional time* on school level for Bahrain and Qatar. In Bahrain, the correlation coefficient yielded 0.81 for mathematics and 0.82 for science. In Qatar, the coefficient amounted to 0.82 for both subjects.

Table 9-32: Final mathematics indicators selected for multilevel analyses and their means and standard deviations

Level	Factor	Factor - Details	Variable (math)	Variable description	Comment	BHR		KWT		OMN		QAT		SAU		ARE		Average	
						Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
School	Input	Resources	JCBG13A	# of books in school library	Recoded to # of books based on category means (0-125000)/ 1 unit = 100 books	39.0	39.6	24.6	34.4	35.5	33.4	71.7	45.5	5.6	13.3	59.5	45.5	39.3	42.9
			F_SHORTAGE_M	Shortage of resources (M)	Factor	-0.1	0.9	-0.2	1.0	-0.1	0.9	0.3	1.2	0.0	0.8	0.1	1.0	0.0	1.0
	Quality	Environment (SLE)	F_SC_EAS	Emphasis of academic success	Factor	-0.1	0.9	-0.5	1.2	0.1	0.8	0.5	0.9	-0.2	0.9	0.1	1.0	0.0	1.0
	Time		F_ABSENCE	Problems with absenteeism	Factor	0.2	1.0	-0.5	1.0	-0.1	1.1	0.3	0.8	-0.3	1.0	0.4	0.8	0.0	1.0
Course	Input	Teacher background	JTBG02	Gender of teacher	Male = 0, female = 1	0.8	0.4	0.9	0.3	1.0	0.2	0.9	0.3	0.5	0.5	0.9	0.3	0.8	0.4
			JTBG04	Teacher's highest education level	Recoded to years of schooling (6 yrs - 21 yrs)	16.2	0.7	15.8	1.5	16.4	1.4	16.1	1.2	15.3	1.9	16.4	1.3	16.0	1.4
			JTBM10	Time spent for professional development (M)	Recoded into days based on catefory means (0-5.63)	3.1	2.1	1.5	1.6	2.1	1.9	2.6	2.2	1.9	1.8	2.3	2.1	2.2	2.0
			F_CONFIDENCE_M	Confidence in teaching strategies (M)	Factor	0.0	1.0	-0.2	1.1	0.2	0.8	0.1	1.0	-0.2	1.1	0.0	1.0	0.0	1.0
		Student composition	F_ESCS_clX	Average economic and sociocultural status	ESCS (course average)	0.0	0.5	0.0	0.4	0.0	0.5	0.0	0.5	-0.1	0.5	0.1	0.6	0.0	0.5
			F_EARLYNUM_clX	Average early numeracy skills	Early numeracy skills (course average)	0.3	0.2	-0.2	0.3	0.0	0.3	0.0	0.4	-0.2	0.3	0.2	0.3	0.0	0.4
			JTSEX_clX	Average gender composition	Range 0 (=100% boys) to 1 (=100% girls)	0.5	0.4	0.5	0.5	0.5	0.1	0.5	0.4	0.5	0.5	0.5	0.4	0.5	0.4
			JNATIONAL_clX	Average composition in terms of immigrants	Range 0 (=100% nationals) to 1 (=100% non-nationals)	0.3	0.3	0.3	0.3	0.2	0.3	0.6	0.3	0.1	0.2	0.6	0.3	0.4	0.3
	Quality of Instruction	Structured teaching	F_CLEARST_M	Clear and structured teaching (M)	Factor	-0.1	1.0	0.0	1.0	0.4	0.7	-0.3	1.1	0.0	1.0	-0.1	1.0	0.0	1.0
		Activation	F_COGNACTIV	Cognitive activation	Factor	0.2	0.8	-0.5	1.2	0.3	0.8	0.2	0.9	-0.6	1.1	0.1	0.9	-0.1	1.0
		Management	JTBG15D	Teaching limited by disruptive students	0 = a lot...2 = not at all	1.0	0.6	0.8	0.6	0.9	0.7	1.2	0.5	0.9	0.6	1.1	0.6	1.0	0.6
		Climate	F_ENVIRONM	Orderly learning environment	Factor	-0.1	1.1	-0.2	1.0	0.1	0.8	0.3	0.8	-0.1	1.1	0.0	1.0	0.0	1.0
		Assessment	JTBM07Z	Verification of homework completion (M)	Sum of frequency categories for homework activities (0-6)	5.1	1.0	4.4	1.8	5.4	0.9	5.2	1.0	5.3	1.1	5.4	1.2	5.1	1.3
	Time		JTBM01	Time spent on subject (M)	Converted from minutes to hours per week	4.6	1.1	3.5	1.0	4.2	1.5	4.9	1.5	3.7	1.1	4.4	1.1	4.2	1.3
			JTDM07Z	Amount of homework assigned (M)	Based on category means from TQM07a/b, resp. TQS06a/b (minutes per week)	1.9	1.2	1.1	0.7	1.6	1.0	1.8	1.1	1.5	0.9	1.6	1.0	1.6	1.0
	Opportunity		JTDM06Z	Amount of topics covered (M)	Average from all topics taught. One unit increase = 10% add. Topics taught	8.6	1.6	8.4	1.5	9.1	1.5	7.6	1.8	8.2	1.4	8.0	1.7	8.3	1.7
Student	Student Characteristics	Student Background	F_ESCS	Economic, social and cultural status (ESCS)	Factor	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.1	1.0	0.0	1.0
			JNATIONAL	Nationality status	0 = national, 1 = non-national	0.3	0.5	0.3	0.4	0.2	0.4	0.6	0.5	0.1	0.3	0.6	0.5	0.4	0.5
			F_EARLYNUM	Early numeracy skills	Factor	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
		Aptitude	F_MOTIV_M	Student likes learning (M)	Factor	0.0	1.0	0.0	1.1	0.0	0.7	0.0	1.1	0.0	1.0	0.0	1.0	0.0	1.0
		Subject motivation	JSBG08	Student's absence from school	# of absences per month based on category means. Range: 0-4	0.0	1.5	0.0	1.5	0.0	1.4	0.0	1.5	0.0	1.6	0.0	1.5	0.0	1.5
	Time		JSBH09BB	Parental help with homework	Converted to times per week. Range: 0-5	0.0	1.5	0.0	1.5	0.0	1.4	0.0	1.6	0.0	1.6	-0.1	1.7	0.0	1.5

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
SD = standard deviation

Table 9-33: Final science indicators selected for multilevel analyses and their means and standard deviations

Level	Factor	Factor - Details	Variable (science)	Variable description	Comment	BHR		KWT		OMN		QAT		SAU		ARE		Average	
						Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
School	Input	Resources	JCBG13A	# of books in school library	Recoded to # of books based on category means (0-125000)/ 1 unit = 100 books	39.7	39.1	24.8	35.0	36.3	33.0	71.3	45.2	5.7	13.4	59.7	45.6	39.6	42.7
			F_SHORTAGE_S	Shortage of resources (M/S)	Factor	-0.1	0.9	-0.3	1.0	-0.1	0.9	0.3	1.2	0.1	0.8	0.1	1.0	0.0	1.0
	Quality	Environment (SLE)	F_SC_EAS	Emphasis of academic success	Factor	-0.1	0.9	-0.4	1.2	0.1	0.8	0.5	0.9	-0.2	0.9	0.1	1.0	0.0	1.0
	Time		F_ABSENCE	Problems with absenteeism	Factor	0.2	0.9	-0.5	1.0	-0.1	1.1	0.3	0.8	-0.3	1.0	0.4	0.8	0.0	1.0
Course	Input	Teacher background	JTBG02	Gender of teacher	Male = 0, female = 1	0.8	0.4	0.9	0.3	1.0	0.1	0.9	0.3	0.5	0.5	0.9	0.3	0.8	0.4
			JTBG04	Teacher's highest education level	Recoded to years of schooling (6 yrs - 21 yrs)	16.5	1.1	16.1	1.0	16.6	1.1	16.2	1.4	13.9	2.9	16.6	1.3	16.0	1.8
			JTBS09	Time spent for professional development (S)	Recoded into days based on category means (0-5.63)	2.5	2.0	1.5	1.6	1.7	1.8	2.4	2.2	1.9	1.9	2.1	2.0	2.0	2.0
			F_CONFIDENCE_S	Confidence in teaching strategies (S)	Factor	-0.1	1.0	0.0	1.0	0.1	0.8	0.3	0.9	-0.2	1.1	-0.1	1.1	0.0	1.0
		Student composition	F_ESCS_clX	Average economic and sociocultural status	ESCS (course average)	0.0	0.5	0.0	0.4	0.0	0.5	0.0	0.5	-0.1	0.5	0.1	0.6	0.0	0.5
			F_EARLYNUM_clX	Average early numeracy skills	Early numeracy skills (course average)	0.3	0.3	-0.2	0.4	0.0	0.3	0.0	0.4	-0.2	0.3	0.2	0.3	0.0	0.4
			JTSEX_clX	Average gender composition	Range 0 (=100% boys) to 1 (=100% girls)	0.5	0.4	0.5	0.4	0.5	0.1	0.5	0.4	0.5	0.5	0.5	0.4	0.5	0.4
			JNATIONAL_clX	Average composition in terms of immigrants	Range 0 (=100% nationals) to 1 (=100% non-nationals)	0.3	0.3	0.3	0.3	0.2	0.3	0.6	0.3	0.1	0.2	0.6	0.4	0.3	0.4
	Quality of Instruction	Structured teaching	F_CLEARST_S	Clear and structured teaching (S)	Factor	0.0	1.1	0.2	0.9	0.4	0.7	-0.2	1.0	-0.2	1.1	-0.2	1.1	0.0	1.0
		Activation	F_COGNACTIV	Cognitive activation	Factor	0.2	0.8	-0.2	1.0	0.3	0.7	0.3	0.8	-0.5	1.2	0.2	0.9	0.1	1.0
		Management	JTBG15D	Teaching limited by disruptive students	0 = a lot...2 = not at all	1.0	0.7	0.9	0.6	0.9	0.7	1.1	0.6	0.9	0.6	1.1	0.6	1.0	0.7
		Climate	F_ENVIRONM	Orderly learning environment	Factor	0.0	1.1	-0.3	1.1	0.0	0.9	0.4	0.8	-0.1	1.1	0.0	1.0	0.0	1.0
		Assessment	JTBS06Z	Verification of homework completion (S)	Sum of frequency categories for homework activities (0-6)	5.3	1.0	2.8	2.7	5.4	1.1	5.1	1.6	4.5	2.2	5.0	1.8	4.7	2.0
	Time		JTBS01B	Time spent on subject (S)	Converted from minutes to hours per week	2.9	0.9	2.0	1.0	3.5	1.0	3.3	1.4	2.0	1.1	3.0	1.1	2.8	1.3
			JTDS06Z	Amount of homework assigned (S)	Based on category means from TQM07a/b, resp. TQS06a/b (minutes per week)	1.1	0.5	0.6	0.6	1.3	0.8	1.1	0.6	0.9	0.5	1.1	0.7	1.0	0.7
	Opportunity		JTDS05Z	Amount of topics covered (S)	Average from all topics taught. One unit increase = 10% add. topics taught	7.3	2.2	8.1	1.7	7.6	1.7	6.6	1.8	8.7	1.3	7.5	1.8	7.6	1.9
Student	Student Characteristics	Student Background	F_ESCS	Economic, social and cultural status (ESCS)	Factor	0.0	1.0	0.0	0.9	0.0	1.0	0.0	1.0	0.0	1.0	0.1	1.0	0.0	1.0
			JNATIONAL	Nationality status	0 = national, 1 = non-national	0.3	0.5	0.3	0.4	0.2	0.4	0.6	0.5	0.1	0.3	0.6	0.5	0.3	0.5
			F_EARLYNUM	Early numeracy skills	Factor	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
		Aptitude	F_MOTIV_S	Student likes learning (S)	Factor	0.0	1.0	0.0	0.9	0.0	0.7	0.0	1.1	0.0	1.2	0.0	1.0	0.0	1.0
		Subject motivation	JSBG08	Student's absence from school	# of absences per month based on category means. Range: 0-4	0.0	1.5	0.0	1.5	0.0	1.4	0.0	1.5	0.0	1.6	0.0	1.5	0.0	1.5
	Time		JSBH09BB	Parental help with homework	Converted to times per week. Range: 0-5	0.0	1.5	0.0	1.5	0.0	1.4	0.0	1.6	0.0	1.6	-0.1	1.7	0.0	1.5

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirate
SD = standard deviation

Both constructs were created based on different options out of the same question (SCQ-16), which might explain the strong relationships. With the exclusion of the *school discipline and safety* on school level to avoid parallel constructs on school and course level, no high inter-correlations above 0.8 remained between the constructs and variables that were kept for use in the multilevel analyses.

9.5 Summary

This chapter described the results from the variable selection process for the final framework and the data reduction procedures undertaken to elaborate on the underlying constructs and to obtain a final set of variables to be used in the subsequent multilevel analyses.

In a first step, the TIMSS 2015 background questionnaires were examined for questions matching the framework elaborated in chapter 6. Altogether, more than 170 options from a total of 9 questions from the school questionnaire, 24 questions from the teacher questionnaire, 10 questions from the student questionnaire, and 8 questions from the parent questionnaire could be matched with the factors of the developed effectiveness framework from a theoretical perspective.

The second step aimed to reduce the number of variables to a smaller set of underlying factors by means of PCA in combination with reliability analyses. The data reduction resulted in four factors on school level: *shortage of resources* for mathematics and science instruction, *emphasis on academic success*, and problems with *absenteeism*. On course level, altogether six factors were created: *Confidence in teaching strategies* related to mathematics and science, *clear and structured teaching* in mathematics and science, *cognitive activation*, and an *orderly learning environment*. Finally, on student level, two factors were obtained: an index on the *ESCS* and a factor describing students' *early numeracy skills*. All factors showed KMO values far above the minimum criteria of 0.5, and significant results for Bartlett's test of sphericity. The internal reliability of the PCA scales reached Cronbach's alpha value of 0.7 or higher. Additionally, three indices for each subject were created based on a sum score: *Preparedness to teach mathematics and science topics*, *verification of homework assignment* related to mathematics and science, and the number of *topics covered* in mathematics and science. All indices here exceeded the previously defined minimum Cronbach's Alpha criterion of 0.5 for exploratory analyses.

In a subsequent step, correlation analyses between the retained predictors (factors, indices, and single variables) and mathematics and science achievement, respectively, were performed to

assess the validity of the construct. Retained components were then revisited once more in the light of theoretical framework, where redundancies in constructs and components were also examined. The purpose was to reach a compromise between including all possible indicators and variables in the subsequent multilevel models on one hand, and on the other hand, obtaining parsimonious models that could still be calculated once all model factors were entered jointly.

Components with a correlation to mathematics and science below 0.2 in any of the six countries were excluded from further analyses, unless literature review gave strong empirical support in favor of keeping the component. The final list of indicators to be kept for the subsequent multilevel analyses consisted of 5 variables related to mathematics, science, or both on school level, 19 variables related to factors on course level, and 7 variables on student level. Additionally, the course averages of the ESCS index, of the early numeracy skills, and of the composition in terms of nationality and gender were calculated and retained for further analyses.

10 RESULTS OF THE MULTI LEVEL ANALYSES

This chapter explores the extent to which factors at the student, course, and school level are associated with student outcomes in mathematics and science. The analyses are guided by the main research questions, which examined the extent to which the different effectiveness factors identified according to the framework proposed in chapter 6 associate with student achievement. The multilevel analyses found in this chapter add to the results of the correlation analyses described in the previous chapter, deemed as the appropriate method to address the research question in more depth. The analyses described below are intended to illuminate the degree to which factors in the school environment, especially those of a malleable nature, of the GCC countries explain variation of student's mathematics and science achievement. The question of which of the framework factors are specifically important, and whether these factors emerge consistently among the region, merits investigation. All multilevel models distinguish between two levels (the student level and the course/school level), as often only one class per school was selected for a participation in the TIMSS assessment, and hence variances between classes and schools cannot be clearly distinguished from each other. The group level (level 2) for the subsequent analyses is the *course/school* level, which distinguishes the different teacher-course combinations in a given school. In the vast majority of classes in the region only one teacher is teaching the whole class. In these cases class-level and course-level are equivalent. If different math or science teachers teach the selected TIMSS students, then the student information was multiplied for each teacher-course combination. A justification of this procedure can be found in section 8.3.3. All school-level variables were then disaggregated and matched to the respective courses of that school.

As the primary focus of the current research is on identifying and interpreting course- and school-level variables on student performance, level 1 student variables were generally entered grand-mean centered into the multilevel models, except for the *nationality status*, which was left uncentered. Level 2 variables were always entered uncentered. In order to allow comparisons with results from similar models of other studies where researchers sometimes decided differently on the centering approach, a separate set of models were calculated using a group-mean centering approach of all level 1 predictors. Results related to the amount of variances explained when applying a group-centering approach can be found in APPENDIX D.

At first, a null model was created, used to identify the proportion of variance between group level and individual level. Subsequent, more complex models focus on two perspectives: firstly, predictor variables that show significant relations with student achievement; and secondly, the

proportion of variance that can be explained by each model. A level 1 student background model is created to determine the amount of variance that can be explained on an individual level, but the model additionally serves as a reference for the amount of group-level variance that can be explained by the addition of level 2 predictor variables in the subsequent models. Only the model described in section 10.4 contains purely level 2 predictors and consequently will be compared directly with the null model. Please refer to the corresponding subsections in section 8.5.2 for more detailed information about the centering approach and the procedures to calculate the proportions of explained variance.

The following sections describe the proportion of variance explained by the different models and also present the explanatory variables on student level and group level, showing significant results in relation to mathematics and science achievement. Section 10.1 will describe the partitioning of the variance between the two levels, section 10.2 will present the results from the student background variable analyses on level 1, section 10.3 will present the final student model including the aggregated student composition variables included on level 2, section 10.4 will show the results from a course/school model without controlling, and section 10.5 finally will present the full country-specific models that include all framework indicators on both levels.

10.1 The Null Model

The null model is the simplest model, as it only includes the outcome variable and no explanatory variables. In order to investigate possible country differences, separate models were calculated for each country subject combination amounting to altogether 12 different null-models.

10.1.1 Mathematics

The results for the six country-specific null models with mathematics achievement as the outcome variable are presented in Table 10-1. The table shows the overall country intercepts as the only fixed effects in the model, the between-course variance (level 2 variance), and the within-course variance (variance on level 1). All estimated parameters are statistically significant at $p \leq 0.05$.

In general, the confidence intervals of the intercepts include the results obtained from the weighted country averages reported in Table 2-6, albeit the intercepts of the null model in Oman are somewhat lower (416 vs. 425 score points).

The results of the null models allow for partitioning of the variance into the two levels. The proportion of variance that is between courses (the level 2 variance) ranges from close to 25% in Bahrain to nearly 60% in the United Arab Emirates.

The deviance values reported in the last row will be used for comparison with the subsequent, more complex models. Deviance is expected to go down with the subsequent, more elaborated models which should better fit the empirical data.

Table 10-1: Null models for mathematics

Effects	Null Model - Mathematics					
	BHR	KWT	OMN	QAT	SAU	ARE
Fixed Effects						
Intercept	453.3 (2.6) **	345.1 (3.8) **	416.4 (3.1) **	443.2 (4.4) **	386.0 (4.6) **	447.3 (2.9) **
Random Effects						
Between-course variance (τ_{00})	1929.3 (213.0) **	3150.0 (258.4) **	2992.7 (237.9) **	3818.0 (349.9) **	2833.7 (268.4) **	6667.3 (311.5) **
Within-course variance (σ^2)	5952.1 (169.4) **	7281.1 (262.2) **	7566.8 (195.2) **	5504.9 (165.6) **	5003.0 (190.5) **	4617.7 (117.4) **
Proportion of variance on course level	24.5%	30.2%	28.3%	41.0%	36.2%	59.1%
Deviance	187464.2	583954.2	644284.6	217346.4	4826690.7	810115.2

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
() Standard errors appear in parenthesis
** $p \leq 0.05$

10.1.2 Science

Table 10-2 shows the results for the country-specific null models using science achievement as the outcome measure. Again, the intercepts for most of the countries are matching the weighted country averages reported in table Table 2-6. Likewise, all parameters are statistically significant at $p \leq 0.05$. Intercepts of the null model for Oman (13 score points difference) and, for science, also in the United Arab Emirates (10 points difference) are somewhat lower than the weighted country averages but are still within the confidence intervals.

The between-course/school variances for science range from 22% in Oman to 51% in the United Arab Emirates. In all countries except Bahrain, the proportion of level 2 variance is somewhat lower for science achievement compared to mathematics achievement.

Table 10-2: Null models for science

Effects	Null Model - Science					
	BHR	KWT	OMN	QAT	SAU	ARE
Fixed Effects						
Intercept	460.0 (3.4) **	330.9 (5.3) **	418.2 (3.8) **	435.2 (4.7) **	393.4 (5.9) **	441.9 (3.4) **
Random Effects						
Between-course variance (τ_{00})	3280.3 (213.0) **	3150.0 (258.4) **	2992.7 (237.9) **	3818.0 (349.9) **	2833.7 (268.4) **	6667.3 (311.5) **
Within-course variance (σ^2)	7856.4 (351.0) **	10375.8 (287.8) **	10482.6 (296.1) **	7808.4 (244.1) **	8197.9 (310.6) **	6363.8 (133.7) **
Proportion of variance on course level	29.5%	23.3%	22.2%	32.8%	25.7%	51.2%
Deviance	192041.9	601529.4	662111.1	223869.4	5036527.9	832874.2

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
() Standard errors appear in parenthesis
** $p \leq 0.05$

10.2 The Level-1 Student Background Models

Once the null models were examined, the student background models were created for both subjects. In this set of models, all student background variables that were retained from previous analyses steps were added as explanatory variables. Please refer to Table 9-32 (mathematics) and Table 9-33 (science) to obtain further information on the coding and certain basic statistics related to the analysis variables included in the current as well as the consecutive models. For the results presented in the subsequent tables, due to space constraints a shorter description will be used. Table 11-1 matches the previously used labels (listed in the column "Description") with the shorter version (listed in the column "Short Description") used for this chapter.

10.2.1 Mathematics

The background models for mathematics are presented in Table 10-3. The upper part of the tables shows the results for the fixed effects of the country-specific models. As all variables except *nationality status* were entered grand-mean centered, the multilevel results for the student background models, but as well also for all further models can be interpreted in comparison to a *national* student whose characteristics in regards to *ESCS*, *early numeracy*, *motivation*, *absenteeism*, and *parental help with homework* are corresponding to the country averages of these variables. For each increase of one unit of a variable, the table shows the corresponding changes in mathematics outcomes. It should be noted that all variables that were combined to scales (on student level, these comprise the *ESCS*, the *early numeracy activities*, and the subject motivation (*student likes learning*)) were standardized through principal component analyses on regional level to a mean of zero, and a standard deviation of one. While the country-specific means of the factors are especially differing from zero, the standard deviations in most cases are still close to one also for the country-specific analyses. Means and standard deviations for all variables used in the multilevel analyses can be found in Table 9-32 for mathematics and in Table 9-33 for science. Single variables were coded using the original metric but with their

lowest category starting with zero. Thus, a student who *likes learning* mathematics about a standard deviation above the average subject motivation would be assumed to score close to 18 points more in Oman compared to the average Omani student. Similarly, a student in Bahrain who reported being absent once more per month than the average student would be assumed to score more than 9 points lower in mathematics.

The ESCS index is significant in all GCC countries, showing the highest absolute value in Oman. *Non-nationals* in all countries except Oman are expected to have significantly higher mathematics achievement, ranging from 21 points in Bahrain to 36 score points in Qatar.

Subject motivation (*student likes learning*) is significantly associated with mathematics outcomes in all countries – reaching close to 18 score points in Oman. The number of absences per month (*absenteeism*) and the *parental help with homework* are also significantly related to mathematics achievement, but the association is negative, meaning that higher values of the predictors are associated with lower achievement. A negative association between *absenteeism* and student achievement is expected due to the coding of the variable, which indicates the number of absences per month. Unexpectedly negative associations with achievement also occur on course level, and will be discussed further in chapter 11.

The lower part of Table 10-3 shows the level 2 variance components explained by the country-specific background models. The variance components explained by the individual student background variables on level 1 range from 5% in Kuwait to 15% in Qatar. No explained portions of level 2 variance components are reported for this model, as a reduction in level 2 variance only is interpretable between models with the same level 1 specification. Raudenbush and Bryk (2002, p. 150) stated that introducing predictors on level 1 changes the meaning of the intercept – and as such, represents the variability for a different parameter. They concluded that as a consequence, the residual level 2 variance may be smaller or even larger compared to the corresponding variance component of the null model and therefore a comparison between the null model and the level 1 model is meaningless. Thus, the level 1 specification of the current model will be retained for the subsequent models (except for one model that only contains level 2 predictors) to allow a meaningful comparison of the variance components which are always regarded in comparison to the level 1 model described here.

The last two lines show the deviance, which is a measure of the appropriateness of the model given the empirical data. Comparing the deviances between null models (Table 10-1) and the student background model shows lower deviance for all countries, which indicates that the student background model fits the empirical data better. The last line in the variance section shows

the extent to which the changes in deviance between the models are significant. Values of 0.05 and below are interpreted as significant, which is the case for all six models.

Table 10-3: Student background models – mathematics

Level	Factor	Details	Explanatory Variable	BHR	KWT	OMN	QAT	SAU	ARE
Student	Student Characteristics	Student background	ESCS	16.7 (1.4) **	11.2 (1.8) **	23.2 (1.6) **	13.0 (1.6) **	8.9 (1.7) **	13.1 (1.0) **
			Nationality	20.6 (3.4) **	23.3 (3.7) **	2.9 (5.1)	36.1 (3.5) **	30.9 (5.9) **	26.1 (2.1) **
		Aptitude	Early numeracy	9.7 (1.4) **	8.5 (1.5) **	12.2 (1.3) **	9.4 (1.5) **	9.5 (2.0) **	7.9 (0.8) **
		Motivation	Student likes learning	8.4 (1.2) **	6.4 (1.3) **	17.8 (1.7) **	9.0 (1.3) **	6.8 (1.7) **	10.0 (0.8) **
	Time		Absenteeism	-9.4 (0.9) **	-5.1 (1.1) **	-7.1 (0.8) **	-8.1 (1.1) **	-4.2 (0.9) **	-7.5 (0.4) **
	Opportunity		Parental help	-3.7 (0.9) **	-2.8 (0.9) **	-3.2 (1.2) **	-4.3 (0.9) **	-3.0 (1.1) **	-3.7 (0.4) **
Variance	Explained variance (Level 1)			12%	5%	13%	15%	7%	11%
Model Fit	Deviance			185404	581339	636712	214212	4793758	801333
	Significance of changes in deviance to Null model			0.000	0.000	0.000	0.000	0.000	0.000

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
() Standard errors appear in parenthesis
** $p \leq 0.05$

10.2.2 Science

The results from the student background model for science are presented in Table 10-4. Science tables are generally designed equally to their math counterparts. Thus, the upper part shows the results of the fixed effects.

While for mathematics, all background variables except the *nationality status* in Oman are significantly associated with achievement in science, the *parental help* with homework in Kuwait is also not significantly related to science outcomes. As is the case for mathematics, the *ESCS* indicator is an especially strong predictor for achievement in Oman, while the *nationality status* seem to be especially important in all other countries. *Early numeracy* skills seem to have some importance for Saudi Arabia, especially in science. Similar to the mathematics results, all background variables except *absenteeism* and *parental help* are positively related to science achievement.

When comparing the magnitude of the mathematics and science effects, it seems that the overall pattern is similar, but that magnitude of the science estimations in general is somewhat higher; this is especially true for the subject motivation and also for the *ESCS* index.

The lower part of Table 10-4 displays the amount of explained variance in level 2 science achievement. The variance components explained by the individual student background variables on level 1 in terms of absolute values are a bit higher compared to the mathematics results. The share of explained level 1 variance ranges from 7% in Kuwait to 18% in Qatar.

The last two lines present the deviance results, which indicate that in addition to the mathematics models, the science level 1 background models also fit the empirical data significantly better than the null models.

Table 10-4: Student background models – science

Level	Factor	Details	Explanatory Variable	BHR	KWT	OMN	QAT	SAU	ARE
Student	Student Characteristics	Student background	ESCS	19.1 (2.6) **	13.9 (3.2) **	29.0 (1.8) **	15.8 (2.2) **	13.7 (1.9) **	16.8 (1.5) **
			Nationality	21.7 (6.9) **	28.5 (7.0) **	8.7 (5.1)	44.9 (4.2) **	29.8 (5.7) **	35.6 (2.7) **
		Aptitude	Early numeracy	8.6 (1.6) **	9.9 (2.9) **	13.9 (1.6) **	11.3 (1.5) **	15.0 (2.5) **	8.3 (0.9) **
		Motivation	Student likes learning	17.8 (2.3) **	20.0 (3.2) **	26.2 (1.9) **	12.2 (1.6) **	7.2 (1.6) **	13.5 (0.9) **
	Time		Absenteeism	-11.0 (1.2) **	-5.1 (1.8) **	-7.0 (1.0) **	-11.2 (1.1) **	-5.6 (1.3) **	-8.6 (0.6) **
	Opportunity		Parental help	-3.2 (1.3) **	-2.5 (1.6)	-2.9 (1.4) **	-4.3 (0.9) **	-4.5 (1.1) **	-4.0 (0.5) **
Variance	Explained variance (Level 1)			14%	7%	14%	18%	8%	13%
Model Fit	Deviance			189652	597832	653558	220182	5000516	822725
	Significance of changes in deviance to Null model			0.000	0.000	0.000	0.000	0.000	0.000

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
() Standard errors appear in parenthesis
** $p \leq 0.05$

10.3 The Student Background Models Including Student Composition Variables on Level 2

In a subsequent step, investigations were made concerning the extent to which the student composition in terms of important student background variables would have an additional effect beyond the effect of those variables on individual level. School composition variables were included on level 2 to more clearly disentangle school effects from home background effects. In the current framework, the student composition is regarded as an input of the processes of learning and teaching in schools and therefore was classified as an element of the *input* factor in the model. See section 3.4.6 for a more in-depth discussion of the student composition effects.

10.3.1 Mathematics

Table 10-5 lists the results from the background model, which, in addition to the student-level variables, includes the composition variables comprising the course averages variables related to the home background of the students.

When comparing the student-level models (Table 10-3), it is evident that the estimates for the level 1 indicators, as expected, have not changed with the inclusion of the class averages for student background variables. The additional entered student composition variables, however, show quite a different pattern among countries. While the average *ESCS* is significantly related to mathematics achievement – with about 19 points in Oman and even up to 75 points in the United Arab Emirates for a change of about 1 standard deviation on the ESCS scale – in Saudi

Arabia, the *ESCS* student composition shows the lowest estimate, and is not significant. The aggregated *early numeracy* skills only show significant effects in half of the countries. The *gender* composition effects seem to be especially important in Oman (albeit with a high standard error) and in Saudi Arabia, but is not significant in Qatar. However, as for the subsequent models, it is noteworthy that the *gender* variable, due to the completely gender-segregated classes in Saudi Arabia, was omitted from the level 1 models. The composition effect in terms of gender estimated by the group aggregate of the gender, therefore, might to some extent be overestimated. The composition effect related to *nationality* is only significant in two countries, namely Bahrain and Oman; interestingly, in these countries it is negatively associated with mathematics achievement. This is especially surprising in Bahrain, as on individual level immigrant students on average achieve significantly higher outcomes than national students; hence, this finding will need further explanation.

When comparing with the level 1 student model described in the previous section, it is evident that the addition of level 2 student background composition variables explained between 12% of the group-level variance in Oman and Saudi Arabia and 28% in Qatar. In addition, the model statistics show a further significant reduction of the deviance, and consequently a better model fit which includes level 2 composition effects.

Table 10-5: Student background models including composition – mathematics

Level	Factor	Details	Explanatory Variable	BHR	KWT	OMN	QAT	SAU	ARE
Course	Input	Student composition	ESCS (avg.)	27.4 (4.4) **	58.4 (7.1) **	19.2 (7.5) **	58.6 (7.6) **	18.5 (9.9)	74.8 (3.8) **
			Early numeracy (avg.)	24.5 (9.7) **	29.9 (13.2) **	24.0 (10.5) **	-12.3 (10.8)	-19.6 (15.6)	6.3 (7.6)
			Gender (avg.)	19.6 (3.6) **	20.5 (5.0) **	48.2 (22.3) **	6.5 (7.0)	36.8 (8.3) **	14.6 (5.3) **
			Non-nationals (avg.)	-27.6 (9.4) **	25.5 (14.8)	-94.0 (14.4) **	14.6 (13.0)	-9.1 (21.6)	-0.7 (6.9)
Student	Student Characteristics	Student background	ESCS	16.0 (1.4) **	10.9 (1.8) **	23.1 (1.6) **	12.1 (1.6) **	8.9 (1.7) **	12.5 (1.0) **
			Nationality	21.9 (3.6) **	22.4 (3.8) **	4.1 (5.1)	35.0 (3.4) **	30.9 (5.9) **	24.9 (2.1) **
		Aptitude	Early numeracy	9.5 (1.4) **	8.4 (1.5) **	12.1 (1.4) **	9.4 (1.5) **	9.5 (2.0) **	7.9 (0.8) **
		Motivation	Student likes learning	8.4 (1.2) **	6.5 (1.3) **	17.8 (1.7) **	9.0 (1.3) **	6.8 (1.7) **	10.0 (0.8) **
	Time		Absenteeism	-9.4 (0.9) **	-5.0 (1.1) **	-7.1 (0.8) **	-8.1 (1.1) **	-4.2 (0.9) **	-7.5 (0.4) **
	Opportunity		Parental help	-3.7 (0.9) **	-2.8 (0.9) **	-3.2 (1.1) **	-4.3 (0.9) **	-3.0 (1.1) **	-3.6 (0.4) **

Variance	Change in explained variance (Level 2)	20%	22%	12%	28%	12%	26%
	Change in explained variance (Level 1)	0%	-2%	0%	0%	0%	0%
Model Fit	Deviance	185310	581180	636648	214059	4793732	800769
	Significance of changes in deviance to L1 background model	0.000	0.000	0.000	0.000	0.000	0.000

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
() Standard errors appear in parenthesis
** $p \leq 0.05$

10.3.2 Science

Similar to the mathematics model, the science results of the current model also show no differences in significance or magnitude of level 1 variables, compared to the student level model displayed Table 10-4.

Compared to the corresponding mathematics model, the pattern in terms of the size of estimates is similar but not identical. The *ESCS* index for science is significant for all countries including

Saudi Arabia, but fewer countries show significant effects in terms of the *early numeracy* composition. The *gender* composition effect is significant in all countries. The pattern of the relative magnitude of effects, in most cases, is roughly similar to the pattern found for mathematics. However, in general the gender composition effect in science is more pronounced. This probably can be explained by the fact that gender differences are generally higher in science when compared to mathematics, as seen Table 9-1 and Table 9-2.

Table 10-6: Student background models including composition – science

Level	Factor	Details	Explanatory Variable	BHR	KWT	OMN	QAT	SAU	ARE
Course	Input	Student composition	ESCS (avg.)	20.0 (6.2) **	69.8 (9.5) **	18.7 (9.0) **	51.3 (8.4) **	38.3 (11.0) **	73.3 (4.3) **
			Early numeracy (avg.)	24.1 (9.2) **	10.9 (11.6)	34.9 (12.4) **	-2.5 (12.9)	-17.9 (19.0)	12.1 (7.9)
			Gender (avg.)	41.7 (5.2) **	36.5 (7.4) **	93.3 (29.7) **	29.0 (7.8) **	73.6 (8.7) **	27.9 (5.2) **
			Non-nationals (avg.)	-25.7 (12.6) **	29.3 (17.6)	-111.7 (17.7) **	2.2 (13.8)	-6.1 (23.5)	7.3 (8.0)
Student	Student Characteristics	Student background	ESCS	18.6 (2.6) **	13.5 (3.3) **	28.9 (1.9) **	15.1 (2.2) **	13.7 (1.9) **	16.0 (1.5) **
			Nationality	22.7 (7.2) **	27.7 (7.0) **	9.7 (5.0)	44.0 (4.3) **	29.8 (5.7) **	34.0 (2.8) **
		Aptitude	Early numeracy	8.4 (1.6) **	9.9 (2.9) **	13.8 (1.6) **	11.3 (1.5) **	15.0 (2.5) **	8.3 (0.9) **
		Motivation	Student likes learning	17.7 (2.3) **	20.0 (3.2) **	26.0 (1.9) **	12.3 (1.6) **	7.2 (1.6) **	13.5 (0.9) **
	Time		Absenteeism	-10.9 (1.2) **	-5.0 (1.8) **	-7.0 (1.0) **	-11.2 (1.1) **	-5.6 (1.3) **	-8.6 (0.6) **
	Opportunity		Parental help	-3.1 (1.3) **	-2.5 (1.6)	-2.9 (1.3) **	-4.2 (0.9) **	-4.5 (1.1) **	-3.9 (0.5) **

Variance	Change in explained variance (Level 2)	15%	27%	11%	24%	29%	30%
	Change in explained variance (Level 1)	0%	0%	1%	0%	0%	0%
Model Fit	Deviance	189570	597698	653468	220085	5000456	822165
	Significant changes in deviance to L1 background model	0.000	0.000	0.000	0.000	0.000	0.000

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
() Standard errors appear in parenthesis
** $p \leq 0.05$

The amount of explained level 2 variance due to the addition of the aggregated student background variables overall is comparable to the mathematics results, ranging from 11% in Oman to 30% in the United Arab Emirates. For Saudi Arabia, however, it is remarkable that group level predictors in science explain more than twice the amount of variance compared to the mathematics results.

10.4 School- and Course-Level Effectiveness Variables Without Controlling for the Student Background

In an intermediate step, school environment variables, teacher characteristics, and variables related to classroom instruction were investigated by means of a separate model, without controlling for the student background. This model is complementary to the previously described student background model, with composition effects, as adding the predictors of both models together will result in the full model described in the subsequent section.

As this model does not contain level 1 variables, the explained group level variance was obtained via comparison with the null model.

10.4.1 Mathematics

Table 9-7 displays the outcomes of the multilevel model with all course- and school-level variables combined, but uncontrolled for student achievement. Once entered jointly, the number of significant effects are quite low, especially in Saudi Arabia where only the resource indicator number of *library books* and an indicator for the quality of instruction, namely *clear teaching*, stay significant. Unexpectedly, some model indicators are significantly negatively associated with mathematics achievement in several countries. This is the case, for example, for the teacher characteristics indicators *Education level* in Oman and the amount of *time for development* in Qatar and in the United Arab Emirates. *Confidence in teaching* seems to be negatively related to achievement in the United Arab Emirates. Finally, the *amount of homework* assigned is negatively associated with achievement in Bahrain.

With the exception of Kuwait, *clear and structured teaching* seems to be the most consistent and also strongest factor associated with mathematics achievement in the region. *Resource*-related input variables show significant effects in five countries, while the number of *topics covered* in the current or last school year still are related to achievement in four countries of the region. Overall, the highest number of significant effects of the uncontrolled model is shown in the United Arab Emirates.

The second part of the table displays the results from the variance analyses. It can be seen that course and school variables explain between 26% of the mathematics outcome variance in Kuwait to even 52% in the United Arab Emirates. All models show a better model fit compared to the corresponding null models.

Table 10-7: Course/ school level model without controlling – mathematics

Level	Factor	Details	Explanatory Variable	BHR	KWT	OMN	QAT	SAU	ARE
School	Input	Resources	Library books	0.2 (0.1) **	0.4 (0.1) **	0.0 (0.1)	0.1 (0.1)	0.4 (0.2) **	0.4 (0.1) **
			Shortage resources	6.9 (2.7) **	3.0 (3.4)	0.2 (3.2)	9.4 (3.0) **	2.6 (4.6)	8.8 (2.7) **
	Quality	Environment (SLE)	Emphasis on success	6.3 (2.7) **	10.4 (3.7) **	-3.2 (3.3)	6.8 (4.1)	5.7 (5.3)	16.2 (2.6) **
	Time		Absenteeism	2.1 (2.7)	1.8 (4.0)	2.9 (2.7)	3.3 (4.7)	5.4 (5.4)	9.1 (2.6) **
Course	Input	Teacher background	Gender of teacher	-0.4 (6.3)	9.7 (11.8)	5.4 (18.5)	-14.8 (9.1)	11.9 (11.9)	-13.0 (7.0)
			Education level	-2.9 (3.6)	-0.8 (2.5)	-9.7 (2.1) **	-0.8 (2.4)	1.6 (1.9)	0.4 (2.3)
			Time for development	0.6 (1.2)	2.5 (2.4)	-0.1 (1.5)	-4.9 (1.6) **	-3.3 (1.9)	-3.7 (1.1) **
			Confidence in teaching	-1.5 (2.8)	-0.3 (3.6)	6.7 (3.4)	-1.6 (4.6)	9.5 (5.1)	-6.2 (2.8) **
	Quality of Instruction	Structured teaching	Clear teaching	11.3 (2.3) **	3.3 (2.8)	21.8 (3.7) **	21.7 (3.6) **	17.1 (4.9) **	15.6 (2.6) **
		Activation	Cognitive activation	5.2 (3.4)	0.7 (3.7)	-4.2 (3.5)	-0.1 (4.7)	-5.4 (4.7)	8.8 (2.8) **
		Management	Disruptive students	2.8 (4.6)	15.5 (5.5) **	3.2 (3.6)	15.6 (7.4) **	13.0 (8.2)	9.8 (4.2) **
		Climate	Orderly environment	8.4 (2.3) **	1.6 (4.0)	7.8 (3.4) **	2.3 (5.5)	-0.7 (4.8)	16.1 (2.8) **
		Assessment	Hmwk. verification	0.2 (2.1)	-4.1 (2.3)	3.5 (2.8)	4.7 (3.7)	5.7 (4.3)	-1.4 (2.2)
	Time		Time spent on subject	1.1 (1.9)	7.6 (3.5) **	0.1 (1.7)	1.4 (2.7)	1.9 (5.7)	1.6 (1.9)
			Amount of homework	-4.2 (2.1) **	16.8 (8.3) **	-3.3 (2.7)	-2.9 (3.3)	-1.3 (4.3)	3.1 (2.3)
	Opportunity		Topics covered	3.9 (1.7) **	1.8 (2.6)	5.7 (1.9) **	6.2 (2.2) **	3.3 (3.2)	5.9 (1.3) **
Variance	Explained variance (Level 2)			44%	26%	39%	43%	28%	52%
Model Fit	Deviance			187374	583911	644213	217249	4826627	809740
	Significance of changes in deviance to Null model			0.000	0.000	0.000	0.000	0.000	0.000

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
() Standard errors appear in parenthesis
** $p \leq 0.05$

10.4.2 Science

Corresponding results for science are shown in Table 10-8. The model shows that in Saudi Arabia, the *gender of the teacher* is the only significant model variable, while in the United Arab Emirates altogether 11 model indicators show significant relations to science achievement. *Clear teaching* again emerges as the most consistent model factor across the region, with significant effects in five countries, and rather higher estimates compared to other model factors. Also for science, certain indicators in some of the countries show counterintuitive results; as was found for mathematics, the amount of *time for development* spent in the United Arab Emirates and the *teacher education level* in Oman is significantly negatively associated with achievement. Additionally, a frequent *homework verification* seems to be negatively associated with achievement in Qatar and in the United Arab Emirates. Moreover, *the time spent* on science instruction in Bahrain and the United Arab Emirates is negatively related to science achievement, but (as would be expected according to the framework) the relation is positive in Oman. Finally, a higher number of science *topics covered* is negatively associated with student outcomes in Kuwait, while the relation is positive for the same indicator in the United Arab Emirates.

Variance components demonstrate that between 33% of the between-group variance in Kuwait and Oman and 53% in Bahrain can be explained by the course- and school-level variables. Compared to the corresponding mathematics results, in four out of six countries, the amount of explained level 2 variance is somewhat higher for the science model, and here especially in Saudi Arabia.

Again, the deviance statistics show a significantly better model fit compared to the corresponding null models for all countries.

Table 10-8: Course/ school level model without controlling – science

Level	Factor	Details	Explanatory Variable	BHR	KWT	OMN	QAT	SAU	ARE
School	Input	Resources	Library books	0.2 (0.1) **	0.4 (0.2) **	0.0 (0.1)	0.1 (0.1)	0.4 (0.2)	0.4 (0.1) **
			Shortage resources	0.1 (3.5)	1.9 (3.9)	3.5 (4.1)	6.0 (2.9) **	-3.3 (6.4)	7.7 (3.0) **
	Quality	Environment (SLE)	Emphasis on success	7.8 (3.7) **	16.8 (4.8) **	-5.4 (4.0)	13.1 (4.0) **	0.6 (6.6)	15.6 (2.9) **
	Time		Absenteeism	7.3 (3.5) **	0.2 (5.3)	3.9 (3.4)	1.6 (4.2)	7.8 (5.7)	11.2 (3.2) **
Course	Input	Teacher background	Gender of teacher	6.6 (8.9)	55.3 (18.1) **	24.5 (18.4)	-2.2 (9.7)	62.0 (13.0) **	-0.6 (8.6)
			Education level	0.1 (3.3)	6.9 (5.1)	-9.9 (3.2) **	2.5 (2.9)	0.9 (1.8)	1.9 (2.2)
			Time for development	-0.1 (1.4)	-2.6 (2.2)	-0.6 (2.1)	-3.0 (1.9)	-2.3 (2.2)	-3.4 (1.3) **
			Confidence in teaching	-3.6 (3.5)	-1.0 (5.2)	1.3 (4.9)	-7.7 (5.5)	-1.7 (5.8)	3.1 (2.7)
	Quality of Instruction	Structured teaching	Clear teaching	14.1 (3.3) **	10.8 (4.8) **	30.2 (4.7) **	30.2 (4.1) **	10.6 (6.0)	24.8 (2.8) **
		Activation	Cognitive activation	3.6 (4.1)	8.0 (5.8)	8.3 (5.5)	20.9 (5.6) **	6.2 (5.3)	6.3 (3.5)
		Management	Disruptive students	10.8 (5.3) **	5.5 (7.3)	5.7 (5.2)	7.3 (7.0)	8.8 (7.6)	15.6 (4.3) **
		Climate	Orderly environment	8.8 (3.1) **	1.9 (4.8)	-0.4 (4.2)	3.0 (5.5)	6.0 (4.3)	13.3 (3.1) **
		Assessment	Hmwk. verification	6.0 (3.2)	-0.2 (3.2)	-4.3 (3.0)	-8.0 (3.0) **	3.8 (4.1)	-3.3 (1.5) **
	Time		Time spent on subject	-10.3 (4.8) **	1.8 (5.1)	10.8 (3.8) **	-5.1 (3.3)	3.4 (4.4)	-5.5 (2.4) **
			Amount of homework	-2.6 (7.1)	10.4 (11.5)	-4.8 (4.8)	2.4 (6.6)	-11.3 (15.7)	-1.6 (4.2)
	Opportunity		Topics covered	1.4 (1.9)	-6.0 (2.3) **	-0.3 (1.9)	0.8 (2.2)	-0.5 (4.0)	5.2 (1.5) **
Variance	Explained variance (Level 2)			53%	33%	33%	50%	48%	52%
Model Fit	Deviance			191984	601455	662050	223773	5036454	832482
	Significance of changes in deviance to Null model			0.000	0.000	0.000	0.000	0.000	0.000

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
() Standard errors appear in parenthesis
** $p \leq 0.05$

10.5 School Effectiveness Variables after Controlling for the Student Background

In a final step, the full country-specific models were built. In addition to the course- and school-related instructional and environmental variables (identified according to the educational effectiveness model), these models also contained the student background factors on level 1, as well as the student composition variables on level 2. As such, they constitute a synthesis of the previous two models described in section 10.3 and 10.4. The full model has two purposes: first, it should be used to investigate how all indicators, identified according to the framework specified in chapter 6, behave when entered simultaneously. Additionally, the final models allow focus to be placed on possible malleable course and school factors that emerge in the different GCC countries after controlling for student background on both levels. According to the OECD's publication "Best practices to assess the value-added of schools" (OECD, 2008), the type of model used here belongs to the *contextualized attainment models* (CAM) group. Such models allow (at least to a certain extent) for the separation of educational from non-educational influences, when using the available cross-sectional data at hand (see more on the use of TIMSS data for EER in section 8.2.2).

10.5.1 Mathematics

The results from the analyses with all components entered jointly are shown in Table 10-9 for mathematics. As expected, all student-level background variables show basically the same estimates and significances as in the separate home background models. The number of significant student composition variables in the final model is a bit different compared to the pure student background model. In Kuwait, in the final model, the average composition of *non-nationals* also becomes significant; while in Oman, the level 2 average of the *early numeracy* skills and the average *gender* composition is not significant anymore, as is the case for the *gender* composition effect in the United Arab Emirates. Moreover, the absolute estimates are somewhat different and include deviations in both directions. The number of significant malleable course and school predictors in the final model is noticeably reduced compared to the course/school model without controlling. The number of significant effects varies between none in Kuwait (previously five) and six (previously eleven) in the United Arab Emirates. Again, *clear and structured teaching* emerged as the single most important component of malleable course level factors, being positively significant in five out of six countries. Other instructional variables are generally significant in all countries except Kuwait, but no regional pattern can be discerned. Similarly, input characteristics, either on school or on teacher level, play a certain role in all countries except Qatar – but with different variables showing significance in the region.

Variables related to the factor *time* show significant associations to mathematics achievement in Bahrain and Qatar, while variables related to the factor *opportunity* on teacher level are related to outcomes in Bahrain, Qatar, and in the United Arab Emirates.

After controlling, some of the model indicators also unexpectedly show a negative association with student outcomes. This affects the *education level* in Bahrain and Oman, and the frequency of *homework verification* in Bahrain. According to the literature review, it was also expected that, being a female teacher might be positively associated with achievement, if at all. However, in Saudi Arabia, and only after controlling for the student background, a difference of about 35 points in favor of male teachers is observed. These unexpected results are further examined and discussed in chapter 11.

Likewise, when comparing significant variable association before and after controlling for the student background, a number of relations become significant only *after* controlling for the background. This affects five variables in Bahrain, three variables in Saudi Arabia, and one

variable in Qatar. One likely reason for these differences is assumed to be the occurrence of so-called *suppression effects,* which also will be discussed in the subsequent chapter.

When comparing the results from the variance analyses with the student background model including student composition (Table 10-5), the additional explained variance differs quite a bit among countries. While in Kuwait the full model cannot really explain much additional variance compared to the background factors (7%), in Bahrain 21% of additional variance can be explained, and in the remaining countries between 15% and 18%.

Despite 16% explained level 2 variance and four school learning environment-related variables being significant in Saudi Arabia, the deviance is only minimally lower compared to the background model, and with an alpha level of 0.05, the difference is not significant – as can be seen in the lower part of Table 10-9.

Table 10-9: Course/ school-level model with controlling – mathematics

Level	Factor	Details	Explanatory Variable	BHR	KWT	OMN	QAT	SAU	ARE
School	Input	Resources	Library books	0.0 (0.1)	0.1 (0.1)	0.0 (0.1)	0.2 (0.1) **	0.1 (0.2)	0.1 (0.0) **
	Quality	Environment (SLE)	Emphasis of success	0.8 (2.1)	5.0 (2.6)	-4.6 (3.1)	5.0 (3.0)	0.2 (4.9)	6.1 (2.2) **
	Time		Absenteeism	5.1 (2.0) **	-0.3 (2.8)	2.1 (2.5)	3.5 (3.2)	3.9 (5.6)	4.4 (2.2) **
Course	Input	Teacher background	Gender of teacher	-1.0 (5.2)	2.0 (8.1)	21.0 (18.5)	-5.9 (6.9)	-34.7 (16.3) **	-4.4 (5.3)
			Education level	-6.8 (3.0) **	-1.9 (2.0)	-6.6 (1.8) **	-0.6 (1.8)	1.8 (2.0)	-3.1 (1.7)
			Time for development	2.8 (0.9) **	0.9 (1.7)	0.2 (1.3)	0.5 (1.2)	-1.9 (1.9)	0.8 (0.9)
			Confidence in teaching	-2.4 (2.4)	-3.5 (3.0)	2.8 (3.0)	-0.5 (3.1)	10.9 (4.8) **	-0.6 (2.2)
		Student composition	ESCS (avg.)	35.5 (4.7) **	49.3 (7.5) **	17.0 (6.9) **	53.4 (7.8) **	12.8 (10.8)	62.4 (4.9) **
			Early numeracy (avg.)	25.4 (9.0) **	25.6 (11.4) **	12.2 (10.7)	-10.9 (10.1)	-14.4 (13.9)	0.4 (7.3)
			Gender (avg.)	15.3 (4.7) **	18.5 (5.0) **	-3.6 (19.0)	-0.1 (6.8)	56.1 (16.8) **	7.3 (5.0)
			Non-nationals (avg.)	-20.3 (8.0) **	34.1 (14.7) **	-75.1 (14.3) **	10.6 (11.8)	-15.4 (22.6)	0.0 (7.7)
	Quality of Instruction	Structured teaching	Clear teaching	3.7 (1.8) **	-2.1 (2.3)	16.0 (3.2) **	8.2 (2.7) **	11.3 (4.8) **	6.9 (2.1) **
		Activation	Cognitive activation	4.6 (3.0)	1.6 (2.9)	-1.7 (3.1)	2.9 (3.7)	-7.8 (4.9)	-0.2 (2.4)
		Management	Disruptive students	1.6 (3.9)	6.3 (4.2)	4.0 (3.4)	0.7 (5.7)	18.3 (7.8) **	5.3 (3.2)
		Climate	Orderly environment	1.0 (1.9)	-2.7 (3.1)	8.0 (3.2) **	-0.2 (4.0)	-0.3 (4.6)	5.3 (2.2) **
		Assessment	Hmwk. verification	-3.7 (1.7) **	-1.5 (1.9)	3.1 (2.4)	1.2 (3.1)	3.2 (4.3)	-0.7 (1.5)
	Time		Time spent on subject	5.6 (1.6) **	0.8 (3.1)	1.0 (1.7)	3.9 (1.8) **	2.9 (5.8)	-0.1 (1.4)
			Amount of homework	-2.1 (1.7)	5.4 (7.3)	-1.4 (2.2)	0.0 (2.4)	-3.6 (4.6)	2.5 (2.3)
	Opportunity		Topics covered	7.0 (1.5) **	2.5 (1.8)	4.5 (2.4)	3.5 (1.7) **	3.5 (3.2)	7.3 (1.0) **
Student	Student Characteristics	Student background	ESCS	16.0 (1.4) **	10.8 (1.8) **	23.1 (1.6) **	12.2 (1.6) **	8.9 (1.7) **	12.5 (1.0) **
			Nationality	21.9 (3.6) **	22.5 (3.7) **	4.0 (5.1)	35.0 (3.4) **	30.9 (6.0) **	24.9 (2.1) **
		Aptitude	Early numeracy	9.5 (1.4) **	8.5 (1.5) **	12.2 (1.3) **	9.4 (1.5) **	9.6 (2.0) **	7.9 (0.8) **
		Motivation	Student likes learning	8.2 (1.2) **	6.5 (1.3) **	17.7 (1.7) **	8.9 (1.3) **	6.8 (1.7) **	10.0 (0.8) **
	Time		Absenteeism	-9.4 (0.9) **	-5.0 (1.1) **	-7.1 (0.8) **	-8.1 (1.1) **	-4.2 (0.9) **	-7.5 (0.4) **
	Opportunity		Parental help	-3.7 (0.9) **	-2.8 (0.9) **	-3.2 (1.2) **	-4.3 (0.9) **	-3.0 (1.1) **	-3.6 (0.4) **

Variance	Change in explained variance (Level 2)	41%	29%	27%	46%	28%	43%
	Change in explained variance (Level 1)	-1%	-3%	0%	-3%	0%	0%
Model Fit	Deviance	185218	581173	636585	214015	4793704	800625
	Significance of changes in deviance to SES model	0.000	0.978	0.000	0.000	0.029	0.000

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
() Standard errors appear in parenthesis
** $p \leq 0.05$

This indicates that perhaps another model, with fewer course- and school-level predictor variables, could better fit the empirical data in Saudi Arabia than the current model displayed here. Similarly, for Kuwait, the final model does not bring any improvement over the student background model, and also none of the framework predictors beyond student background influences show significant associations with mathematics outcomes. Due to the overarching goal of a regional comparative exploratory analysis, however, and to find possible relations with malleable educational factors in the region, the author chose to keep the models comparable among countries.

10.5.2 Science

Table 10-10 shows the final results for the science analyses with all predictors entered simultaneously. Again, the United Arab Emirates emerge as the country with the highest number of effects that remained significant (altogether nine), while Kuwait and Saudi Arabia only show two significant relations with achievement for malleable course- and school-level variables.

Table 10-10: Course/ school level model with controlling – science

Level	Factor	Details	Explanatory Variable	BHR	KWT	OMN	QAT	SAU	ARE
School	Input	Resources	Library books	0.1 (0.1)	0.0 (0.1)	0.0 (0.1)	0.1 (0.1) **	-0.2 (0.3)	0.1 (0.1) **
	Quality	Environment (SLE)	Emphasis of success	1.9 (3.2)	11.8 (3.7) **	-5.1 (3.6)	5.1 (3.2)	-1.3 (5.8)	7.5 (2.6) **
	Time		Absenteeism	9.3 (3.0) **	-2.2 (3.9)	4.0 (3.1)	3.5 (3.4)	7.4 (5.3)	5.6 (2.5) **
Course	Input	Teacher background	Gender of teacher	-4.0 (8.0)	37.5 (13.2) **	14.9 (15.3)	2.7 (7.9)	34.2 (16.2) **	8.7 (6.8)
			Education level	-1.3 (3.0)	1.0 (4.0)	-5.6 (2.8) **	1.6 (2.3)	0.5 (1.8)	-0.8 (1.6)
			Time for development	0.9 (1.3)	-1.6 (1.8)	1.6 (1.8)	1.1 (1.3)	-2.3 (2.1)	2.7 (1.1) **
			Confidence in teaching	-3.8 (2.9)	-1.6 (4.1)	-0.4 (4.3)	1.7 (3.8)	-6.7 (6.6)	4.2 (2.0) **
		Student composition	ESCS (avg.)	14.4 (6.6) **	63.3 (9.2) **	23.8 (8.3) **	47.0 (7.8) **	35.7 (11.8) **	60.2 (5.5) **
			Early numeracy (avg.)	26.6 (9.2) **	13.1 (16.9)	20.2 (11.2)	-7.0 (11.2)	-16.4 (18.6)	7.4 (9.0)
			Gender (avg.)	35.1 (8.0) **	30.1 (7.7) **	55.1 (27.2) **	11.2 (7.6)	33.9 (18.1)	17.0 (4.7) **
			Non-nationals (avg.)	-22.1 (12.6)	27.3 (27.5)	-112.3 (14.8) **	-3.8 (13.2)	-20.5 (26.4)	9.5 (10.3)
	Quality of Instruction	Structured teaching	Clear teaching	5.0 (2.9)	-1.2 (4.4)	21.3 (4.3) **	13.3 (3.9) **	1.6 (6.2)	6.9 (2.1) **
		Activation	Cognitive activation	1.1 (3.3)	5.4 (4.2)	6.2 (5.3)	7.0 (4.2)	3.6 (5.6)	-4.6 (2.5)
		Management	Disruptive students	5.0 (4.7)	-2.3 (7.0)	7.0 (4.5)	5.4 (5.1)	14.7 (8.2)	2.9 (3.8)
		Climate	Orderly environment	3.8 (2.6)	-3.4 (4.0)	-1.5 (3.9)	1.0 (3.8)	3.1 (5.1)	6.3 (2.6) **
		Assessment	Hmwk. verification	5.2 (2.8)	0.3 (2.8)	-4.4 (2.7)	-2.4 (2.3)	5.6 (3.8)	-0.7 (1.2)
	Time		Time spent on subject	-8.9 (4.2) **	-2.7 (4.8)	9.7 (3.1) **	-2.5 (2.6)	8.7 (3.9) **	-3.5 (1.6) **
			Amount of homework	-3.9 (6.2)	1.4 (9.5)	-2.8 (3.6)	-4.1 (5.2)	-18.0 (14.0)	-2.9 (3.1)
	Opportunity		Topics covered	1.4 (1.7)	-2.9 (1.9)	0.6 (1.9)	-3.4 (1.6) **	-2.2 (3.6)	3.9 (1.2) **
Student	Student Characteristics	Student background	ESCS	18.7 (2.7) **	13.5 (3.3) **	28.9 (1.9) **	15.1 (2.2) **	13.7 (1.9) **	16.0 (1.5) **
			Nationality	22.7 (7.2) **	27.7 (7.1) **	9.8 (5.0) **	44.1 (4.3) **	29.8 (5.7) **	34.1 (2.8) **
		Aptitude	Early numeracy	8.4 (1.6) **	9.9 (2.9) **	13.8 (1.5) **	11.3 (1.5) **	15.0 (2.5) **	8.3 (0.9) **
		Motivation	Student likes learning	17.5 (2.3) **	19.9 (3.2) **	25.9 (1.9) **	12.1 (1.6) **	7.2 (1.6) **	13.4 (0.9) **
	Time		Absenteeism	-10.9 (1.2) **	-5.0 (1.8) **	-7.0 (1.0) **	-11.1 (1.1) **	-5.6 (1.3) **	-8.6 (0.6) **
	Opportunity		Parental help	-3.1 (1.3) **	-2.5 (1.6)	-2.9 (1.3) **	-4.2 (0.9) **	-4.5 (1.1) **	-3.9 (0.5) **

Variance	Change in explained variance (Level 2)	31%	38%	24%	51%	41%	42%
	Change in explained variance (Level 1)	-2%	0%	-2%	-5%	0%	-1%
Model Fit	Deviance	189548	597677	653440	220048	5000442	822057
	Significance of changes in deviance to SES Model	0.161	0.157	0.028	0.002	0.612	0.000

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
() Standard errors appear in parenthesis
** $p \leq 0.05$

A comparison of results with the student model shows some differences related to the level 1 predictors in Oman. While in general the magnitude of the effects is similar, the estimate for the *nationality* effect is a bit higher (and its standard error slightly lower), which leads to a significant relation between the *nationality* variable and science achievement in the final science model. On the other hand, when all variables are entered jointly, the number of significant relations in terms of student composition variables becomes lower. In Bahrain, the composition in terms of *nationality status* is no longer significant in the final model, as is the case in Oman for the composition in terms of *early numeracy* and in Qatar and Saudi Arabia for the composition in terms of *gender*.

For science, and after controlling for student background, the table shows the *time spent on instruction* as the variable which is correlated with science achievement in most of the countries – albeit, interestingly, there seems to be a negative relation in Bahrain and in the United Arab Emirates. *Clear and structured teaching* is still significant in half of the countries, namely Bahrain, Qatar, and the United Arab Emirates. All other course- and school-related variables only

show significant effects in one or two countries of the region. As was the case for mathematics, no regional pattern can be discerned. Input characteristics play a certain role in all countries except Oman on school level, while teacher background characteristics seem to be important in all countries except Bahrain and Qatar, but again with different variables being significant in the different countries.

For science, variables related to the factor *time* show significant relations to outcomes in Bahrain, Oman, Saudi Arabia, and in the United Arab Emirates, while the predictor subsumed under the factor *opportunity* on teacher level is related to outcomes in Qatar and in the United Arab Emirates.

Similar to mathematics, for science, some of the model indicators also exhibit unexpected negative relations to student achievement, but to a lesser extent. In science, this affects the *time spent on teaching* in Bahrain and in the United Arab Emirates, and the number of *topics covered* in Qatar. These results will be further examined in the discussion section in the subsequent chapter.

Similar to the mathematics results, a number of variables become significant only after controlling for the background. This affects one variable in Bahrain, Oman, and Qatar, and two variables in Saudi Arabia and in the United Arab Emirates. This is assumed to be related to *suppression effects,* that will be further discussed in the next chapter.

When comparing the results from the variance analyses with the student background model including student composition (Table 10-6), the additional explained variance differs to some extent among countries. While in Kuwait the full model only explains an additional 11% of the science variance, in Qatar 27% additional variance can be explained. As was the case for the mathematics models, deviances in Saudi Arabia and Kuwait, and for science also in Bahrain, are only marginally lower compared to the student background model, and with an alpha level of 0.05, the differences in both countries are not significant .

10.6 Summary

This chapter described the results of the multilevel analyses. Altogether five different sets of models were analyzed: null models, level 1 student-level background models, background models including student composition variables on level 2, school- and course-level variables without controlling, and the final models with all framework variables entered jointly. The results of the null model analyses showed that the amount of between-group variances varied quite a

lot between the GCC countries, ranging from about 25% in Bahrain to nearly 60% in the United Arab Emirates in mathematics, and from about 22% in Oman to 51% in the United Arab Emirates for science.

The level 1 student background models showed significant associations between model variable and student outcomes across the region for nearly all variables and both subjects. Only the *nationality status* in Oman for both subjects, and the *parental help with homework* in Kuwait for science, was not significantly related to achievement. The amount of individual variance explained spans from 5% in Kuwait to 15% in Qatar in mathematics, and from 7% to 18% in for science in the same countries.

Not all of the student composition variables on level 2 which were entered into the subsequent set of models were significantly related to achievement in all of the countries of the region, and the predictors worked differently across the region. While the ESCS index was significantly related to outcomes in all countries except for mathematics in Saudi Arabia, and the *gender* composition was significant in all countries except for Qatar in mathematics, the *early numeracy skills* and the *nationality status* were only significantly associated to outcomes in two to three countries. The student composition models could explain between 12% (Oman) and 28% (Qatar) of level 2 variance compared to the level 1 background models in mathematics, and even up to 30% for science in the United Arab Emirates.

In a separate step, models including school- and course-level variables without controlling for the student background were built. *Clear teaching*, as an indicator for the *Quality of instruction,* emerged most consistently as an important variable across the region. The United Arab Emirates showed 11 significant relations (from altogether 16 variables) in both subjects, the highest number of statistically significant associations, while Saudi Arabia on the other end only showed two significant variables for mathematics and one for science. While variables from all main factors of the framework (*Input, Quality, Time*, and *Opportunity*) exhibited significant relations to mathematics and science achievement in at least one country, the pattern across countries is rather heterogeneous and not necessarily consistent across subjects. The school/course model without controlling for student's background could explain between 26% of the level 2 variance in Kuwait and 52% in the United Arab Emirates. In science, the explained amount of variance ranged from 33% in Kuwait and Oman to 53% in Bahrain.

In the final models, all framework variables were entered jointly. Adding the student background variables in some countries further reduced the number of significant relationships be-

tween school- and course-level variables; interestingly, especially in Bahrain and predominantly in mathematics (but also partly in other countries and in science), certain variables that were *not* significantly related to achievement in the model without controlling show significant associations with achievement in the final model. These effects may partly be explained by so-called *suppression effects*. Concerning the student composition variables, the *ESCS* indicator emerged as being significantly related to achievement in all countries and, except for Saudi Arabia, also in both subjects. The composition in terms of *nationality* is significant in Bahrain, Kuwait, and Oman for mathematics and also for science in Oman. While in Kuwait a higher share of non-nationals is positively related to students' outcomes, interestingly, in Bahrain and Oman, a negative association was detected – a finding which merits further investigation. The most relations between school- and course-level variables and achievement after controlling (leaving aside the student composition effects) can be found for Bahrain in mathematics (seven) and for the United Arab Emirates in science (nine).

After controlling, *clear teaching* as an indicator for the *Quality of instruction* also emerged as an indicator significantly associated with mathematics achievement in five countries of the region. In science, the amount of *time spent on subject* was significant for four countries, although interestingly, in Bahrain and in the United Arab Emirates, the relation is negative. Also after controlling, variables from all main factors of the framework showed significant relations to mathematics and science achievement at least one country, but no regional pattern could be discerned. While compared to the background models, including the composition variables, the full models cannot explain much more variance for mathematics in Kuwait, the amount of additional explained variance reaches 7% in Kuwait and 21% in Bahrain. For science, the additional explained variance ranges from 11% in Kuwait to 27% in Qatar.

11 DISCUSSION AND CONCLUSIONS

11.1 Introduction

This chapter is structured as follows. Firstly, a short summary of the study is presented (11.2), followed by the answers to the research questions (11.3). A broader discussion of the findings is then presented (11.4). The chapter continues with an outline regarding how the current research may contribute to scientific and practical knowledge (11.5), and offers recommendation for policy-makers and in regard to the TIMSS assessment (11.6). Thereafter, the limitations of the study are delineated (11.7), and further research suggestions are given (11.8), before final conclusions are drawn (11.9).

11.2 Summary

The current study tried to explain achievement differences of grade four students in the Gulf Cooperation Council countries (GCC) from a perspective of educational effectiveness research (EER) based on secondary analyses of TIMSS 2015 data.

In a first step, the educational context of the region was examined, as described in chapter 2. It was found that in spite of the similarities in terms of values, history, and languages, the region also exhibits larger variations in certain conditions which may affect learning and teaching in the GCC countries. One major difference is related to the share of non-nationals in the population, a figure which is highest in the United Arab Emirates and Qatar and lowest in Saudi Arabia. Other major differences between countries relate to the degree of conservatism in regards to religion, tradition, and family orientation, which are expected to influence learning opportunities – as well as the opportunities to participate in economy and society – differently.

To better understand the learning environments in the region from the perspective of EER, educational effectiveness literature, with a special focus on factors influencing student achievement (chapter 3), and subsequently theoretical frameworks and models of educational effectiveness, were examined (chapter 4). A literature review, in combination with the only comparable data source allowing for a regional comparison of the region, namely the TIMSS 2015 data (described in chapter 7), led to the following two main research questions: *To what extent does TIMSS 2015 reflect essential factors in terms of educational effectiveness research?* and *According to the framework specified, which educational factors are most effective from the perspective of EER with regard to learning outcomes on primary level in the GCC countries?*

To answer these questions, a theoretical framework was built that considered the special conditions in the region and simultaneously tried to take into account the limitations of the cross-sectional data at hand. The development of the study-specific framework is described in chapter 6. It is based on Creemers' comprehensive model of educational effectiveness (Creemers, 1994), which contains all essential features of an integrated effectiveness model, but for this research purpose was enhanced by other theoretical constructs and models in the field – mainly through adding an input dimension and refining certain elements.

The analysis was based on TIMSS 2015 data, which is a cross-sectional and multi-purpose assessment and as such not specifically designed for EER. However, the TIMSS assessment framework does seek to collect policy relevant data about the context for learning, and considers how educational opportunities are provided and how students use these opportunities (Mullis & Martin, 2016, p. 4). While considerable consequential overlap with predictors important for effectiveness research can be expected, certain limitations, such as controlling for non-educational home background influences, need to be taken into account. The steps undertaken to obtain a suitable, and at the same time parsimonious, student background measure is described in section 8.4.

To answer the research questions, variables from the TIMSS questionnaires were categorized according to the model factors of the developed framework, and factor, reliability, and correlation analyses were conducted to reduce the number of indicators and to work out the underlying constructs. Finally, different sets of multilevel models were performed in order to identify predictors that can explain variances in mathematics and science achievement, especially at the course and school level. Details of the procedures applied can be found in chapter 8. The results of the variable selection process, as well as for the factor-, reliability, and correlation analyses, can be found in chapter 9. These procedures finally resulted in 4 variables on school level, 16 variables on course level (including the aggregated student-level variables), and 6 variables on student level that were included in the subsequent multilevel analyses for mathematics and science, respectively. The results of the multilevel models are presented in chapter 10.

11.3 Answering the Research Questions

The following section explores in detail the research questions posed in chapter 5.

R1: To what extent does TIMSS 2015 reflect essential factors in terms of educational effectiveness research?

This research question was split up in two sub-questions, which are presented and answered separately in the subsequent sections.

- How should an EER framework that takes into account recent findings of educational effectiveness, the special educational conditions in the Gulf area, and the restrictions imposed by using cross-sectional large-scale assessment data be constructed?

To answer this research question, literature about EER and about the educational context in the Gulf region was examined. Related findings led to the decision to base construction of the framework for the current study on Creemers' integrated model of educational effectiveness (Creemers, 1994), as it places specific emphasis on the classroom processes of teaching and learning, and comprises all essential features of an integrated effectiveness model. Additionally, the main indicators of Creemers' model were well researched and empirically validated in several studies. However, Creemers' model was developed at the beginning of the 1990s, after which further developments in EER resulted in more differentiated and refined models, such as the dynamic model of educational effectiveness (Creemers & Kyriakides, 2008) and other more elaborated constructs, especially in the field of instructional effectiveness. The dynamic model was not deemed to be well-suited for use with the cross-sectional large-scale assessment data at hand. Instead, the framework was based on Creemers' model, but was complemented by more recent research findings in terms of instructional quality and climate. Taking into consideration the limitations of the data at hand and the specific situation in the Gulf area, the author decided to include a set of input factors on class and on school level beyond the observable teacher behavior already included in Creemers' work. As supported by the literature, it was expected that resource-related factors, important preconditions for effective teaching and learning, might partly be unequally distributed in the region despite its prosperity. Furthermore, it was assumed that analyses based on self-reported answers to background questionnaires would benefit from additional input information, related to teacher characteristics and qualification. Moreover, these input variables are partly malleable by government policies, and therefore relevant for policy-makers. Especially in terms of teacher characteristics, a large variation was expected in the Gulf region, as the teaching force consists to a large extent of expatriate teachers

from a variety of countries. Finally, the student composition in terms of different background variables was regarded as an important input factor on course level. The class composition was expected to influence the learning environment in the class beyond the effect of the student's individual background, and usually cannot be controlled by the teacher. Thus, from a theoretical perspective, the derived framework should be suitable to allow for an explorative study in educational effectiveness factors in the Gulf region. Having specified the framework, the next step was to match the developed framework with the TIMSS 2015 available data, which resulted in the subsequent sub-question.

- Can TIMSS 2015 grade four student, teacher, and school questionnaire data be used to give empirical support for the developed educational effectiveness framework in the GCC countries, using mathematics and science achievement as outcome variables?

To answer this question, background questionnaire data from principals, teachers, students, and parents were categorized according to the main factors and the sub-factors specified in the delineated framework. Altogether 9 questions from the principal questionnaire, 24 questions from the teacher questionnaire, 10 questions from the student questionnaire, and 8 questions from the parent questionnaire, many of which included several options, were regarded as matching the constructed framework from a theoretical perspective. Details are listed in APPENDIX A. For all main factors of the framework, which included the factors *input* (including the student background), *quality*, *time*, and *opportunity,* corresponding questionnaire variables could be matched. Nevertheless, the number of available indicators for the different model factors was unevenly distributed among the TIMSS background questionnaires. On one hand, several suitable indicators for the *input* factor, on school and teacher level, for the *quality of instruction* and for the *student background* could be identified. On the other hand, framework-related information pertaining to the main factor *quality* on school level was rather scarce: no information was available regarding the sub-factor *quality of instruction,* which comprises rules and agreements about classroom instruction, monitoring, and professional development. In addition, *the use of time* on school level, which includes regulations concerning the management of time, homework, etc., could only be matched using a proxy related to problems with *absenteeism*. Moreover, the only indicator found for *opportunity to learn* on school level was a rather generic question related to the *policies related to tracking* (SCQ-10), which showed no meaningful correlation with achievement in any of the GCC countries and thus ultimately was removed from the model. However, taking into account that the TIMSS assessment was developed as a multi-purpose study and not specifically designed for EER, and considering the rich array of needed information on class and student level, adequate coverage of the framework was

achieved overall. Further reflection on the theoretical relevance, as well as data reduction techniques such as factor analyses, index creation, and correlation analyses, were used to reduce the vast amount of more than 170 initial variables from 50 questions judged to be possibly relevant to their major underlying framework. The assignment of questionnaire items, along with a short comment about their treatment in data preparation procedures, can be found in Table 11-1 below.

Table 11-1: Factors identified from questionnaires according to the specified framework

Level	Factor	Factor - Details	TIMSS Question	Description	Short Description	Comments (see App. B for more details)
School	Input	Resources	SCQ-11	Number of computers in school	# of computers	dropped (low correlation)
			SCQ-13/A	Availability of school library and # of books	Library books	
			SCQ-14	Shortage of resources (M/S)	Shortage resources	factor
	Quality	Environment (SLE)	SCQ-15A-E,K-M	Emphasis on academic succes	Emphasis on success	factor
			SCQ-16D-J	School discipline and safety	School discipline and safety	dropped -> related variable on course level
	Time		SCQ-8A/B/C	Instructional time	Instructional time	dropped (low correlation)
			SCQ-16A/B & 17A/B	Problems with absenteeism	Absenteeism	
	Opportunity		SCQ-10A/B	Policies related to tracking	Tracking policies	dropped (low correlation)
Course	Input	Teacher background	TQ-G1	Teaching experience (years)	Teaching experience	dropped (low correlation)
			TQ-G2	Gender of teacher	Gender of teacher	
			TQ-G4	Teacher's highest education level	Education level	
			TQ-G5A/B	Teacher majored in edu. and subject (M/S)	Teacher majored in subject	TIMSS index used (ATDM/S05)/ dropped (low correlation)
			TQ-M2/S2	Confidence in teaching (M/S)	Confidence in teaching	index
			TQ-M10/S9	Time spent on professional development (M/S)	Time for development	
			TQ-M11/S10	Preparedness to teach subject (M/S)	Preparedness to teach	dropped (low correlation)
		Student Composition	HQ-20A/23A/13/SQ-G4	Average economic and sociocultural status	ESCS (avg.)	factor (course average)
			HQ-8A-C	Average early numeracy skills	Early numeracy (avg.)	factor (course average)
			SQ-G1	Average gender composition	Gender (avg.)	(course average)
			HQ17A/B & SQG6A/B	Average composition in terms of non-nationals	Non-nationals (avg.)	(course average)
	Quality of Instruction	Structured teaching	SQ-MS2 & MS5 (A/B/E/F/I)	Clear and structured teaching (M/S)	Clear teaching	(student perception - course average)
		Activation	TQ-G14	Cognitive activation	Cognitive activation	factor
		Management	TQ-G15D	Limitation of teaching (disruptive students)	Disruptive students	
		Climate	TQ-G6	Emphasis on academic success	Emphasis on academic success	dropped -> related variable on school level
			TQ-7D-H	Orderly learning environment	Orderly environment	factor
		Assessment	TQ-M7C/S6C	Verification of homework assignment (M/S)	Hmwk. verification	index
			TQ-M8A/S7A	Monitoring progress	Monitoring progress	dropped (low correlation)
	Time		TQ-M1/S01B	Teaching time spent on subject	Time spent on subject	
			TQ-M7A/B	Amount of homework assigned (M/S)	Amount of homework	
			TQ-M10/S9	Time spent on professional development	Professional development	
	Opportunity		TQ-M6/S5	Number of topics covered (M/S)	Topics covered	index
Student	Student Characteristics	Student Background	SQ-G1	Gender	Gender	Not used on L1 because of single-sex classes
			HQ13/SQ-G4	Books at home	Books at home	used for ESCS
			SQ-G5	Home possessions	Home possessions	dropped (low correlation)
			HQ-20A/B	Highest level of parental education	Highest education level	used for ESCS
			HQ-23A/B	Highest occupational level	Highest occupational level	used for ESCS
			HQ17A/B & SQG6A/B	Parents born in country	Nationality	Father born in ctry. used as indicator for the nationality
		Aptitude	HQ-8A-C	Early numeracy activities (number sense)	Early numeracy	factor
		Subject motivation	SQ-MS1/MS4	Student likes learning (M/S)	Student likes learning	factor
	Time		SQ-G8	Student's absence from school	Absenteeism	
	Opportunity		HQ-9BB	Parental help with homework	Parental help	

Note. SCQ – School questionnaire/ TQ –Teacher questionnaire / SQ – Student questionnaire /HQ – Home questionnaire

The subsequent sections will briefly describe the different variables and indicators that were kept for subsequent analyses steps, based on the current framework and in combination with the findings from the literature review.

School-level indicators

Related to the main factor of *input* on school level, two indicators for the factor *educational resources* were kept: an indicator of the *number of books in school library* (SCQ-13/13A), and the principals' judgement regarding *shortage of resources* (i.e. the extent to which the school's

capacity to provide instruction is affected by a shortage in resources for mathematics and science instruction, respectively; SCQ-14). The school learning environment (SLE), as a measure for the *quality* factor on school level, was characterized by items related to *emphasis on academic success* (SCQ-15).

Course-level indicators

On course level, a set of teacher characteristics and qualifications was kept as measures for the *teacher background,* which is defined as a category of the main factor *input*. The *gender of teacher* (SQG-2) was kept, due to indications from Ridge (2014) that different working situations and motivations of female and male teachers could be part of the explanation for the gender gap in the region. Additionally, the *confidence in teaching strategies* (here as a proxy for pedagogical knowledge) and the amount of *time spent for professional development* were kept as important *teacher background* characteristics.

The *student composition* here also was regarded as an *input* factor assumed to have a potential input related to the learning atmosphere according to the discussion in section 3.4.6. For the current study, the course averages of the student composition in terms of their *ESCS,* average *early numeracy skills*, average *gender* composition, and average composition in terms of *nationality status* were used.

The *quality of instruction* was divided into five separate sub-factors. *Clear and structured teaching* was measured by the class averages of related student answers to relevant teaching activities (SQ-MS2/5). While *cognitive activation* is a more recent factor on the agenda, the empirical evidence of associated learning gains is growing (Chapman et al., 2015; Hattie, 2009; Klieme & Rakoczy, 2003) and its importance is also under discussion in the GCC countries (Khan, 2015). As *classroom management* influences student attention and ultimately their *time on task*, it has empirical support concerning its relation with student outcomes (Brophy & Good, 1986; Doyle, 1985). The TIMSS 2015 questionnaires did not assess *classroom management* of teachers directly; therefore, a proxy indicating *teaching limited by disruptive students* (TQG-15D) was selected here. The *supportive climate* on course level was characterized by a TIMSS question related to the *orderly learning environment* of the school. Assessment and feedback strategies are repeatedly shown to have positive effects on motivation and learning gains (Creemers & Kyriakides, 2008; Teddlie & Reynolds, 2000). As an indicator for *assessment,* in this content an index based on the frequency of *verifying (correcting, discussing, and monitoring)* homework was constructed for the purpose of this study (calculated from TQ-M7C/S6C).

Time for learning was measured by two variables, namely the *amount of teaching time per week* (TQ-M1/S01B) and the *amount of homework assigned* (calculated from TQ-M7A/B for mathematics and TQ-S6A/B for science).

Opportunity to learn on course level was collected via the *average amount of topics covered up to the current school year* TQ-M6/S5).

Student level indicators

The main factor *student characteristics* was divided into three sub-factors, namely *student background*, *aptitude*, and *subject motivation*. The main indicator for the student background was the *ESCS* index, based on parental education (HQ-20A/B), parental occupation (HQ-23A/B), and the number of books at home (SQ-4/HQ-13), as described in section 8.4. Due to the huge differences among the GCC countries in terms of the percentage of non-nationals in the population, and due to the partly quite high differences in terms of achievement between both student groups, the *nationality status* was additionally included as an important variable to predict student achievement. As the region is also generally characterized by huge gender differences in favor of girls, it had made sense to include student's gender here. However, as in Saudi Arabia all schools are single-sex schools, and the other countries also show a high share of single-sex schools, the author decided to include the gender variable only on level 2 for the sake of comparability of the results across different countries.

International large-scale assessments are usually not able to capture any measures of students' *aptitude,* albeit *aptitude* is seen as an important predictor for student achievement (Reynolds, 1991; Reynolds & Walberg, 1991). For the current study, students' *early numeracy skills* before entering primary education, as judged by their parents, were included as a proxy measure for students' *aptitude*.

As a proxy for the *time used* by the students, a question related to the *number of absences from school* was administered in the assessment (SQ-G8). The *opportunities used* by students were measured by a question related to frequency of the *parent's assistance with their child's homework* (HQ-9). This variable, to a certain extent, could also be used to measure social interaction; correspondingly, social capital within a family was seen as relevant background information.

In summary, the TIMSS 2015 questionnaire variables exhibit a satisfactory coverage of the constructed framework in terms of indicators which were empirically demonstrated to be related to student outcomes, according to the effectiveness literature. From a theoretical perspective for all factors, except for the *quality of instruction* factor on school level, matching TIMSS

2015 variables or proxies could be found. However, questionnaire variables were not evenly distributed – a stronger focus was found to be placed on *input* related variables as well as on *quality of instruction* on teacher level. The factor *opportunity* on school level was finally dropped from the model, as the only matching variable related to *tracking policies* did not show meaningful correlations in any of the countries.

R 2: According to the framework specified, which educational factors are most effective from the perspective of EER with regard to learning outcomes on primary level in the GCC countries?

The research question is split in three parts, beginning with:

- How do the different educational effectiveness factors identified associate with students' mathematics and science achievement in the different GCC countries, when controlling for the home background?

Table 11-2 summarizes the results from the full models of the multilevel analyses, wherein all variables related to the framework were entered simultaneously. The table shows the model factors and details on the y-axis and the significant relations between predictor variable and achievement of the six countries on the x-axis. Cells for model variables that are positively related with outcomes in both subjects are marked in dark grey. Variables that are only significantly related to mathematics achievement are marked in light grey, and variables that are only significantly related to science achievement are marked in medium grey. Plain color markings indicate positive variable associations with achievement, while a dotted pattern indicates a negative relation to achievement. A few dark grey-colored cells with diagonal lines indicate significant relations to both subjects but in different directions for both subjects (see also the legend below the table). The sums below the table list the number of significant effects (independent from the subject that is concerned).

Table 11-2 : Significant indicators using mathematics and science as outcome variables

Level	Factor	Factor - Details	Explanatory Variable	BHR	KWT	OMN	QAT	SAU	ARE
School	Input	Resources	Library books				+		+
	Quality	Environment (SLE)	Emphasis of success		+				+
	Time		Absenteeism	+					+
Course	Input	Teacher background	Gender of teacher		+			+/-	
			Education level	-		-			
			Time for development	+					+
			Confidence in teaching					+	+
		Student composition	ESCS (avg.)	+	+	+	+	+	+
			Early numeracy (avg.)	+	+				
			Gender (avg.)	+	+	+		+	+
			Non-nationals (avg.)	-	+	-			
	Quality of Instruction	Structured teaching	Clear teaching	+		+	+	+	+
		Activation	Cognitive activation						
		Management	Disruptive students					+	
		Climate	Orderly environment			+			+
		Assessment	Homework verification	-					
	Time		Time spent on subject	-/+		+	+	+	-
			Amount of homework						
	Opportunity		Topics covered	+			-/+		+
Student	Student Characteristics	Student background	ESCS	+	+	+	+	+	+
			Nationality	+	+	+	+	+	+
		Aptitude	Early numeracy	+	+	+	+	+	+
		Motivation	Student likes learning	+	+	+	+	+	+
	Time		Absenteeism	-	-	-	-	-	-
	Opportunity		Parental help	-	-	-	-	-	-
				17	12	13	11	13	17

Notes. BHR = Bahrain, KWT = Kuwait, OMN = Oman, QAT = Qatar, SAU = Saudi Arabia, ARE = United Arab Emirates
** $p \leq 0.05$

Legend:

(positive)	(negative)	
+	-	** Mathematics only
+	-	** Science only
+	-	** Mathematics + Science
+/-	-/+	** Science (pos.) & Mathematics (neg.)/ Science (neg.) & Mathematics (pos.)

For a better overview, each educational level will be covered separately:

Factors associated with achievement on student level

On student level, altogether six variables were selected based on the categorization of TIMSS questionnaire variables according to the model factors of the developed framework, which can be seen in the lower section of Table 11-2.

All variables related to the main factors *student characteristics*, *time used*, and *opportunity used* show a significant relationship to mathematics and/or science achievement in all GCC countries. The factor *background* was conceptualized by two variables, the *ESCS* index and the *nationality status*. The *ESCS* is positively related to achievement in both subjects in all six countries. The before-mentioned variables related to the socio-economic background have consistently found to be associated with students' cognitive outcomes, since the seminal studies of Coleman et al. (1966) and Jencks (1972), who even asserted that they explain nearly all the variance in student achievement, and consequently concluded that schools would not make any difference. Indications supporting strong associations of background variables have often been

found by researchers, including Baker et al. (2002), Cervini (2009), Ehmke and Siegle (2005), Jungbauer-Gans (2004), McConney and Perry (2010), and Sirin (2005). Moreover, in multilevel analyses conducted in 34 TIMSS and PIRLS countries and benchmarking entities, Martin and Mullis (2013) reported that their home background indicator was the strongest predictor for achievement in nearly all countries. For the Gulf region, the importance of the students' socio-economic background was empirically supported by Ridge et al. (2013) in a study on early male school dropouts in the United Arab Emirates, and for Saudi Arabia by Wiseman, Al Sadaawi, and Alromi (2008), based on an analyses of TIMSS data.

Regarding the *nationality status,* significant relations in favor of non-nationals are found with both subject outcomes in five out of six countries; in Oman, non-nationals only perform significantly higher for science achievement. The results mirror the overall differences between nationals and non-nationals which are prominent in both subjects, but even more in science for all GCC countries, with the exception of Oman (see Table 9-3 and Table 9-4). The disparities in terms of *nationality status* will be further discussed in section 11.4.

The variable *early numeracy skills*, which is related to students' competencies in counting, recognizing, and writing numbers before entering primary school, was used as an indicator for the factor *aptitude* and is also related to achievement in all countries and in both subjects. In this context, *aptitude* can be understood as the prior knowledge of students, as reported by the students' parents. In the literature, *aptitude* is regarded as an important predictor for student achievement, as students with higher *aptitude* would need less time for learning (Reynolds, 1991; Reynolds & Walberg, 1991; Teddlie & Reynolds, 2000).

The factor *time used* or *time on task* is characterized by the rates of *absenteeism* indicated by the students. A higher rate of *absenteeism* allows less time for learning, and consequently is significantly negatively related to mathematics and science outcomes. This is consistent with the expectations and findings that students who use more time on task make better educational progress (Carroll, 1963; Creemers, 1994). The underlying concept of *time on task* is discussed in section 3.3.2.

An interesting finding is the negative association between achievement and the factor *opportunities used,* which describes the students' use of the experiences that were offered in the instructional process (Creemers, p. 118). As an indicator for *opportunities* on student level, the variable *parental help with homework* was selected. A possible explanation for the negative association with achievement might be that parental help is offered to weak students to a greater extent, which would switch the causation: instead of more parental help leading to higher

achievement, lower achievement would lead to a higher amount of parental help. This result is quite similar to findings by Cho (2010) in an educational effectiveness study; comparing the educational systems of Korea and South Africa using TIMSS 2003, the author found that "extra tutoring" was negatively related to achievement in South Africa. The author also explained this finding by the need of extra tutoring for students that were lagging behind.

Factors associated with achievement on course level

The researcher was specifically interested in educational factors on the course level, as it is the classroom where teaching and learning predominantly takes place. Out of 16 variables categorized into the four main factors on course level, namely *input*, *quality of instruction*, *time for learning*, and *opportunity,* 14 showed a significant relation with mathematics or science in at least one country, as can be seen in the middle section of Table 11-2.

The factor *input* is divided further in two components: *Teacher background*, which comprises the teacher characteristics and qualifications, and *Student composition*. *Student composition* will be handled first, as this block contains the main student background indicators already discussed in the section on student-level variables; in this context, however, they are rather included as course-level aggregates. According to the literature, many scholars assume that certain student background variables have a relation with student achievement beyond the individuals' background (for example, Baumert, 2006; Coleman et al., 1966; Harker & Tymms, 2004); nevertheless, disagreement about the magnitude and exact nature of the effect remains (see section 3.4.6 for a further discussion of this topic). For the current study, these variables were categorized as *input* variables as they constitute, similarly to the input in terms of teacher characteristics or educational resources, a kind of input for the processes related to teaching and learning in schools. In many more recent educational effectiveness studies these student composition variables are included in the models to allow for a better disentanglement of educational school influences from influences outside school (for example, Kyriakides & Charalambous, 2005; Lamb & Fullarton, 2001; Martin & Mullis, 2013). Likewise, this research project includes student composition variables in the final model, allowing for a more comprehensive controlling of out-of-school influences. The most consistent effect was found for the average *ESCS* index, which shows a significant composition effect in relation to science achievement in all countries, and also to mathematics achievement in all countries except Saudi Arabia. The average *Early numeracy skills*, on the other hand, only display significant relations to outcomes in Bahrain and Kuwait (and in the latter, only in mathematics). A *gender* composition effect, indicating that higher percentages of female students in the courses are associated with higher

achievement, can be found in all countries except Qatar – albeit not always in both subjects. However, this effect may be overestimated, due to the fact that no corresponding gender variable was included on level 1 of the models. In general, for the Gulf region, these results support Reynolds et al. (2014, pp. 208–209), who reviewed the related literature and found that most studies indicated a high average achievement, a high proportion of girls, and a high average socio-economic status to have positive effects on achievement.

A very interesting and unexpected result is the negative association between the percentage of non-nationals in the course and achievement in Oman (for both subjects) and Bahrain (mathematics only). While Oman is the only Gulf State where non-nationals, on average, did not score significantly higher than nationals, the difference in favor of non-nationals for Bahrain is significant: 12 points in mathematics and 13 points in science (Table 9-3 and Table 9-4). Still, results here indicate that, on average, a higher share of nationals in the courses in both countries is related to higher achievement, and with high differences in terms of absolute values as can be seen from Table 10-9 and Table 10-10. When analyzing TIMSS 2015 mathematics data related to the schools attended by nationals and non-nationals (Table 9-7 and Table 9-8), it is evident that national students who attend schools without any non-nationals are scoring exceptionally high while non-nationals in non-national-only schools score relatively low – but only in Bahrain and in Oman. In other GCC countries (except for Saudi Arabia where such schools don't exist), students in non-national-only schools score far higher than nationals in nationals-only schools. Therefore, the author hypothesizes that the negative association between the share of non-nationals is based on a group of elite schools attended by nationals only, while schools for non-nationals seem to be held to a lower standard.

In regard to the *teacher background,* which is the second category subsumed under the main factor *input,* after controlling for students' home background, the teachers' gender was still significantly associated with achievement in favor of female teachers in Kuwait and Saudi Arabia in science, even when jointly measured with other teacher characteristics such as their *education level* or *time spent for professional development*. To a certain extent, this may be an indication supporting assumptions made by Ridge (2014) that female teachers are often better educated, integrated, and more highly motivated when compared to male teachers who are often expatriates working in precarious situations (Ridge, 2014). However, causation could also function in reverse: Saudi Arabia and Kuwait are the countries with the highest share of gender-segregated schools, and in such schools, female teachers are predominantly teaching higher achieving girls. An interesting finding for Saudi Arabia, after controlling for the background indicators, is that the direction of the association in mathematics changes: after controlling for

the students' composition in terms of *gender*, an association in favor of male teachers was found. Additional analyses (Table 9-9 and Table 9-10) revealed that the gender of the teacher mattered far more in boys-only schools. While overall results indicate that female teachers (who mainly teach higher-achieving female students) generally accomplish, on average, higher outcomes; for mathematics in Saudi Arabia, both gender groups achieve lower achievement when taught by female teachers – but the difference in boys' schools is by far larger (an insignificant 10 score points in girls' schools versus a significant 48 points in boys' schools). One speculation for this finding is that due to the conservative nature of the society, and the traditional gender roles of the Saudi society, it might be more challenging for female teachers to assert themselves when teaching in boys' schools. The results from the final model for mathematics for Saudi Arabia also show a significant positive relation in terms of *gender* composition in favor of girls, and problems with *disruptive students*. This pattern related to the teachers' gender in Saudi Arabia might show a certain similarity to the difficult role of foreign teachers in the region, who are partly "perceived by their national students and parents as inferior because they come from poorer Arab nations" (Ridge, 2014, p. 119). Further investigation could be useful in unpacking the relation with classroom disorder and female teachers' job satisfaction in boys' schools. However, the question of why this finding holds only for mathematics might then be raised. In science, female teachers for both gender groups are associated with a higher achievement of about 30 score points. Moreover, results should be regarded with caution – as the tables also show that the rates of teachers of a different sex than their students in Saudi Arabia are very low. In boys' schools, only 5% of teachers are female, while in girls' schools, only 3% of teachers are male.

Interestingly, while *teachers' highest education level* shows significant relations with mathematics outcomes in Bahrain and Oman, the relation in both cases is negative. The *TIMSS 2015 Mathematics Teacher Almanacs* (Foy, 2017, pp. 4–5) reveal that the mean achievement by education level drops, starting with a bachelor degree (albeit this group comprises the majority of teachers in both countries: 86% in Bahrain and 66% in Oman), and continues to drop with higher degrees. The same pattern can be seen in Oman for the science results. This drop is not seen as consistently in other countries like Kuwait and Qatar, or at least is far less pronounced. In the literature, findings related to the association of formal qualifications – like the highest *educational level* – with achievement seems to be somewhat inconsistent as discussed in section 3.3.5.1, and do not always clearly indicate a higher achievement as associated with a higher formal education level. Findings of Blömeke et al. (2016, p. 21), which indicated the ISCED level as the strongest predictor for student achievement across countries in their TIMSS 2011

analyses, hence cannot be confirmed with the results of the current analyses. A negative relation of the teacher's level of education with science achievement was also reported by Anderson (2012) based on TIMSS 2007 data, which is notable for Dubai (a city in the United Arab Emirates) – which in that cycle of TIMSS participated as a benchmarking participant. Anderson related these findings to the type of postgraduate education teachers receive in the United Arab Emirates, assuming that higher education there would not comprehensively cover science content and pedagogy, and therefore may not be effectively transferable to the classroom. While this could be part of the explanation for the current findings in Bahrain and in Oman, in this case an additional explanation is hypothesized: throughout the GCC region, the teaching profession in general is regarded as a low status profession, especially for males, and often even as a "profession of last resort" (Ridge, 2014, p. 117). Highly educated staff in particular might therefore feel overqualified, and hence less motivated to teach – instead, feeling motivated to find a higher-level administrative position, or a better position in other employment areas, instead of focusing on the teaching profession. On a more general level, problems with dissatisfied teachers moving away from the teaching profession to seek higher-status jobs is reported by Ridge (2014, p. 26) for Kuwait, and it is safe to hypothesize that there is a similar effect found in other countries of the region. Dr. Al Awadi, the Bahrainian TIMSS research coordinator, also supported the notion that teachers who enter the teaching profession with a bachelor degree would often later lose interest in teaching, and rather migrate towards the private sector after obtaining higher degrees during the course of their teaching career (Dr H. Al Awadi, personal communication, February 15, 2018).

The amount of *time spent for professional development* only seems to play a relevant role in Bahrain (for mathematics) and in the United Arab Emirates (for science). It was found that professional development may improve student achievement, but it seems that only longer training programs have a measurable influence in classroom culture and practice (Supovitz & Turner, 2000; Yoon et al., 2007). In addition, the content and quality of the training also matter, according to Supovitz et al. (2000); this information, however, cannot be retrieved from the TIMSS questionnaires. Blömeke et al. (2016) found development activities particularly important for Asian and Arab countries, a finding that can only partially be confirmed with the current analyses. Teachers' *confidence in teaching strategies*, which was used as a proxy for their pedagogical knowledge, only shows significant relations with mathematics in Saudi Arabia and with science in the United Arab Emirates. In the literature, pedagogical knowledge is fairly consistently related with student performance (see for example Ashton & Crocker, 1987;

Evertson et al., 1985; Monk, 1994), albeit teacher knowledge was often only indirectly measured by *teacher's formal education*, as discussed separately in the section above. More recent research combines pedagogical knowledge with subject matter knowledge to a common dimension of *pedagogical content knowledge* (Baumert et al., 2010; Hill et al., 2008; Tatto et al., 2012). The content perspective was measured in TIMSS 2015 by questions related to teachers' confidence in teaching the various TIMSS topics. However, in this context, content knowledge was not found to be related to achievement during the preparatory correlation analyses in any of the countries under consideration, and therefore was dropped from the final multilevel analyses step.

The main factor *quality of instruction* was further categorized into the five dimensions: c*lear and structured teaching*, *cognitive activation*, c*lassroom management*, *supportive climate*, and *assessment*. While each factor except for *cognitive activation* showed a relation in at least one country, the most consistent variable that emerged was the aggregate of student variables related to the concept of *clear and structured teaching*. The variable was found to be significantly associated with achievement in all countries except Kuwait, albeit only in Oman, Qatar, and in the United Arab Emirates for both subjects, while Bahrain and Saudi Arabia showed significance only for mathematics. Strong support for the importance of well-structured whole-class teaching has been found in several intervention programs and classroom observation studies and reviews (Chapman et al., 2015; Good & Grouws, 1979; Hattie, 2009; Mortimore et al., 1988; Muijs & Reynolds, 2000); it seems that direct teaching methods, as described in section 3.3.5.2, have also found justification in the era of constructivism.

However, *cognitive activation* strategies, which according to the definition of Klieme and Rakoczy (2003, p. 335), require higher-order thinking skills and enable students to really understand what was taught to them, did not emerge as a significant variable in any of the GCC countries. While these more recent constructivist approaches are discussed in the Gulf region, at the time of writing such practices have only found limited utilization and teachers are still trained in the "traditional" way (Khan, 2015, p. 9). In other regions, however, there is quite some indication for the importance of learning gains based on cognitive activation strategies. Klieme and Rakoczy (2003, p. 336), for example, identified cognitive activation in a review of data from the TIMSS video study as one of three major dimensions associated with student outcomes; however, no Gulf country was included in the study. In the Gulf region, such teaching practices seem to be in nascent stages of implementation. The author further notes that a broader set of items related to cognitive activation was only included in the most recent cycle

of TIMSS – and that items might better capture this rather complex concept after some elaboration in future cycles of the assessment.

Teaching limited by disruptive students was used as a proxy for *classroom management* capabilities of the teachers, but only showed a significant relation to mathematics achievement in Saudi Arabia. Classroom management, which affects students' attention and thus is regarded as an important precondition for their *time on task*, has repeatedly been described in the literature as an important factor with influence on students' learning gains (Brophy & Good, 1986; Doyle, 1985). The *supportive climate* was conceptualized via questions related to an *orderly learning environment*. This variable was significant for Oman in both subjects. An orderly climate is also partly associated to teachers' classroom management skills, and in general is an important precondition for effective learning (Levine & Lezotte, 1990; Mortimore et al., 1988; Sammons et al., 1995); correspondingly, in the multilevel analyses of Martin and Mullis (2013), orderly climate was also found to be one of the most important factors related to achievement in Oman, Qatar, and the United Arab Emirates. Finally, the author created an index from teachers' strategies to handle *homework verification* as a measure for the factor *assessment*. A stronger focus on *homework verification* was found to be negatively related to achievement in Bahrain. *Assessment,* according to the literature, is important for students' learning gains, if it is used in a formative way to identify students' needs and to adjust teaching approaches (Creemers & Kyriakides, 2008; Teddlie & Reynolds, 2000). However, overemphasis has shown a decreasing effect (Bangert-Drowns, Kulik, & Kulik, 1991; Black & Wiliam, 1998). Another explanation for the negative association could be that a special emphasis on homework verification is mainly needed for weak students, which would switch the causation direction – indicating that homework is more strictly verified in classes with mainly weak students.

Time for Learning

Two variables were kept to measure *time for learning*. Firstly, the *instructional hours per week* spent teaching mathematics and science, respectively; secondly, the average *amount of homework assigned*. In the final model, while the latter was not significantly related to outcomes in any country, the number of instructional hours emerged after controlling for the background as a predictor for student achievement in all countries except Kuwait. Only in Bahrain was the result significant for both subjects; for Oman and Saudi Arabia, significant results were seen for science and for Qatar in mathematics. Interestingly, in Bahrain and in the United Arab Emirates, the amount of *time for learning* was negatively correlated to student achievement in science – a finding which merits further investigation. In general, more time allowed for learning

should have a positive influence on student learning; hence, the amount of time available for instruction has been an important factor in effectiveness models since research performed by Carroll (1963) over 50 years ago. While most empirical studies indeed show positive results, Elley (1992, p. 40) using reading literacy data, indicated a negative relation between the number of school days and reading achievement beyond a threshold of around 180 days per year – but did not provide an explanation for these counter-intuitive findings. Concerning the overall time allowed for learning, it should be noted that the TIMSS questionnaire only collects information about the gross amount of time available on school and teacher level. Stronger relationships related to the *time for learning,* therefore, may be found if the questionnaire included more detailed information about the amount of available time that is actually used for effective teaching. Additional findings were provided by the analysis on effective teaching time carried out by Sandoval-Hernández et al. (2013), which is shortly described in section 3.3.2. Moreover, time for learning on its own is not sufficient; it rather needs to be filled with opportunities. This could also be a possible explanation for the negative results in Bahrain and the United Arab Emirates. If extended learning time is of poor quality, for example because policies on extended school time resulted in employment of less prepared teachers due to the shortage of teachers in these countries, then more time might not necessarily add positively to student outcomes (Hincapie, 2016). Empirical evidence about an association between homework and outcomes is less conclusive as can be seen from the discussion in section 3.3.2.

Opportunity to learn

Students*' opportunity to learn* was measured by the number of TIMSS-specific *topics covered* up until the time of testing. While *topics covered* was positively related with achievement in the United Arab Emirates for both subjects, and for Bahrain and Qatar in mathematics; it was negatively related to achievement for science in Qatar. In this context, *opportunity to learn* is essentially understood as the curriculum alignment of classroom practices (Scheerens, 2016, p. 55), but here only can be evaluated as the alignment to the TIMSS framework. As such, higher coverage would be expected to be associated with higher student outcomes, as for example reported from the results of different IEA studies (Comber & Keeves, 1973; Postlethwaite & Wiley, 1992). On the other hand, in order to master a topic properly, a certain amount of time and practice is necessary – as laid out by Bloom (1968) in his model of mastery learning. From Table 2-4, it is possible to derive from column 8 (Science – "TCMA") the coverage of Gulf countries' curriculum in regard to the TIMSS science test, which shows curriculum coverage of 100% only for Qatar. A hypothesis for the negative relation to achievement, which cannot be further verified here, is that all these topics cannot be sufficiently covered in

an in-depth manner during the available instructional time; teachers who are able to restrict their teaching to more important concepts and topics, therefore, might manage to accomplish better TIMSS assessment results for their students.

Factors associated with achievement on school level

On school level, all three factors show significant associations with achievement in the region (see the upper section in Table 11-2). Similarly to the course level, on school level an *input* factor was also included in the research framework. Here, the *input* rather focuses on the *availability of educational resources*, as described by Creemers (1994, pp. 105–106), albeit Creemers included these into his process factors. Research evidence regarding the association of resources with achievement is rather mixed. As detailed in the discussion in section 3.3.6.5, it can be concluded that educational resources might be a precondition for effectiveness to a greater extent in developing countries. While the GCC countries rate among the richest countries in the world (see Table 2-1), wealth tends to be distributed unequally, and does not always reach students. Ridge (2014, p. 53), for example, described for Qatar that although 4.1% of the total income is routed to the educational sector, much of the money benefits staff at the Ministry of Education even while some schools face severe shortages and high numbers of student per class. The results show that the number of *library books* in the school, used in this context as an indicator for *educational resources*, is still significantly associated with outcomes in Qatar and in the United Arab Emirates, for both subjects, after controlling for the home background; this lends some support to potential effects surrounding educational resources and partly unequal distribution of resources in both countries.

As the TIMSS questionnaires did not include any questions about rules and procedures related to the quality of instruction, the *quality factor* here only consisted of an indicator for the *school learning environment,* which was measured by several question options related to the *emphasis on academic success*, as rated by the principal. The learning environment seems to be relevant in the United Arab Emirates for the outcomes of both subjects, and for science in Kuwait. Similar to the discussion about *supportive climate* on course level, an orderly atmosphere and positive disciplinary climate, in combination with a high emphasis on academic success, have been found to be important preconditions for effective learning in nearly all effectiveness reviews. Sammons (1999) and Levine and Lezotte (1990), for example, found that schools wherein members of the staff are committed to a school-wide mission focusing on academic improvement have frequently proved to be more effective. Policies related to creating a positive school

learning environment, and actions taken for improving it, constitute one of the main pillars of the dynamic model developed by Creemers and Kyriakides (2008).

The *time* factor was conceptualized by the degree to which the principal sees *absenteeism* of teachers and students as a problem in his or her school. The variable was coded in such a way that fewer problems are associated with higher expected achievement. *Absenteeism* seems to be a certain problem in Bahrain and in the United Arab Emirates. These findings are in line with the statements of participants in a Gulf seminar on TIMSS 2011 data in Qatar (Khan, 2015), where it was repeatedly reported by ministry officials of GCC countries that *absenteeism* (in addition to bullying) was a huge problem in their schools.

- Do effectiveness factors operate in a similar way in the region for both subjects, and can a regional pattern been identified?

When looking at the distribution of significant model factors for the different levels, it is possible to derive from the lower section of Table 11-2 that all six variables related to the main factors on student level are significantly related to achievement in all of the six GCC countries under investigation. The direction of the association is equal in all countries for each of the variables, and, with only two exceptions, an association for both measured outcomes can be determined. Thus, on student level, factors in general operate in a similar way across countries. Nevertheless, the strength of the relation is somewhat different among countries (see Table 9-28 for mathematics and Table 9-29 for science) and there are two notable exceptions: *nationality status* and *gender* (the latter, on student level, was not included in the final model for the sake of comparability): In Oman, the average achievement results from TIMSS 2015 show no significant differences in achievement in both subjects between nationals and non-nationals, in contrast to the pattern found in all other countries (see Table 9-3 and Table 9-4). For the students' gender, on the other hand, there are strong differences, especially for Saudi Arabia in science which amounted to 79 score points (see Table 9-2); no significant gender differences were found for mathematics in Qatar and in the United Arab Emirates (see Table 9-1).

The course level, in contrast, rather shows a heterogeneous pattern among the six countries (see the middle section of Table 11-2). Five countries show significant associations between the *ESCS* index and the outcomes in both subjects; in Saudi Arabia, the association is only significant for science. The most consistent finding among variables related to *quality of instruction* is that *clear and structured teaching*, as reported by the students, was found to be associated with higher achievement in all countries except for Qatar – albeit partly only for mathematics outcomes. Further, the *time factor* plays a role in all countries except Qatar, but mostly only for

one subject (and partly with a negative relation to achievement). While no common pattern can be discerned, and there is quite some variability in the conditions of the different countries (more details can be found in the discussion section below), we can see that in all countries except Qatar there are associations between *teacher background* variables and student outcomes. All countries show effects in terms of their *student composition,* again with a somewhat different variable pattern. Variables related to the factors *quality of instruction* and *time* are relevant in all countries except in Qatar, and the measure for *opportunity to learn* still shows significant associations in three countries.

With regard to the number of significant relations to achievement, the analyses show that in Qatar only one model variable (namely, the *gender of teacher*), and only for science, exhibits a relation to student outcomes beyond variables related to the home background; on the other hand, six model predictors are significantly related to achievement in the United Arab Emirates. Thus, it can be concluded that the framework works quite differently throughout the region, and seemingly functions better in the United Arab Emirates and Bahrain than in Kuwait – where TIMSS variables were only able to explain variance and significant relations to achievement by student background characteristics.

Similarly, no regional pattern concerning the subject-specific outcomes and certain model factors can be discerned, apart from student-level factors. While certain variables, such as the *ESCS* indicator or *clear and structured teaching*, tend to be relevant in most countries for both subjects, other factors show quite a heterogeneous pattern in terms of subject-specific associations. In general, it seems that many predictor variables are more strongly associated with mathematics education in Bahrain, while it tends to be the opposite case in the United Arab Emirates.

In terms of significant relations between the three model variables and student outcomes on school level, no regional pattern across countries can be found either. It rather seems that the educational conditions in all six countries vary to a certain extent – despite their wide historical and cultural similarities, which will be discussed further in section 11.4.

- To what extent do the educational effectiveness factors identified for the region explain differences between the GCC countries, after controlling for the student background?

Overall, the countries of the Gulf area differ quite significantly in terms of the predictors explaining student outcomes in mathematics and science, but also in the extent to which different groups of factors can explain achievement differences in the region. The main outcomes related

to the variance components of the calculated multilevel models are summarized in Table 11-3 for mathematics and in Table 11-4 for science. In the second column, the tables display the percentage of between-course/school (or simply level 2) variance obtained from the null models. In the third column, the results for the student background models, including level 2 composition variables, are displayed. Finally, the fourth column shows the results from the final models wherein all model variables are entered jointly into the multilevel analyses. The tables show that the United Arab Emirates exhibits the largest variance components between schools (59% for mathematics and 55% for science), followed by Qatar. The least achievement variance between courses occurs in Bahrain in both subjects (24% for mathematics, and 29% for science). The tables show that the levels of between-course/school variance for all countries are fairly similar across subjects. The mathematics figures for Oman, Qatar, and Saudi Arabia are well in line with the between-school variances reported by Martin and Mullis (2013, pp. 139–140) for multilevel analyses based on TIMSS 2011 data of those countries, except that they reported a somewhat lower between-school variance for the United Arab Emirates (45%). The international average between-school variance for the 34 countries included in their analyses came to 26%; this indicates that, especially for the United Arab Emirates and Qatar, the amount of variance between courses/schools is on the higher end. Comparing the between-course variances with the science analyses conducted by Martin and Mullis (2013) on TIMSS 2011 data, the estimates match well for Oman, are about 10% lower for Qatar and Saudi Arabia, and about 10% higher for the United Arab Emirates. The international average for the proportion of between-school science variance reported by Martin and Mullis (2013) came to 25%. However, comparisons should only be made with caution, as for the current analyses the *course* level – with disaggregated school-level variables and separate courses for multiple teachers of the selected students – were defined as the group level, while in Martin and Mullis (2013) the school-level was specified as the level 2 component. Besides, Martin and Mullis used analyzed data from a previous cycle of TIMSS.

A possible explanation for the high between-group variances in the United Arab Emirates and Qatar is conceivably related to the remarkably high share of non-nationals (see Table 2-3) and their children in schools of these countries (see Table 9-3), and will be discussed further in the discussion section (section 11.4).

The amount of between-course/school variance that can be explained by student composition variables ranges in mathematics from 12% in Oman and Saudi Arabia to 28% in Qatar. For science, the explained variance ranges from 11% in Oman to 30% in the United Arab Emirates (column 3). The variance components explained by the background model is similar across

subjects, with the exception of Saudi Arabia where these variables explain considerably more variance in science compared to mathematics. The difference in the amount of explained variance is again assumed to be related to a certain extent to the share of non-nationals and associated heterogeneous backgrounds.

Column 4 shows the variance components explained by the full model. The differences to the previous background model range from only 7 percentage points in Kuwait to 21 percentage points in Bahrain for mathematics, and from 11 percentage points in Kuwait to 27 percentage points in Qatar. In general, the additional amounts of variance explained by further input and school environment variables, after controlling for the background, is somewhat limited; in general, more than 50% of the level 2 variance remains unexplained. A comparison of the findings with the educational effectiveness analyses of Martin and Mullis (2013) reveal, for the four GCC countries that were included in their analyses, a somewhat lower amount of explained variance for the full model of the current analyses in both subjects, especially for Qatar and the United Arab Emirates. On the other hand, the explained group variances of the full model in Saudi Arabia, and for the Home Background model in Oman, are somewhat higher.

Notably, in addition to the somewhat different specifications of the group-level as discussed above, student-level variables for the current analyses were centered on the grand-mean, as opposed to the analyses by Martin and Mullis (2013) wherein a group-centered procedure was applied. Moreover, Martin and Mullis' analyses were based on the previous cycle of TIMSS (TIMSS 2011), and conditions might have changed during the four year interim. Nevertheless, when applying a group-centered approach for the current analyses (see APPENDIX D), the explained variance components for both the student background model, including composition variables, as well as for the final model are in general somewhat higher when compared to the corresponding results reported by Martin and Mullis (2013). A higher share of explained variance, particularly for the background models in this study, could be explained through the addition of gender composition and the nationality composition on group level. Furthermore, a more comprehensive approach in the identification of effectiveness-enhancing factors which are specifically relevant for the Gulf area was followed here.

Consulting the effectiveness literature, it seems that after controlling for intake factors for both subjects, the amount of group-level variance in the region explained by school educational factors occasionally exceeds findings from earlier effectiveness research – amounting to levels of around 5 to 15% (for more information on the magnitude of school effects, refer to section 3.4.2.). It also appears that the shares of variance explained using a group-centering approach

better match the ranges indicated in the literature review; however, no information about the centering approaches in these comparison studies was available to the author. The fact that the proportions of explained variances profoundly depend on the centering approach chosen highlights the importance of a clear indication of the chosen approach for EER studies.

In general, it can be deduced that school-level variables will likely show less variance than teacher- or even student-level variables, as schools can be assumed to be more homogenous than teachers or home characteristics (Rutter, 1983, pp. 3–4). This implies that lower-level variables can usually be assumed to have greater effects, and explain more variance. From a policy perspective, however, low percentages of variance explained through school characteristics might also be relevant for school improvement efforts (Teddlie & Reynolds, 2000, p. 98). Notably, the course/school models without controlling explained between 26% and 53% of the level 2 variance (Table 10-7 and Table 10-8). The variance explained by the final model was obtained by a conservative approach, first entering student variables and assuming that all such variables are solely affected by home background effects.

Table 11-3: Mathematics variance components and variance explained on group level

Country	% of variance in mathematics achievement that is between courses	% of between-course variance attributable to level 1 + 2 home background	% of between-course variance attributable to home background and full model
Bahrain (BHR)	24	20	41
Kuwait (KWT)	30	22	29
Oman (OMN)	28	12	27
Qatar (QAT)	41	28	46
Saudi Arabia (SAU)	36	12	28
United Arab Emirates (ARE)	59	26	43

Table 11-4: Science variance components and variance explained on group level

Country	% of variance in science achievement that is between courses	% of between-course variance attributable to level 1 + 2 home background	% of between-course variance attributable to home background and full model
Bahrain (BHR)	29	15	31
Kuwait (KWT)	32	27	38
Oman (OMN)	29	11	24
Qatar (QAT)	35	24	51
Saudi Arabia (SAU)	34	29	41
United Arab Emirates (ARE)	55	30	42

11.4 Discussion

The conditions influencing education in the Gulf area are shaped by rapid societal and economic changes, and are fueled by the great wealth accumulated after World War II via export of natural resources. However, natural resources are declining; this forces the region to diversify its economies and to generate other sources for future income. In the region, a good education – especially in STEM subjects (science, technology, engineering, and mathematics) – is seen as an important key in successfully achieving the intended leap from formerly resource-based economies to globally interacting knowledge societies. Western countries have a longer history of developing mass schooling systems; a major objective for GCC countries, therefore, is the *qualification* of the student body in terms of job-related skills in order to make their economies competitive on a global market, and to offer individuals changes for the better in terms of good working conditions and higher salaries (Fend, 2006). The affluent rulers of the Gulf monarchies, on the other hand, created social welfare systems especially to support their own people through employment and other gratifications, maintaining closer bonds to their kin as a result (Colton, 2011). This included provision of education free of charge, but also included strict control of the ruling elite over institutions and learning – allowing more focus to be put on such social functions of schooling, as described by Fend (2006) as *enculturation* and *legitimization* (via

academic Islamic and Arabic studies), rather than necessarily *qualifying* the youth to face the challenges of modernization. Often, access to wealth, prestige, or power also depended (and still depends) more on an individual's connections to the ruling families than on his or her economic or cultural capital. This created societies in which "Privilege or disadvantage is determined by class, gender, ethnicity, and national origin, while religious affiliation is another significant social marker" (Moghadam & Decker, 2010, p. 75).

Along with modernization and economic growth, the demand of expatriate labor forces increased; accordingly, the educational sector also heavily depended on expatriate teachers. While non-nationals in most of the countries represent more than half of the population, they tend to not be integrated, living completely separated under different and often precarious conditions, and with only limited opportunities to participate in society (Fargues, 2011, p. 247). As such, non-nationals are still regarded as the lowest social class and furthest from the ruling parties, even if some (frequently, Western) consultants and technical specialists do accumulate a certain measure of wealth and influence. In addition, the special situation of girls and women in the traditionally patriarchic GCC countries is important for the educational context in the region. While girls now have universal access to primary and secondary education in all GCC countries, and sometimes even represent the majority in tertiary education, they still face restriction in terms of job opportunities, and they are often pushed towards fields like education, nursing and public administration (Ridge, 2014, p. 146). The special educational contexts described above also might help to better interpret some of the findings made during the current research.

Scholars offer different explanations regarding the comparatively poor overall performance of the region in international large-scale assessments. BouJaoude and Dagher (2009, p. 3), summarizing their book about science education in the Arab world, found that "Teaching suffers from an overemphasis on teacher-centered approaches and dissemination methods that encourage rote-memorization and neglect the development of critical thinking, problem-solving and inquiry skills." Similarly, the *Arab Human Development Report 2003* stated that "In Arab countries, however, lectures seem to dominate. Students can do little but memorise, recite and perfect rote learning" (UNDP, 2003, p. 69). UNESCO's Education For All regional report offered: "What seems to be the common denominator in the various Arab states is important shortcomings in quality education (learning levels, curricula relevance) and especially external efficiency (relevance of training to the labor market needs)" (UNESCO, 2012a, p. 29). A lack of relevance of the content taught to the needs of the job market is also supported by Bahgat (1999, p. 130), who described two prominent characteristics of the public education of the GCC countries:

firstly, the curricula are dominated by Islamic and Arabic studies; secondly, on all levels of education, more emphasis is put on academic learning than on vocational and technical training. While most of the GCC countries, in their modernization programs, have also shifted focus towards devoting more time to mathematics and science, analyses of the TIMSS 2015 data show that the lowest-achieving countries of Kuwait and Saudi Arabia are also the countries with the lowest amount of time for mathematics and science per week; differences for science are stronger among countries. Whereas in Kuwait, science is only taught on average for 123 min and in Saudi Arabia for 113 minutes per week, average teaching time for science in Oman reaches 213 minutes. These values can be derived from the almanac statistics of question TQS-01B (Foy, 2017, G4 science teacher almanac). While problems on the curriculum level are recognized, and are currently being addressed by the GCC countries (see section 2.3), implementation of new standards in the education system still poses many challenges. As per Aziz (2016, p. 41), in his conclusions about the development of curriculum standards in the GCC countries, "It is human nature to resist change, even when it is necessary." Similarly, the high share of expatriate, mainly male, teachers is mentioned as a factor influencing achievement outcomes. Western-style mass schooling only started in the 1970s, but then expanded very quickly, resulting in a vast amount of expatriate teachers that needed to be recruited from other Arab countries – mainly Egypt, Syria, Jordan, and Palestine (Ridge, 2014, p. 109). Students within these countries exhibit low overall achievement in mathematics and science, which consequently can be assumed to reflect teacher education and preparation in those countries. For example, an examination of the TIMSS mathematics results from 2011 reveals, for grade eight achievement, 406 score points for Jordan, 404 score points for the Palestinian Authority, and 380 points for Syria (Mullis, Martin, Foy, & Arora, 2012). The last cycle Egypt participated in was TIMSS 2007, obtaining 391 points (Martin, Mullis, & Foy, 2008). All these results are located below the average GCC mathematics achievement of 417 score points, as shown in Table 2-6. Ridge (2014, pp. 116–117) stated that "The quality of students entering teacher education programs in these countries is uncertain and is then exacerbated by equally deficient teacher education programs." Expatriate teachers, many of whom are rather subject specialists who don't obtain education degrees, often lack the practical components of education (Ridge, 2014, p. 117), bringing influences from a variety of different curricula and transferring mainly teacher-centered approaches. Additionally, expatriate teachers, as well as other non-nationals, receive low wages compared to their national counterparts; suffer from high job insecurity; and receive poor promotion opportunities, often resulting in low motivation and the temptation to make extra money by private tutoring or other activities (Ridge, 2014).

Another common characteristic in the region is the high gender gap in favor of girls for all countries in science, and for four out of six countries also for mathematics – as displayed in Table 9-1 for mathematics and Table 9-2 for science. While the gender variable was not included in the multilevel model on level 1 (to allow for model comparison with Saudi Arabian data wherein all schools are gender segregated) the multilevel results showed that a higher share of girls in courses is associated with higher achievement in all countries except Qatar. The gender gap in the region is often attributed to motivational factors. Ridge (2014), for example, claimed that boys have easier access to well-paid public-sector jobs that require little education. Additionally, as per Ridge, cultural norms require the male to financially and emotionally take care of his family in case of family break down or polygamy – which could explain early male dropouts. She concluded that a better education allows girls to partly overcome cultural norms that formerly restricted them to the household.

Gender differences are not only affected by different roles and opportunities in the Gulf societies, but also might be influenced by different opportunities to learn based on gender-specific differences related to the teachers. Teaching in the GCC countries is usually seen as a "low status" profession; as such, it is mostly avoided by national men who usually enjoy more employment opportunities. This leads to a situation in which most female teachers are nationals, and most male teachers are expatriates. Particularly in single-sex schools, students are mainly taught by teachers of the same sex. In consequence, boys, especially those in single-sex schools, are often taught by lower-paid, lower-motivated, and less-educated expatriate teachers (Ridge, 2014). These findings fit results of the current analyses, which indicated that in Kuwait and Saudi Arabia female teachers are associated with higher student outcomes; the data cannot give indication about the direction of the association, however. Causation could also be in reverse – thus, it is also possible that the gender differences are due to the fact that the majority of (better-achieving) girls are taught by female teachers. However, for mathematics, a higher achievement in Saudi Arabia is associated with male teachers, which requires a different explanation. As the "gender gap" of the teachers is far higher in boys' schools, the author hypothesizes that, due to the traditional gender roles in Saudi Arabia, female teachers might have a specifically hard time in boys' classes (for more detail, refer to the section about factors associated with achievement on course level earlier in this chapter). A somewhat different, or possibly complementary, hypothesis for the large gap in favor of girls is discussed in the master thesis of Anderson (2012). From her research on TIMSS 2007 science data for the United Arab Emirates, she concluded that girls see their female teachers as role models and, in the absence of other professional opportunities, want to become teachers – as for them, the profession might have greater prestige

than for boys; she concluded that "This could explain why girls persist in school longer, have greater achievement levels in science, and transition to tertiary education at greater rates than boys in the ARE" (Anderson, 2012, pp. 69–70). While both explanations discussed above might contribute to the large gender differences in the region, effects are specifically seen in the national populations. Consequently, it can be expected that the gender gap for non-national children, for both subjects, is generally considerably lower. This assumption is supported by the preliminary analyses presented in Table 9-5 for mathematics and Table 9-6 for science.

Another common characteristic in the region is the higher achievement of non-nationals in all countries except Oman. In Oman, the immigrant workforce is described as having the lowest educational level of all GCC countries (Baldwin-Edwards, 2011, p. 50), which might explain lower differences to the national population. While in Oman both genders of the non-national population outperform Omani boys, Omani girls have the by far highest achievement of all groups. In general, *nationality status* in the region can be assumed to affect educational opportunities, as foreigners do not much benefit from the welfare systems of the GCC countries. Literature in regard to the study of why non-nationals outperform national populations is rather scarce. A possible hypothesis, based on the literature review related to the conditions of non-nationals in the GCC countries postulated here, is that obtaining a good education is vital for non-nationals, but less important for national citizens. While nationals benefit from the national welfare system as well as from nationalization policies, non-nationals have to leave the country if they become unemployed. Hence, their only chance to stay in a GCC country is to find an employer. Findings of Wiseman, Alromi, and Alshumrani (2013), who analyzed TIMSS 2007 data, fit this hypothesis: they reported a stronger connection between doing well in science and getting a desired job in the labor market in Saudi Arabia for non-national students than for nationals or students with only one Saudi parent. Similarly, the analyses displayed in APPENDIX C (Table C-1 for mathematics and Table C-2 for science for science) empirically support such a hypothesis to a certain extent. The analyses show that in all countries except Bahrain and Oman, achievement of the immigrant population is higher for all levels of parents' highest level of education, albeit standard errors due to a low number of students in some of the cells are quite high, and therefore differences are not statistically significant in all the groups.

While the above-mentioned section describes common underlying patterns in the region, results also indicate larger differences in the countries; multilevel analyses could not discern a common regional pattern of predictors of educational effectiveness in the region. Likewise, variance components between classes vary across the region, and both the amount of explained variance by background factors and the variance explained by the full model vary quite substantially

among the GCC countries. Reasons for these differences can only be hypothesized – as many different and overlapping factors may be at play.

GCC countries such as the United Arab Emirates and Qatar, who have the highest share of non-nationals, also show more between-group variances compared to countries with a lower share of non-nationals. It is reasonable to assume that a society with such a high immigrant population as the United Arab Emirates will be more heterogeneous, due to the influence of different cultures. Non-national students are often not admitted to the public school system in the Gulf countries, resulting in increased development of the private school sector (see Table 2-3 for enrollment to the private sector in primary education), which follows different curricula and, according to Ridge (2014, p. 30) lacks governmental regulations – especially in the United Arab Emirates. Again, these factors may add to increased between-school differences. On the other hand countries with a lower share of non-nationals, paired with a higher degree of conservatism such as in Saudi Arabia, feature highly centralized curricula and high degrees of control by the ministries of education which is likely to reduce the variation between classes and schools.

Examination of the home background models including composition variable reveals that in most countries, the models explain at least half of the overall group-level variance explained by the full model. In terms of absolute values, the home-background model explains the least amount of variance in Oman and in Saudi Arabia, with only 11 and 12% (only for mathematics), respectively, indicating that the home background conditions in both countries are more similar than in the remaining GCC countries – which also could partly be due to a lower share of immigrants.

Especially in Saudi Arabia, the correlation of variables usually used for creating SES indices with mathematics and science achievement is also relatively low (see Table 8-4 and Table 8-5). On the other hand, in all countries, the correlation between *nationality status* and mathematics achievement is quite high, with coefficients nearing (and in Saudi Arabia even surpassing) the magnitude of the *ESCS* index (see Table 9-28). An analysis of the variance in reading achievement explained by various aspects of family background based on PISA 2009 data (OECD, 2010b, Figure II.2.4) shows similar results. For both GCC participants (Dubai [United Arab Emirates] and Qatar), the immigrant status, as well as the language spoken at home, emerged as the two most important single aspects. Gender, however, was not included in the PISA analysis. Relating these findings to the description of the determinants for social stratification in the Gulf area described above, it is hypothesized here that factors influencing the student's background, especially in more traditional societal strata, are partly different compared to those in

the West. For immigrant workers, especially when coming from the West, it would be reasonable to assume that differences in economic and cultural capital determine social position within the immigrant population; and thus, that similar processes for cultural reproduction, as described for Western countries, are at work. Conditions for resident populations will likely be different. With possible exceptions for portions of the newly emerging middle class composed of professionals, access to power, prestige, and wealth cannot simply be reached by a good education and a well-paying job; rather, access is dependent on connections to the ruling family. These connections, in turn, assure well-paid employment (predominantly in the public sector), benefits from social gratifications (such as free access to all levels of the education system), etc. In consequence, and within the boundaries of origin and gender, the accumulation of social capital by creating a strong and extended network might be very important. It is therefore assumed that inclusion of additional factors of the family context, like those used by Smits and Huisman (2012) for Arab countries outside the GCC area (see also section 3.3.4.3), could further improve the models – especially in more conservative societies of the region.

In general, the degree of conservatism in regard to religion, tradition, and family orientation, which is especially pronounced in Saudi Arabia with its strict Wahhabi religion, is expected to influence the educational context in many ways. The *opportunity to learn,* for example, might be affected by excluding certain content-related topics that don't fit with the beliefs of the predominant religion, especially in science (for example, the theory of evolution). Correspondingly, the TIMSS test curriculum matching analyses (TCMA) indicated only 40% coverage of the science items for Saudi Arabia and the United Arab Emirates in their national curricula, while 100% coverage was achieved for mathematics (see Table 2-4). Likewise, the *time on task* for mathematics and science may be limited by the amount of time devoted to other subjects, such as religious studies and the like.

Another factor that might lead to more variability in the region is the introduction of educational reforms, which seem to work differently across the six countries. It can be assumed, for example, that implementation of reforms would be easier for smaller countries, which only have a limited number of schools, and higher pressure for reforms would be applied in countries with limited natural resources, or in which resources are expected to run out in near future. In Bahrain, for example, organizational and educational reforms, such as the decentralization of the school system, have started at an earlier stage than elsewhere: more comprehensive programs, such as an inclusive program to provide equal opportunities for girls, expatriates, and disabled children have been implemented (Ridge, 2014). Recently implemented reforms in the region are briefly described in section 2.1. In larger countries with high oil revenues, such as Kuwait

and Saudi Arabia (the countries with the lowest achievement in the region), substantial reforms lag behind. As Sakr (2008) concluded for these countries, "it seems that long educational histories have made the bureaucratic legacy an impediment to far-reaching initiatives, while ideological disputes prevent the emergence of new ideas." Indeed, it can be seen from when comparing TIMSS 2011 and 2015 results that in the smaller countries especially, but also in Oman (whose oil reserves soon are expected to be depleted; Baldwin-Edwards, 2011, p. 52), more progress was made in terms of mathematics and science outcomes.

While the above-mentioned explanations might all contribute to the large differences detected in the achievement results and the fairly different associations of effectiveness-enhancing factors and outcomes across countries, the current study does not allow for the disentanglement of the different factors, or for judgments regarding their relative importance.

Reflection on the Framework used

The analysis results of the TIMSS 2015 data show significant relations between predictor variables and achievement for the major dimensions *Input*, *Quality*, *Time*, and *Opportunity* of the framework used. Moreover, results showed that indicators from all educational levels were associated with student outcomes, which supports the notion that educational influences in the Gulf area are also multilevel. These findings of the current analyses can be regarded as providing further empirical support for the validity of Creemers' model, which was used as the starting point for the research framework of the current study.

Additionally, the results empirically supported the justification of adding an explicit *input* dimension beyond the process variable dimension. Similar to research performed for other regions of the world, the current analyses showed that student composition effects explained additional variance in all Gulf education systems; correspondingly, variables related to the students' background were significantly related to outcomes throughout the region. Moreover, the multilevel analyses revealed that in all countries, except Qatar, teacher background characteristics showed a significant association with achievement – even when entered together with all process characteristics variables. The effectiveness model of Creemers (1994), as well as the dynamic model (Creemers & Kyriakides, 2008), restricted the models "to the classroom factors that may have a direct impact on student learning through the actions of the teacher that can be observed in classrooms" (Creemers & Kyriakides, 2008, p. 217). In this context, the author assumed that data from self-reported background questionnaires, as available from the large-scale assessment data at hand, cannot fully cover teacher behavior in the classroom – an assumption that is supported by the findings related to significant associations of teacher characteristics

with achievement, once entered into the model jointly with the teacher behavior variables. Additionally, certain teacher background characteristics, such as the pedagogical content knowledge, are malleable to a certain extent and as such convey important information for policymakers.

The current study, however, could only give limited empirical evidence for the differentiation of the factor *quality of instruction* into five sub-dimensions which were created after the review of more recent literature on effective instruction. The most consistent finding was the importance of a *clear and structured teaching* approach, as reported by the students. The newly introduced items related to *cognitive activation* were only significant in Qatar in a model without controlling for the home background. Unfortunately, TIMSS 2015 did not include items directly addressing the classroom management. As a proxy, *problems reported with disruptive students* were used, but such reports may occur not only because of the lack of teachers' classroom management abilities, but also depend on the student composition (which here at least partly was controlled for), as well as policies and support on higher levels. Information related to school policies could not be retrieved from the TIMSS questionnaires in most cases. Based on the importance of the construct as discussed in chapter 3.3.5.2 the author supports assertions by Blömeke et al. (2016) that classroom management is a vital dimension of instructional quality, and additional questions therefore should be included in further cycles of TIMSS. Moreover, measurement of the *assessment* dimension could not be fully covered by teachers' activities in terms of homework verification. It is therefore likely that the TIMSS questionnaires, in some areas, lack more appropriate indicators which would allow for a more clear distinction between the different sub-dimensions. As discussed in section 8.2.1, this is clearly a disadvantage to be taken into account when performing secondary analyses of data that was collected for a (partly) different purpose.

Notably, although the main model variables in the region were found to be related to student achievement, variables beyond the student background characteristics could only explain a limited amount of unexplained variance – albeit, the range is in line with results from earlier effectiveness studies. Furthermore, except on student level, no regional pattern could be discerned, and countries like Kuwait and Saudi Arabia only show a few significant relations to achievement outside the newly added *input* dimension. In these countries, the deviance statistics do not indicate a better fit of the more complex model over the mere background variable model. This shows that empirical support for the model used is limited, and the model seems to work best in the United Arab Emirates, the country with the highest share of immigrant students and a high share of between-group variance.

Suppression effects

An additional finding related to the multilevel models is that controlling for the student background sometimes strengthened the associations between predictors and achievement, which led, in turn, to several significant relations which did not occur in the uncontrolled model. When comparing the uncontrolled course- and school-level model in section 10.4 with the full model in section 10.5, several new significant relations are shown to have occurred. This affected five predictors in Bahrain and three in Saudi Arabia; one variable for Qatar in mathematics; and for science, one variable each in Bahrain, Oman, and Qatar, and two variables each in Saudi Arabia and in the United Arab Emirates. Similarly, when comparing the student models including composition (section 10.3) with the full model, for Kuwait a positive composition effect in terms of *nationality status* occurs for mathematics.

Such results are assumed to be explained by the so-called *suppression effect*. Suppression effects are quite common in social research, and occur in multiple regression equations when a suppressor variable increases the predictive validity of one or more other variable(s) (Conger, 1974). A variable can improve the association of other variables with the outcome variable, and thus can act as a suppressor even when there is no direct relationship between suppressor and outcome. Thus, suppressors are predictor variables that, while not directly correlated to outcomes, are strongly correlated with other predictor variables – which in turn are associated with the outcome variable. Pandey and Elliott (2010, p. 35) argued that keeping suppressor variables would give more accurate regression coefficients of the independent variables, improve the overall predictive power of the model, and enhance accuracy of theory building. They concluded that "…the risks associated with excluding a relevant variable are much greater than the risks associated with including an irrelevant variable." When comparing the tables listed for uncontrolled and controlled models in the comparative effectiveness analyses by Martin and Mullis (2013), similar effects can be detected for two out of four GCC countries participating in the study: in Oman, the predictor *schools are safe and orderly* is only associated with reading achievement after controlling for students' background (Exhibit 3.24). In Qatar, the same predictor is significantly associated with mathematics in the school environment model, not significant in school environment and instruction model, and again significant in the controlled model.

In the current analyses, few predictors were removed during the preliminary analyses in order to keep the complex models reasonably parsimonious. The ultimately evaluated and interpreted models were always the full models, which included all model predictors jointly.

11.5 Contribution to Scientific and Practical Knowledge

International comparative assessments such as TIMSS, PIRLS, or PISA play an important role in the monitoring of student achievement trends, and in assisting policymakers in making informed decisions to improve their educational systems. Thus, all GCC countries have participated in the TIMSS assessment in grade four and grade eight, and are using the results to plan several intervention strategies in their countries.

While findings from such large-scale assessments are only of limited help to teachers regarding their daily teaching practices, they do allow for some in-depth insights related to areas of concern, in turn allowing certain weaknesses to be addressed and possibly facilitating a reduction of inequality among certain subgroups of the population. Comparisons between countries that share many common characteristics in terms of their education system, history, and values – as is the case for the GCC countries – are especially valuable, as they facilitate learning from successful teaching and learning practices of other countries facing similar educational conditions in the region. Policymakers and researchers in the region can use the data and resulting analyses to better understand the current state of their education systems, in order to delineate suitable interventions. In this regard, analyses of the variance components can help to detect differences between subgroups of the populations, as a starting point in implementing interventions intended to reduce inequality.

The current dissertation detailed a comprehensive conceptual framework which was originally based on the well-established work by Creemers (1994), but also integrated more recent model developments and research findings – for example from Creemers and Kyriakides (2008), Klieme and Baumert (2001), and Nilsen et al. (2016). The framework recognizes the multilevel influences of educational factors on learning, and emphasizes the factors *quality*, *time*, and *opportunity* from a perspective of teaching and learning theory. Recognition of the fact that major processes in learning and teaching take place in the classrooms is increasing, resulting in a stronger focus of researchers and policy-makers in the quality of instruction and the educational climate. The current framework tried to address these needs by elaboration of this area of study. For example, there is increased focus on more modern constructivist approaches of learning and teaching (here subsumed under the term of *cognitive activation*), a factor which is also gaining importance in the Gulf region (Khan, 2015). GCC countries have acknowledged that often, the still-predominant teaching forms in the region, characterized by memorizing, reciting, and rote learning (UNDP, 2003, p. 69), might need to be adjusted in order to compete in a globalized world. Moreover, while considering the special conditions in the Gulf Area, the

framework and analysis approaches tried to take the limitations of using large-scale cross-sectional data for EER into account as far as possible. As such, the framework and methods tried to summarize important findings from EER in a manner both generic and suited for application to large-scale assessment data in other regions of the world.

By applying a framework based on constructs and theories of EER that were mainly developed in the Western Hemisphere and empirically validated in Western and Asian countries to the states of the GCC, this research project also contributes to the international dimension of EER. The project helped to shed a bit more light regarding whether the major factors of Creemers's integrated effectiveness framework (*quality, time, opportunity*, which here were supplemented by an *input* component) work similarly in a different region. Thus, it contributes to the evidence of the generalizability of theoretical constructs and models predominantly developed in the Western world. Research in this area might help to identify similarities and differences across cultures, and in consequence, help policymakers to focus on malleable educational effectiveness factors – while at the same time preventing simple translations of findings from one culture to another, without taking cultural contexts into account.

11.6 Recommendations

11.6.1 Policy recommendations for the region

Improving the school learning environment

The study revealed that variables related to the school and classroom climate, such as *emphasis on academic success*, *orderly environment*, or *disruptive students*, are an issue in several countries of the region. Therefore, establishment of an orderly school environment, with an atmosphere conducive for learning, is an important precondition for further learning gains of many students in the region. Policymakers in the region must therefore further emphasize the importance of a culture which values the benefits of education as a condition for the further transformation of the GCC monarchies towards knowledge-based societies. Moreover, school policies related to a constant improvement of the quality of instruction and the school learning environment should be established, and the implementation carefully monitored.

Improving the effective teaching time

While several GCC countries implemented policies to extend the *teaching time* for mathematics and science, absenteeism of teachers and students seemed to be an issue in many countries of

the region, while concurrently being an important predictor of achievement in several countries. Even if the total instructional time for mathematics and science nowadays might reach an international level in many GCC countries, the extent to which the available time is really used for effective instruction is still questionable. During a regional seminar on TIMSS outcomes attended by the author, ministry participants repeatedly complained that "schools are overwhelmed with different kinds of projects" (Khan, 2015), indicating that too many extra-curricular activities can result in insufficient time left for regular learning. Thus, effective teaching time in the region might be increased by reducing the administrative burden for teachers and encouraging them to carefully balance the time used for extracurricular activities with the time used for the instruction of core subjects. Establishment of clear policies and consequences in cases of absenteeism could also be evaluated.

Improving quality of teaching

The current research confirmed that *clear and structured teaching* has a significant association with achievement, at least for mathematics, in nearly all countries of the region. A teaching style in which the teacher actively transfers the content in smaller units, with clearly specified goals and ample time for practice, and monitors progress through good questioning strategies, has been found in EER to be one of the main pillars for effective instruction (Brophy & Good, 1986; Creemers & Kyriakides, 2008; Doyle, 1985). In primary education especially, schools should ensure at the beginning that a basic "corpus of knowledge" be made available, as a precondition, before other types of knowledge and skills can be developed (Creemers, 1994; Creemers & Kyriakides, 2008). Consequently, a strong focus should be set on proper teacher training programs, supervision, and monitoring, to further improve the quality of instruction in the region, also in terms of "basic" mathematics and science skills. The importance of this concept is underscored by the fact that GCC countries still depend, to a large extent, on expatriate teachers who might not bring all necessary qualifications and cultural awareness needed for the teaching profession in a culturally different environment. Further, making the teaching profession more attractive, especially to national men, might help to overcome the overly strong dependence on expatriates teachers, and thus from the associated subsequent problems discussed earlier.

Learning from good practices and successful schools

Analyses showed large diversity in terms of the level of educational outcomes, with more than a standard deviation of difference between country averages between the lowest- and highest-performing countries in the region. Likewise, there are still large disparities between different

subgroups of the population. Regional literature comparing schools in terms of the best practices of more successful schools after controlling for intake, at least in English, seems to be rather scarce, however. Research investigating the reasons behind the frequently-high differences between national students and non-nationals could not be found by the author. More comparative educational research within the region, to further elaborate on those factors that work well in the given cultural context to reduce the large disparities in the region, are therefore suggested. While learning from successful countries might be helpful, a simple transfer of curricula and teaching practices from successful countries in other parts of the world might be difficult to implement, due to highly different educational contexts in the different regions.

11.6.2 Recommendations regarding the TIMSS assessment in regard to educational effectiveness research

TIMSS is an international comparative multi-purpose assessment in mathematics and science. The TIMSS assessment is ideally suited for international comparative EER, as it is based on an international curriculum framework developed in collaboration with participating countries, and as it allows for the investigation of teaching and learning in the classrooms through its cohort-based approach.

Recommendations related to the TIMSS questionnaires

The *time for instruction* and *time on task,* respectively, are important components to be considered in terms of EER. However, research shows that measuring a dimension of time could be improved if information about the *effective teaching* time is collected – instead of measures about the overall time students spent in school to learn a certain subject. On school level, the number of instructional days is often reduced by extracurricular activities, such as sport events, partial closure due to examination activities, etc. Additionally, in some of the countries, schools might need to be closed for a longer amount of time due to natural disasters, problems with the heating and cooling systems, or strikes. The author therefore suggests to ask more specifically for the time the selected school provided instruction, rather than for a general measure of time – which, often, might also be prescribed by the ministries and thus wouldn't differ among schools. On class level, more specific questions regarding the amount of time teachers really spent teaching could be asked. Analysis of the TALIS survey has shown that in most countries, one in four teachers lose more than 30% of their time for other activities (OECD, 2009, p. 88). Similarly, an analyses conducted by Sandoval-Hernández et al. (2013) using PIRLS 2006 data showed an overall higher association with effective teaching time, compared to the overall time

available to the students. Notably, however, a related discussion of the author with the members of the TIMSS 2019 Questionnaire Item Review Committee (QUIRC) meeting conducted in July 2018 in Oslo revealed that asking teachers about their time use is not that straight forward, due to the sensitivity of the issue and many teachers' incapability of summing up percentages correctly.

More recent EER frameworks, such as the dynamic model (Creemers & Kyriakides, 2008), recognize the importance of school policies related to the *quality of instruction*, the *school learning environments*, and the *use of time* as important preconditions for an effective teaching and learning in the classrooms. Likewise, it is acknowledged that such policies only have an effect if they are evaluated according to the specific weaknesses occurring, and if corrective measures are taken afterwards. Related questions are currently not included in the TIMSS school questionnaire (albeit these topics are partly reflected in the curriculum questionnaire), and therefore some elaboration in this area could be seen as beneficial in improving the coverage of modern effectiveness frameworks, and thus in obtaining more valid information regarding educational effectiveness-enhancing factors from the school level.

Concerning the factor *quality of instruction* on class level, while good coverage was found otherwise, no items directly related to *classroom management* were found. *Classroom management* is seen as an important dimension of instructional quality, and as such, was also included as one of the main factors for instruction in the dynamic model (Creemers & Kyriakides, 2008).

As the questionnaires cannot be endlessly extended, due to administration costs and the risk of higher non-response rates, the author suggests to examine the extent to which redundancy of constructs in other areas could be reduced. For example, the questionnaires contain a fairly large number of items related to the *input* dimension of the described framework, especially related to the topics of school resources and teacher characteristics.

In the specific context of effectiveness research in the region under consideration, it would be useful to include context factors, such as the level of modernization or patriarchy, and variables related to the social capital of a family, to more comprehensively describe non-educational influences. Furthermore, variables helping to better describe organizational differences in the school system, such as information related to different types of schools (e.g., public/private/religious), would be helpful.

Complementation with qualitative data collection methods

To gain further insight into the mechanisms at play, factors should be examined beyond simple quantitative measurement. Effectiveness research shows that qualitative investigations can help in finding out when, and under what conditions, a certain factor can improve learning. Consequently, Creemers and Kyriakides' dynamic model (2008) not only prescribes measurement of each factor's quantitative considerations, but also collection of information about its qualitative nature. Availability of additional qualitative data would allow for a better interpretation of relationships found, and make findings easier to understand and use for practitioners and policy-makers. While studies such as the TIMSS video study (Stigler et al., 2000) are quite costly and time intensive, optional components of classroom observation, or interview based-methods, could be used to complement quantitative dimensions.

11.6.3 Recommendations concerning further research on educational effectiveness in the region

Two recommendations for researchers working on EER-related questions are listed below.

Documenting the centering approach in multilevel analyses and the procedures to calculate the explained variance

While researchers often use the variance explained in multilevel models as a kind of effect size, the results are often not comparable among different studies – as many studies lack detailed information on the centering approach for the predictors used in their analyses. As was shown by repeating the different multilevel models (originally calculated with level 1 predictors centered on the grand-mean) using a group-mean centered approach, the amount of explained variance differs largely between the two approaches. The amount of explained variance in the grand-mean centered approach is considerably lower; for example, in the home-background model, it often only obtained half of the share of explained variance when compared to a group-centered approach. In addition, different methods can be used to calculate the variance explained from the different models. In order to allow for valid comparisons between different studies, both the centering approach applied, and the procedures to calculate the explained variances, should be clearly documented in all EER studies.

Inclusion of gender and nationality status in measures of student background for the Gulf region

The current project could show that gender and nationality status play an important role as determinants of an individual's role in the Gulf societies. It is therefore recommended to place a special focus on these variables for creating student background indicators, and for research questions related to the SES in the region. Moreover, additional community context factors and variables related to the social capital of a family might help to more comprehensively describe non-educational influences in the region.

11.7 Limitations

The use of international large-scale assessment data for analyses in the field of EER introduces certain limitations, despite all associated benefits for comparative research. Firstly, the approach described here is only of a quantitative nature, and the data is self-reported; therefore, it may be biased, for example due to effects of social desirability. In general, quantitative analyses on educational effectiveness would benefit by further complementary investigations of the causal mechanisms behind educational effectiveness through qualitative research, such as by use of classroom observations and in-depth interviews, as suggested above.

Additionally, IEA large-scale assessment data is of a cross-sectional nature, and thus obtains measures only at a certain point in time. The study does not include a real measure of aptitude, such as an intelligence measure, or a measure about students' prior knowledge. This limits the magnitude of educational influences which can be investigated, and results will consequently depend on the extent to which the different factors can be disentangled through "controlling out" non-educational background influences. As a rough proxy for aptitude, the current study used students' early numeracy skills when entering primary education, as judged by their parents. Moreover, the data in general does not allow for making real causal inferences about factors associated with effectiveness, as only associations between variables can be measured. To investigate causal relationships, longitudinal studies are needed. Notably, moreover, international large-scale studies are multi-purpose studies, and choice regarding variables that can be used to examine questions in the field of EER is restricted. This means that certain important aspects that might be relevant to the constructed framework might not have been asked of principals, teachers, students, and their parents, and hence are not available in the current data. While the questionnaires offered indicators for nearly all of the framework's sub-factors, the

quality of indicators is different, and some suggestions for improvement in this regard have been made in the previous section.

Additionally, while great efforts were made to provide equivalent translations between different languages (while for most students of the region the study material was administered in Arabic, in certain private schools an English adaptation was used), and the selected region is influenced by common historical events, culture, and beliefs, effects of cultural invariance for certain constructs cannot be completely excluded. Respondents with different cultural backgrounds, which in this context necessarily affects both national populations and non-national populations (whereas the non-national populations can be separated into groups with different cultural backgrounds), might exhibit different response behaviors. Consequently, results should be interpreted with a measure of caution.

11.8 Further Research

The current study focused on two cognitive outcome measures: namely, mathematics and science achievement. However, to obtain more of an in-depth understanding of the extent to which effectiveness indicators work in general for the different outcome criteria, further studies of educational effectiveness could benefit by inclusion of other important outcome measures of schooling, and by expansion of the analyses to include non-cognitive outcome measures, such as attitudinal or behavioral characteristics of the students. To gain deeper insight into the relations between different effectiveness indicators, same-level and cross-level interaction should be investigated. More detailed investigation into the interaction effects between educational effectiveness indicators and gender and *nationality status*, respectively, is of special interest – as both variables show large disparities in terms of achievement, and because, due to the missing integration of non-nationals into the Gulf societies, it can be assumed that effectiveness factors might work differently. Of special importance is the quasi-longitudinal design of the TIMSS assessment, in which fourth graders are assessed four years later in grade eight, as well as the fact that all GCC countries participated in both grades; investigation of the changes related to the behavior of effectiveness factors from one cohort to the next, therefore, could add to the validity in terms of the consistency of effects over time.

To gain a deeper understanding of the functioning of the mechanisms at work, qualitative studies which investigate, in more detail, what exactly happens in the classrooms – for example, via video studies or administration of classroom observation records – should also be conducted and connected with quantitative analyses through mixed-methods approaches.

11.9 Conclusion

The current research project attempted to investigate educational effectiveness factors in the GCC countries, based on a framework rooted in Western paradigms of EER. The analyses provided evidence supporting the assertion that educational effectiveness factors based on quality, time, and opportunity, as well as on the added input factor, also operate at different educational levels in the GCC countries. For most of the GCC countries, the analyses showed significant associations of instructional factors, factors related to the school climate, and the time of instruction with achievement, even when controlling for the home background. However, it also became evident that in more traditionally-oriented Gulf societies with less influx of non-nationals, between-group variances were lower, and school environments and instructional factors, with the exception of clear and structured teaching, also seemed to be less pronounced. The most important factor was the student background, especially when gender and *nationality status* were included in the background model.

In the Gulf region, living conditions vary considerably across different subgroups of the population. While non-nationals generally are found in the lower classes in the Gulf societies, and often live under precarious job conditions, they are vital for the economic functioning of the Gulf societies, particularly in the context of their ambition to transform into "knowledge societies." Moreover, non-nationals outnumber the national population in several countries. As non-national students don't benefit as much from welfare state policies, their only chance of obtaining a good job is through a good education. Similarly, a good education allows girls to at least partly overcome the restrictions surrounding a traditional life spent as wife and mother. For a certain share of the younger men from traditional families, however, good education – especially in demanding subjects such as mathematics or science – are not especially valued, and Arabic and Islamic studies still seem to be more highly-valued. In this context, the objective of education is often different; factors related to the origin and gender of a person might be more important in determining an individual's status, and might more strongly affect aspirations in terms of education. The results provide evidence that research based on Western effectiveness paradigms do have some justification in the Gulf area, especially as Western types of schooling is seen in the region as a model with which to compete on a global market. On the other hand, traditional schooling in the area also partly had different functions to fulfill in comparison to the West; moreover, social stratification and reproduction works differently, at least to some extent. Thus, when performing EER in the region, the different societal conditions in the patriarchic Gulf monarchies related to the position of women in society, the importance and role of

the non-national populations, as well as consequences relating to the motivation of the different groups of the population in achieving a good education, should be taken into account.

12 SUMMARIES

12.1 English Summary

One of the most important functions of modern education systems is the qualification of the student body. Student body qualification results in benefits not only for individuals – for example, by providing a better chance for good work conditions, higher salaries, and enhanced participation in the society – but also for society at large, by making the economy more competitive on the global market. As such, improving the quality of education is also an important topic on the agenda of the GCC, an intergovernmental union which politically and economically unites Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates. The GCC represents a region currently experiencing extensive and rapid economic and social changes, while transforming from traditionally-oriented oil monarchies into knowledge societies open for global competition on the international market. The GCC provided the focus for this dissertation as its member countries exhibit many commonalities – such as similar social and cultural values, religious beliefs, and historical events; moreover, all share a common language, and their educational contexts largely differ from the West.

All the GCC countries are classified to be among the wealthiest countries in the world, and – mainly due to the export of natural resources – the region managed to vastly improve quantitative dimensions of schooling within the last few decades. However, qualitative dimensions of education are still lagging behind when being compared on international level. GCC countries are still among the lowest-ranking countries in international large-scale assessments such as TIMSS (Mullis, Martin, Foy, & Arora, 2012), PIRLS (Mullis, Martin, Foy, & Drucker, 2012), and PISA (OECD, 2016a). Additionally, the region still shows large disparities in terms of gender in favor of girls, reaching up to nearly 80 score points for science in Saudi Arabia in grade four. In most of the GCC countries, large achievement gaps also appear between the national populations and the mostly higher-achieving non-national populations, reaching a difference of 110 score points in the United Arab Emirates for science.

The purpose of this study was to explore achievement differences of primary school students in the GCC countries concerning mathematics and science, from the perspective of an educational effectiveness framework. Achievement differences were investigated by means of secondary analyses of data from the IEA's Trends in International Mathematics and Science Study (TIMSS) 2015. TIMSS is a large-scale international comparative assessment which is conducted every four years, focusing on mathematics and science achievement in grades four and

eight. Additionally to the test instruments, TIMSS students, their teachers, and their school principals are requested to answer questionnaires about their educational contexts for learning mathematics and science. In grade four, the parents of sampled students are also administered a background questionnaire to provide complementary data on students' home environments. All six GCC countries participated in both grades of TIMSS 2015.

While highly standardized international large-scale assessments such as TIMSS provide a good opportunity to dig deeper into the international dimensions of educational effectiveness research (EER), the fact that TIMSS is a multi-purpose study of a cross-sectional nature, and as such not specifically designed for EER, had to be taken into account. Concerns relating to the use of large-scale assessments for EER were addressed, to the greatest extent possible, as detailed in this thesis.

Two main research questions were formulated in order to attain the research objectives. The first question is: *To what extent does TIMSS 2015 reflect essential factors in terms of educational effectiveness research?*

To address the research questions, as a first step after a comprehensive literature review of EER, a suitable research framework was built. The research framework was rooted in Creemers' (1994) comprehensive model of educational effectiveness, but was complemented by elements from the dynamic model (Creemers & Kyriakides, 2008), and also partly by Scheerens' model (Scheerens, 1992). Additionally, the research framework included considerations from more recent research in the area of instructional effectiveness (Helmke, 2009; Klieme & Baumert, 2001; Nilsen et al., 2016; Seidel & Steen, 2005). The new model differs from Creemers' approach in three main aspects: firstly, by addition of an *input* dimension; secondly, by reclassification of sub-components related to the *quality* category on classroom level (and, to a lesser extent, also on school level); and thirdly, by inclusion of more recent research findings related to the elements of *instructional quality* and *school climate*. The framework was targeted to the cross-sectional data at hand, and focused on the more generic aspects of educational effectiveness in the region under consideration (i.e., the GCC region). In a second step, about 170 variables from the TIMSS questionnaires were categorized according to the model factors of the theoretical framework, and principal component, reliability, and correlation analyses with mathematics and science outcomes were used to elaborate on the underlying constructs, and to specify a regional model of important factors parsimoniously. These procedures resulted in 5 variables on school level, 19 variables on course level, and 7 variables on student level that were retained for further analysis steps.

While the strength of the correlations between model variables and outcomes varied by country and subject, overall, results indicated that the TIMSS 2015 questionnaire variables exhibit a satisfactory coverage of the constructed framework, in terms of indicators which were empirically demonstrated to be related to student outcomes, according to the effectiveness literature consulted. From a theoretical perspective for all factors, except for the *quality of instruction* factor on school level, matching TIMSS 2015 variables or proxies could be found. However, variables were not evenly distributed – a stronger focus was placed on *input* related variables, as well as on *quality of instruction* on teacher level. The factor *opportunity* on school level had to be dropped from the final model, as the only matching variable related to *tracking policies* did not show meaningful correlations in any of the countries.

The second research question is: *According to the framework specified, which educational factors are most effective from the perspective of EER with regard to learning outcomes on primary level in the GCC countries?*

To answer this question, this study used multilevel modeling techniques to deconstruct the total achievement variance into within- and between-course/school-level parts. The main objective was to quantify the relationship of school-, teacher- and course-level factors, identified according to the proposed model, with student achievement, while controlling for non-educational home background influences. Altogether five different sets of models were analyzed for each of the six GCC members: null models, level-1 student-level background models, background models including student composition variables on level 2, school- and course-level variables without controlling, and the final models with all framework variables entered jointly.

Student background factors emerged as the most consistent predictors of achievement in all six countries. While in general all six factors operate similarly across the region, the strength of the association differed somewhat; for example, no significant differences in Oman could be found related to the *nationality status*, while for mathematics, no significant *gender* differences were found in Qatar, nor in the United Arab Emirates. On course level, *clear and structured instruction* and the *amount of teaching time* emerged as the most consistent factors across the region after controlling for home background influences. However, while predictors of all main model factors (*input, quality, time, and opportunity*) were significantly related with mathematics and science achievement in one or more countries, a regional pattern in terms of common regional factors could not be discerned. With regard to the number of significant relations to achievement, the analyses showed that in Qatar only a single model variable (namely, the *gender of teacher*) demonstrated a relation to student outcomes beyond variables related to the home

background, while a total of six model predictors (out of 12) were found to be significantly related to achievement in the United Arab Emirates for at least one of the subjects. On school level, all three of the retained model factors showed significant relations with achievement in the region after controlling for the home background, but again, no regional pattern across countries could be discerned.

An additional finding related to the multilevel models was that controlling for the student background sometimes strengthened the associations between predictors and achievement, which led, in turn, to several significant relations which did not occur in the uncontrolled model. It was assumed that such results could be explained by the so-called *suppression effect*.

The results of the variance decomposition analyses showed that the United Arab Emirates exhibited the largest variance between courses (59% for mathematics and 55% for science), while the least between-course variance occurred in Bahrain (24% for mathematics, and 29% for science). The amount of level-2 variance that could be explained by student composition predictors ranged from 12% in Oman and Saudi Arabia to 28% in Qatar for mathematics, and from 11% in Oman to 30% in the United Arab Emirates for science.

The final models with all factors entered jointly explained between 27% of the level-2 variance in Oman and 46% in Qatar for mathematics, and between 24% in Oman and 51% in Qatar for science. In most countries, approximately half of the explained level-2 variance was due to the composition effects of student background variables entered on level 2; the additional amount of variance explained by course- and school-level factors ranged from 7% in Kuwait for mathematics to 27% for science in Qatar.

While the results show certain similarities in the educational contexts of the region, which result in relatively low overall levels of achievement and generally high disparities in terms of gender and *nationality status*, the educational conditions across countries in certain aspects also differed to a large extent.

Explanations regarding the comparatively poor overall performance of the GCC countries in international large-scale assessments often point to shortcomings in the quality of education, and a lack of relevancy to labor market needs (for example Bahgat, 1999; BouJaoude & Dagher, 2009; UNDP, 2003; UNESCO, 2012a). The high gender disparity in favor of girls is partly attributed to lower motivation of boys and their (often expatriate male) teachers (Ridge, 2014). A review of the available literature suggested that the higher achievement exhibited by the non-national populations may partly also be explained by motivational factors, as a good education

seems to be less important for nationals who benefit from the national welfare systems and nationalization policies; non-nationals, on the other hand, are obliged to leave the country once they become unemployed.

Various explanations were discussed as potential contributors to the large variation across countries in terms of the number and strength of significant effectiveness-enhancing factors, but also in terms of the different shares of explained model variance in the region. One important difference across countries might be related to the different shares of non-national populations, and, as a partly-related factor, to the independent development of the private sector – which might explain the higher between-group variances in countries like the United Arab Emirates and Qatar. In addition, degrees of conservatism with regard to religion, tradition, and family orientation can be assumed to result not only in different educational opportunities for boys and girls, but also to influence the wider educational context in myriad ways. Another factor that might lead to more variability in the region is the introduction of educational reforms, which seem to operate quite differently across the six countries. While all of these factors can be assumed to contribute to the large variations across GCC countries, the current study does not allow for a disentanglement of the different factors, nor for judgments regarding their relative importance.

Based on the outcomes of the analyses, several policy recommendations were made, including recommendations relating to improvements in the school learning environment, effective teaching time, and quality of teaching. Regarding the assessment and future research, recommendations regarding improvement of the measurement of important EER predictors in the TIMSS questionnaires, and recommendations for further EER research in the region, were also given.

12.2 German Summary

Eine der wichtigsten Funktionen moderner Bildungssysteme ist die Qualifizierung ihrer Schüler. Die Ausbildung der jungen Generation hat nicht nur Vorteile für den Einzelnen, z. B. durch bessere Chancen auf gute Arbeitsbedingungen, höhere Löhne und eine bessere Teilhabe an der Gesellschaft, sondern resultiert auch in Vorteilen für die Gesellschaft als Ganzes, indem sie ihre Wirtschaft auf dem Weltmarkt wettbewerbsfähiger macht. Daher ist die Verbesserung von Bildungsqualität auch ein wichtiges Thema auf der Agenda des Golf-Kooperationsrats (GCC), einer zwischenstaatlichen Vereinigung, die Bahrain, Kuwait, Oman, Katar, Saudi-Arabien und die Vereinigten Arabischen Emirate politisch und wirtschaftlich zusammenschließt. In den Län-

dern des GCC finden gegenwärtig umfangreiche und rasche wirtschaftliche und soziale Veränderungen statt, wobei sich die traditionell orientierten Ölmonarchien in Wissensgesellschaften verwandeln und sich mehr und mehr dem globalen Wettbewerb auf dem internationalen Markt stellen. Die vorliegende Dissertation fokussiert auf den Raum des GCC, da ihre Mitgliedstaaten viele Gemeinsamkeiten aufweisen, wie etwa ähnliche soziale und kulturelle Werte, religiöse Überzeugungen und historische Ereignisse. Zudem teilen sie eine gemeinsame Sprache, und ihr Bildungskontext unterscheidet sich stark von dem des Westens. Die GCC-Länder gehören zu den wohlhabendsten Ländern der Welt, und vor allem aufgrund des Exports natürlicher Ressourcen gelang es der Region, die quantitativen Dimensionen der Schulbildung in den letzten Jahrzehnten erheblich zu verbessern. Qualitative Aspekte der Bildung liegen im internationalen Vergleichen jedoch immer noch zurück. Die GCC-Länder gehören noch immer zu den Ländern mit den niedrigsten Plätzen im Ranking der internationalen Large-scale assessments wie TIMSS (Mullis, Martin et al., 2016), PIRLS (Mullis, Martin, Foy, & Drucker, 2012) oder PISA (OECD, 2016a). Darüber hinaus weist die Region nach wie vor große Ungleichheiten auf in Bezug auf das Geschlecht zugunsten von Mädchen, die bei den Naturwissenschaften in der vierten Klasse in Saudi-Arabien bis zu nahezu 80 Punkte erreichen. In den meisten Ländern des Golf-Kooperationsrates treten auch große Leistungsunterschiede zwischen der nationalen Bevölkerung und der meist bessere Ergebnisse erzielenden ausländischen Bevölkerung auf, die in den Vereinigten Arabischen Emiraten in den Naturwissenschaften eine Differenz von 110 Punkten erreicht.

Es war das Ziel dieser Studie, Leistungsunterschiede von Grundschülern in den GCC-Staaten in Bezug auf Mathematik und Naturwissenschaften aus der Perspektive von Bildungseffektivität zu untersuchen. Leistungsunterschiede wurden mittels Sekundäranalysen basierend auf Trends in International Mathematics and Science Study (TIMSS) 2015 Daten untersucht. TIMSS ist eine groß angelegte internationale und vergleichende Untersuchung, die alle vier Jahre durchgeführt wird und sich auf Mathematik und naturwissenschaftliche Leistungen in der vierten und achten Klasse konzentriert. Zusätzlich zum Mathematik- und Naturwissenschaftstest werden TIMSS-Schüler, ihre Lehrer und ihre Schulleiter gebeten, Fragebögen über ihre Schul- und Unterrichtskontexte für das Lernen von Mathematik und Naturwissenschaften auszufüllen. In der vierten Klasse wird auch den Eltern der getesteten Schüler ein Hintergrundfragebogen vorgelegt, um ergänzende Informationen zum häuslichen Schülerhintergrund zu erhalten. Alle sechs GCC-Länder nahmen in 2015 an beiden TIMSS-Jahrgängen teil.

Während hoch standardisierte internationale Large-scale assessments wie TIMSS eine gute Gelegenheit bieten, tiefer in die internationalen Dimensionen der Bildungswirksamkeitsforschung

(EER) einzutauchen, muss die Tatsache mit berücksichtigt werden, dass TIMSS eine Querschnittsstudie ist, die für unterschiedliche Zwecke entwickelt und als solche nicht speziell für EER konzipiert wurde. Kritikpunkte bezüglich einer Verwendung von Large-scale Assessments für EER wurden in dieser Arbeit soweit wie möglich Rechnung getragen.

Zur Erreichung der Forschungsziele wurden zwei Hauptforschungsfragen formuliert. Die erste Frage lautet: *Inwieweit berücksichtigt TIMSS 2015 wesentliche Faktoren der Bildungseffektivitätsforschung*?

Um die Forschungsfragen zu beantworten, wurde in einem ersten Schritt nach einer umfassenden Literaturrecherche der EER ein geeigneter konzeptioneller Rahmen (d. h. ein Framework) erstellt. Das Framework basierte auf Creemers (1994) umfassendem Modell der Bildungseffektivität, wurde aber durch Elemente aus dem dynamischen Modell (Creemers & Kyriakides, 2008) und teilweise aus dem Scheerens-Modell (Scheerens, 1992) ergänzt. Darüber hinaus berücksichtigte das Framework neuere Forschungen auf dem Gebiet der Unterrichtseffektivität (Helmke, 2009; Klieme & Baumert, 2001; Nilsen et al., 2016; Seidel & Steen, 2005). Das neue Modell unterscheidet sich von Creemers Ansatz in drei wesentlichen Aspekten: Erstens durch Hinzufügen einer *Input*-Dimension; zweitens durch Reklassifizierung von Unterkomponenten bezüglich der *Qualitäts*kategorie auf Klassenebene (und in geringerem Maße auch auf Schulebene); und drittens durch die Einbeziehung neuerer Forschungsergebnisse hinsichtlich von Elementen der *Unterrichtsqualität* und des *Schulklimas*. Das Framework wurde auf die Verwendung der vorliegenden Querschnittsdaten hin ausgerichtet und fokussierte auf allgemeinere Aspekte der Bildungseffektivität in der betrachteten Region (d. h. den GCC-Ländern). In einem zweiten Schritt wurden etwa 170 Variablen aus den TIMSS-Fragebögen entspechend der Modellfaktoren des Frameworks kategorisiert. Dann wurden Hauptkomponenten-, Reliabilitäts- sowie Korrelationsanalysen der Variablen mit mathematischen und wissenschaftlichen Leistungsdaten durchgeführt, um die zugrundeliegenden Konstrukte herauszuarbeiten und ein regionales Modell wichtiger Faktoren möglichst einfach zu spezifizieren. Diese Prozeduren ergaben 5 Variablen auf der Schulebene, 19 Variablen auf der Kursebene und 7 Variablen auf der Studentenebene, die für weitere Analyseschritte beibehalten wurden.

Während die Stärke der Korrelationen zwischen Modellvariablen und Ergebnissen je nach Land und Schulfach unterschiedlich ausfiel, ergaben die Ergebnisse, dass die TIMSS 2015 Fragebogenvariablen eine zufriedenstellende Abdeckung des erstellten Frameworks bezüglich von Indikatoren aufweisen, die gemäß der konsultierten Literatur zur Bildungseffektivität empirisch nachweisbar mit den Schülerergebnissen in Zusammenhang stehen. Aus einer theoretischen

Perspektive heraus konnten für alle Faktoren, mit Ausnahme der Qualität des Anleitungsfaktors auf Schulebene, passende TIMSS 2015 Variablen oder Proxies gefunden werden. Allerdings waren die Variablen nicht gleichmäßig verteilt – ein stärkerer Fokus im Fragebogen wurde auf inputbezogene Variablen, sowie auf Variablen zur Messung der Qualität des Unterrichts auf Lehrerebene gelegt. Der Faktor *Lerngelegenheiten* auf Schulebene musste aus dem endgültigen Modell entfernt werden, da die einzige passende Variable, die sich auf *Richtlinien zur Zuordnung von Schülern* in verschiedene Gruppen bezog, in keinem der Länder aussagekräftige Korrelationen aufwies.

Die zweite Forschungsfrage lautet: *Welche Bildungsfaktoren haben aus Sicht des EER im Hinblick auf Lernergebnisse in der Primarstufe der GCC-Staaten gemäß dem vorgegebenen Framework die höchste Effektivität?*

Um diese Frage zu beantworten, wurden in dieser Studie Multilevel-Modellierungstechniken angewendet, um die gesamte Varianz zwischen den Schülerleistungen in Anteile innerhalb der Kurse und zwischen den Kursen/ Schulen zu zerlegen. Dabei bestand das wesentliche Ziel darin, das Verhältnis von Schul-, Lehrer- und Kursniveau-Faktoren, die entsprechend des vorgeschlagenen Modell identifiziert wurden, mit der Schülerleistung zu quantifizieren und gleichzeitig die nicht-schulischen häuslichen Einflüsse zu kontrollieren. Für jedes der sechs GCC-Mitglieder wurden insgesamt fünf verschiedene Modellreihen analysiert: Nullmodelle, Level 1 Schülerhintergrundmodelle, Hintergrundmodelle mit aggregierten Variablen für den Schülerhintergrund auch auf Level 2, Variablen auf Schul- und Kursniveau ohne Kontrolle des Hintergrunds, und die endgültigen Modelle mit allen Framework-Variablen zusammen.

Schülerhintergrundfaktoren traten als diejenigen Prädiktoren für Schülerleistung hervor, die in allen sechs Ländern annähernd einheitlich wirkten. Während im Allgemeinen alle sechs Faktoren in der Region in gleicher Richtung wirken, unterschied sich die Stärke der Assoziationen ein wenig: So konnten zum Beispiel keine signifikanten Unterschiede in Oman in Bezug auf den *Nationalitätenstatus* festgestellt werden. Auf der Kursebene zeigten sich eine *klare und strukturierte Anleitung* und die verfügbare *Unterrichtszeit* als die am einheitlichsten wirkenden Faktoren in der gesamten Region, nachdem der Schülerhintergrund kontrolliert wurde. Während Prädiktoren für alle wichtigen Modellfaktoren (*Input, Qualität, Zeit und Lerngelegenheiten*) signifikant mit Mathematik und Naturwissenschaften in einem oder mehreren Ländern korreliert waren, konnte kein regionales Muster in Bezug auf gemeinsame regionale Faktoren festgestellt werden. Hinsichtlich der Anzahl signifikanter Faktoren zeigten die Analysen, dass in

Katar nur eine Variable (nämlich das *Geschlecht des Lehrers*) einen Bezug zu den Schülerergebnissen über den häuslichen Hintergrund hinaus aufweist, während in den Vereinigten Arabischen Emiraten insgesamt sechs Modell-Prädiktoren (von 12) signifikant mit Schülerleistungen assoziiert waren. Auf Schulebene zeigten die drei übrig behaltenen Modellfaktoren signifikante Zusammenhänge mit den Schülerleistungen in der Region auch nach der Kontrolle des Schülerhintergrunds, aber auch hier konnte kein regionales Muster zwischen den Ländern festgestellt werden.

Ein weiterer Befund in Bezug auf die Mehrebenen-Modelle war, dass die Kontrolle des Hintergrunds der Schüler in einigen Fällen Assoziationen zwischen Prädiktoren und Schülerleistung verstärkte, was in der Folge zu mehreren signifikanten Korrelationen führte, die zuvor im unkontrollierten Modell nicht auftraten. Es wurde hier angenommen, dass solche Ergebnisse durch den sogenannten *Suppressionseffekt* erklärt werden können.

Die Ergebnisse der Varianzzerlegung zeigten, dass die Vereinigten Arabischen Emirate zwischen den Kursen die größten Varianzkomponenten zeigten (59% für Mathematik und 55% für Naturwissenschaft), während der Anteil der Varianz zwischen den Kursen in Bahrain am geringsten war (24% für Mathematik und 29% für Naturwissenschaft). Der Anteil von Level 2 Varianz, der durch Prädiktoren der Schülerzusammensetzung erklärt werden konnte, lag im Bereich von 12% in Oman und Saudi-Arabien bis zu 28% in Katar für Mathematik, und von 11% in Oman bis zu 30% in den Vereinigten Arabischen Emiraten für Naturwissenschaften.

Die endgültigen Modelle mit allen Faktoren zusammen erklärten zwischen 27% der Level 2 Varianz in Oman und 46% in Katar für Mathematik und zwischen 24% in Oman und 51% in Katar für Naturwissenschaft. In den meisten Ländern war etwa die Hälfte der erklärten Level-2-Varianz auf die Kompositionseffekte von den auf Level 2 aggregierten Schülerhintergrundvariablen zurückzuführen. Die zusätzliche Varianz, die sich aus den Kurs- und schulischen Faktoren ergab, reichte von 7% in Kuwait für Mathematik bis zu 27% für Naturwissenschaft in Katar.

Die Ergebnisse zeigen, dass trotz gewisser Ähnlichkeiten in den Bildungskontexten der Region, die in relativ geringen Gesamtleistungsniveaus und in der Regel hohen Unterschieden in Bezug auf Geschlecht und Nationalität resultierten, sich die Bedingungen sich in den einzelnen Ländern in bestimmten Aspekten auch stark unterschieden.

Erklärungen für die vergleichsweise schwache Gesamtleistung der GCC-Länder in internationalen Large-Scale Assessments werden oft mit Unzulänglichkeiten in der Bildungsqualität und

mit mangelnder Relevanz für die Bedürfnisse des Arbeitsmarktes in Zusammenhang gebracht (z.B. Bahgat, 1999; BouJaoude & Dagher, 2009; UNDP, 2003; UNESCO, 2012a). Die hohe geschlechtsspezifische Ungleichheit zugunsten von Mädchen ist teilweise auf geringere Motivation von Jungen und ihren (oft aus dem Ausland stammenden männlichen) Lehrern zurückzuführen (Ridge, 2014). Die verfügbare Literatur deutet darauf hin, dass auch eine höhere Leistung der ausländischen Bevölkerung teilweise durch motivationale Faktoren erklärt werden kann: Eine gute Bildung für Staatsangehörige, die von den nationalen Sozialsystemen und Verstaatlichungspolitiken profitieren, erscheint weniger wichtig; Ausländer hingegen sind verpflichtet, das Land zu verlassen, sobald sie arbeitslos werden.

Verschiedene Erklärungen wurden diskutiert, die möglicherweise zu den großen Unterschieden zwischen den Ländern in Bezug auf die Anzahl und Stärke signifikanter effizienzsteigernder Faktoren, aber auch in Bezug auf die Erklärung sehr unterschiedlichen Anteile der erklärten Modellvarianz in der Region beitragen. Ein wesentlicher Unterschied zwischen den Ländern könnte mit den unterschiedlichen Anteilen ausländischer Bevölkerungsgruppen zusammenhängen und – dadurch teilweise bedingt – mit der unabhängigen Entwicklung des Privatsektors, was die größeren Unterschiede zwischen den Gruppen in Ländern wie den Vereinigten Arabischen Emiraten oder Katar erklären könnte. Darüber ist anzunehmen, dass der Grad des Konservatismus in Bezug auf Religion, Tradition und Familienorientierung zu unterschiedlichen Bildungschancen für Jungen und Mädchen führt und den Bildungskontext in vielfältiger Weise beeinflusst. Ein weiterer Faktor, der zu mehr Variabilität in der Region führen könnte, ist die Einführung von Bildungsreformen, die in den sechs Ländern recht unterschiedlich zu funktionieren scheinen. Obwohl alle diese Faktoren zu den großen Schwankungen in den GCC-Ländern beitragen können, erlaubt die aktuelle Studie weder eine Entflechtung der verschiedenen Faktoren noch eine Beurteilung ihrer relativen Bedeutung.

Basierend auf den Ergebnissen der Analysen wurden verschiedene politische Empfehlungen abgegeben, darunter solche hinsichtlich Verbesserungen im schulischen Lernumfeld, effektive Unterrichtszeit und Qualität des Unterrichts. Im Hinblick auf die Bewertung und zukünftige Forschung wurden auch Empfehlungen zur Verbesserung der Messung wichtiger EER-Prädiktoren in den TIMSS-Fragebögen sowie für weitere EER-Forschung in der Region gegeben.

REFERENCES

About Us | IEA (2018, January 15). Retrieved from http://www.iea.nl/about-us

Adick, C. (1992). Modern education in 'non-Western' societies in the light of the world systems approach in Comparative Education. *International Review of Education*, *38*(3), 241–255. https://doi.org/10.1007/BF01101431

Ake, C. F. (2005). Rounding after Multiple Imputation with Non-Binary Categorical Covariates. *Paper presented at the annual meeting of the SAS Users Group International, Philadelphia, PA*, 1–11. Retrieved from http://www2.sas.com/proceedings/sugi30/112-30.pdf

Al-Ani, W. (2016). Alternative education needs in Oman: Accommodating learning diversity and meeting market demand. *International Journal of Adolescence and Youth*, *22*(3), 322–336. https://doi.org/10.1080/02673843.2016.1179204

Al-Awadhi, H. (2016). Bahrain. In *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science* (pp. 1–10). Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/encyclopedia/

Albarazi, Z. (2017). *Regional report on citizenship: The Middle East and North Africa (MENA)*. San Domenico di Fiesole, Italy: Global Citizenship Observatory.

Alkhateeb, H. M. (2001). Gender Differences in Mathematics Achievement Among High School Students in the United Arab Emirates, 1991-2000. *School Science and Mathematics*, *101*(1).

Allison, P. D. (2005). Imputation of Categorical Variables with PROC MI. *Paper presented at the annual meeting of the SAS Users Group International, Philadelphia, PA*, 1–14. Retrieved from https://pdfs.semanticscholar.org/848a/a4861e1415d13a80873c814f7a0b3a8ed378.pdf

AlMaskari, Z., AlMawali, F., AlHarthi, A., & AlRasbi, A. (2016). TIMSS 2015 Encyclopedia | Oman. In *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science.* Retrieved from http://timssandpirls.bc.edu/timss2015/encyclopedia/countries/oman/

Alshumrani, S., Alromi, N. H., & Wiseman, A. W. (2014). *Education for a Knowledge Society in Arabian Gulf Countries* (First edition). *International Perspectives on Education and Society: Volume 24*. Bingley, England: Emerald Group Publishing Limited.

Anderson, E. W. (2012). *Is there a crisis for boys? Gender Differences in Student Achievement and Teacher Training Characteristics in the Gulf Cooperation Council Countries* (Theses and Dissertations. Paper 1394). Retrieved from https://preserve.lehigh.edu/cgi/viewcontent.cgi?article=2394&context=etd

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in Reading and How Children Spend Their Time Outside of School. *Reading Research Quarterly*, *23*(3), 285–303. Retrieved from http://www.jstor.org/stable/748043?seq=1#page_scan_tab_contents

Antecol, H., Eren, O., & Ozbeklik, S. (2012). The Effect of Teacher Gender on Student Achievement in Primary School: Evidence from a Randomized Experiment. *Discussion Paper Series, IZA DP No. 6453*. Retrieved from http://ftp.iza.org/dp6453.pdf

Ardent. (2015). *GCC Education Sector: A growing opportunity*. Retrieved from http://www.ardentadvisory.com/files/GCC-Education-Sector-Report.pdf

Ashby, W. R. (1961). *An Introduction to Cybernetics*: Chapman & Hall. Retrieved from https://books.google.de/books?id=GYnuAAAAMAAJ

Ashton, P., & Crocker, L. (1987). Systematic Study of Planned Variations: The Essential Focus of Teacher Education Reform. *Journal of Teacher Education*, *38*(3), 2–8. https://doi.org/10.1177/002248718703800302

Atherton, J. S. (2011). Language Codes. Retrieved from http://www.doceo.org.uk/background/language_codes.htm

Aziz, H. (2016). Science and Mathematics Education in the GCC Countries. *Teacher Education and Curriculum Studies*, *1*(2), 39–42. Retrieved from http://quspace.qu.edu.qa/bitstream/handle/10576/5132/10.11648.j.tecs.20160102.13.pdf?sequence=1&isAllowed=y

Backhaus, K. (2011). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung* (13., überarb. Aufl.). *Springer-Lehrbuch*. Berlin [u.a.]: Springer.

Bahgat, G. (1999). Education in the Gulf Monarchies: Retrospect and Prospect. *International Review of Education*, *45*(2), 127–136.

Baker, D. P., Goesling, B., & LeTendre, G. K. (2002). Socioeconomic Status, School Quality, and National Economic Development: A Cross-National Analysis of the "Heyneman-Loxley Effect" on Mathematics and Science Achievement. *Comparative Education Review*, *46*(3), 291–312. https://doi.org/10.1086/341159

Baldwin-Edwards, M. (2011). *Labour immigration and labour markets in the GCC countries: national patterns and trends*: Kuwait Programme on Development, Governance and Globalisation in the Gulf States.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of Frequent Classroom Testing. *The Journal of Educational Research*, *85*(2), 89–99.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research*, *61*(2), 213–238. https://doi.org/10.3102/00346543061002213

Barbar, Z., Gardner, & Andrew. (2016). Circular Migration and the Gulf States. In C. Solé, S. Parella Rubio, T. Sordé-Martí, & S. Nita (Eds.), *United Nations University Series on Regionalism: v. 12. Impact of circular migration on human, political and civil rights: A global perspective.* [Place of publication not identified]: Springer International Publishing.

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. https://doi.org/10.3102/0002831209345157

Baumert, J. (Ed.). (2006). *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit ; vertiefende Analysen im Rahmen von PISA 2000* (1. Aufl.). Wiesbaden: VS Verl. für Sozialwiss.

Baumert, J., Bos, W., & Watermann, R. (2000). Mathematische und naturwissenschaftliche Grundbildung im internationalen Vergleich. In J. Baumert, W. Bos, & R. H. Lehmann (Eds.), *TIMSS/III Dritte Internationale Mathematik- und Naturwissenschaftsstudie — Ma-*

thematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Band 1 Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit. VS Verlag für Sozialwissenschaften.

Baumert, J., & Schümer, G. (2001). Schulformen als selektionsbedingte Lernmilieus. In J. Baumert (Ed.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 454–467). Opladen: Leske + Budrich.

Baumert, J., Watermann, R., & Schümer, G. (2003). Disparitäten der Bildungsbeteiligung und des Kompetenzerwerbs. *Zeitschrift für Erziehungswissenschaft*, *6*(1), 46–71. https://doi.org/10.1007/s11618-003-0004-7

Becker, R., & Lauterbach, W. Bildung als Privileg – Ursachen, Mechanismen, Prozesse und Wirkungen. In *Becker, Lauterbach (Ed.) 2010 – Bildung als Privileg* (pp. 11–49). https://doi.org/10.1007/978-3-531-92484-7_1

Bernstein, B. (1971). *Class, Codes and Control* (Vol. 1). London: Paladin.

Bill, J. A. (1984). Resurgent Islam in the Persian Gulf. *Foreign Affairs*, *63*(1), 108. https://doi.org/10.2307/20042088

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Blömeke, S., Olsen, R. V., & Suhl, U. (2016). Relation of Student Achievement to the Quality of Their Teachers and Instructional Quality. In T. Nilsen & J.-E. Gustafsson (Eds.), *Teacher Quality, Instructional Quality and Student Outcomes: Relationships Across Countries, Cohorts and Time* (pp. 21–50). Springer Verlag.

Bloom, B. S. (1968). Learning for Mastery: Instruction and Curriculum. Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1. *Evaluation Comment*, *1*(2), 1–12.

Bos, K. (2002). *Benefits and limitations of large-scale international comparative achievement studies: The case of IEA's TIMSS study*. Retrieved from https://ris.utwente.nl/ws/files/6081834/thesis_K_Bos.pdf

Bos, K., & Kuiper, W. (1999). Modelling TIMSS Data in a European Comparative Perspective: Exploring Influencing Factors on Achievement in Mathematics in Grade 8. *Educational Research and Evaluation*, *5*(2), 157–179. https://doi.org/10.1076/edre.5.2.157.6946

Boscardin, C. K., Aguirre-Munoz, Z., Stoker, G., Kim, J., Kim, M., & Lee, J. (2005). Relationship Between Opportunity to Learn and Student Performance on English and Algebra Assessments. *Educational Assessment*, *10*(4), 307–332. https://doi.org/10.1207/s15326977ea1004_1

Bosker, R. J., & Scheerens, J. (1989). Issues in the interpretation of the results of school effectiveness research. *International Journal of Educational Research*, *13*(7), 741–751. https://doi.org/10.1016/0883-0355(89)90025-6

Bosker, R. J., & Witziers, B. (1996). The magnitude of school effects, or: Does it really matter which school a student attends. *Annual Meeting of the American Educational Research Association, New York.*

Boslaugh, S. (2007). *Secondary data sources for public health: A practical guide*. *Practical guides to biostatistics and epidemiology*. Cambridge: Cambridge University Press. Retrieved from http://www.langtoninfo.co.uk/web_content/9780521870016_excerpt.pdf

Bossert, S. T., Dwyer, D. C., Rowan, B., & Lee, G. V. (1982). The Instructional Management Role of the Principal. *Educational Administration Quarterly*, *18*(3), 34–64. https://doi.org/10.1177/0013161X82018003004

Boudon, R. (1974). *Education, Opportunity, and Social Inequality: Changing Prospects in Western Society*: Wiley-Interscience, 605 Third Avenue, New York, New York 10016 ($12.50).

Boudon, R. (1981). *The Logic of Social Action: An Introduction to Sociological Analysis*. London, Boston: Routledge & Kegan Paul.

BouJaoude, S., & Dagher, Z. R. (2009). Introduction: Science Education in Arab States. In S. BouJaoude & Z. R. Dagher (Eds.), *Cultural perspectives on science education. Handbooks: v. 3. The World of Science Education.* Rotterdam, Boston: Sense Publishers.

Bourdieu, P. (1983). Ökonomisches Kapital, kulturelles Kapital, soziales Kapital. *Kreckel [Hrsg.], Soziale Ungleichheiten*, 183–198.

Bourdieu, P. (1986). The Forms of Capital. *Richardson, J. G. (ed.) Handbook of Theory and Research for the Sociology of Education. Greenwood Press, New York, pp. 241-258.*

Bourdieu, P., & Passeron, J.-C. (1977). *Reproduction in education, society and culture*: London, Sage Publ.

Brandsma, H. P., & Knuver, J.W.M. (1989). Effects of school and classroom characteristics on pupil progress in language and arithmetic. *International Journal of Educational Research*, *13*(7), 777–788. https://doi.org/10.1016/0883-0355(89)90028-1

Brese, F., & Mirazchiyski, P. (2013). *Issues and Methodologies in Large-Scale Assessments: Special issue 2: Measuring students family background in large scale international education studies*. *IERI monograph series*.

Brewer, D. J. (2007). *Education for a new era: Design and implementation of K-12 education reform in Qatar*. Santa Monica, CA: RAND RAND-Qatar Policy Institute.

Brookover, W. B. (1979). *School social systems and student achievement: Schools can make a difference*. *Praeger scientific*. New York: Praeger.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievment. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York, London: Macmillan; Collier Macmillan.

Brophy, S. G. (1985). Interactions of male and female students with male and female teachers. In L. C. Wilkinson & C. B. Marrett (Eds.), *Gender Influences in Classroom Interaction.* Burlington: Elsevier Science.

Buchmann, C. (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. In A. C. Porter & A. Gamoran (Eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement* (150-197). Washington, D.C.: National Academies Press.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte und erw). *PS Psychologie*. München [u.a.]: Pearson Studium. Retrieved from http://www.worldcat.org/oclc/846401787

Campbell, J., Kyriakides, L., Muijs, D., & Robinson, W. (2003). Differential Teacher Effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education*, *29*(3), 347–362. https://doi.org/10.1080/03054980307440

Caro, D. H., & Cortés, D. (2012). Measuring family socioeconomic status: An illustration using data from PIRLS 2006. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *5*, 9–33.

Caro, D. H., Sandoval-Hernández, A., & Lüdtke, O. (2014). Cultural, social, and economic capital constructs in international assessments: an evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement*, *25*(3), 433–450. https://doi.org/10.1080/09243453.2013.812568

Carrington, B., & Skelton, C. (2003). Re-thinking 'role models': Equal opportunities in teacher recruitment in England and Wales. *Journal of Education Policy*, *18*(3), 253–265. https://doi.org/10.1080/02680930305573

Carroll, J. B. (1963). *A model of school learning*. Cambridge, Mass. Retrieved from http://www.worldcat.org/oclc/81722006

Carroll, J. B. (1989). The Carroll Model: A 25-Year Retrospective and Prospective View. *Educational Researcher*, *18*(1), 26–31. https://doi.org/10.3102/0013189X018001026

Cervini, R. A. (2009). Class, school, municipal, and state effects on mathematics achievement in Argentina: a multilevel analysis. *School Effectiveness and School Improvement*, *20*(3), 319–340. https://doi.org/10.1080/09243450802664404

Chan, T. C. (1979). The Impact of School Building Age on Pupil Achievement. Greenville, SC: Office of Schools Facilities Planning.

Chapman, C., Muijs, D., Reynolds, D., Sammons, P., & Teddlie, C. (Eds.). (2015). *The international handbook of educational effectiveness and improvement: Research, policy and practice*. *The Routledge international handbook series*.

Cho, M.-O. (2010). *A comparison of the effectiveness of science education in Korea and South Africa: A multilevel analysis of TIMSS 2003 data*. Retrieved from http://up-etd.up.ac.za/thesis/available/etd-10102011-120955/

Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B., & Perlwitz, M. (1991). Assessment of a Problem-Centered Second-Grade Mathematics Project. *Journal for Research in Mathematics Education*, *22*(1), 3. https://doi.org/10.2307/749551

Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education*: Routledge.

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology*. (94), 95–120.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. *OE: 38001 (suppl.)*. Washington]: U.S. Dept. of Health, Education, and Welfare, Office of Education; [for sale by the Superintendent of Documents, U.S. Govt. Print. Off.].

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, *6*(4), 330–351. https://doi.org/10.1037/1082-989X.6.4.330

Colton, N. (2011). *Social Stratification in the Gulf Cooperation Council States*. Kuwait Programme on Development, Governance (Research Paper No. 14). London.

Comber, L. C., & Keeves, J. P. (1973). *Science Education in Nineteen Countries: An Empirical Study*: Wiley. Retrieved from https://books.google.de/books?id=sMwiQQAACAAJ

Conger, A. J. (1974). A Revised Definition for Suppressor Variables: a Guide To Their Identification and Interpretation. *Educational and psychological measurement*. (34), 35–46. Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/001316447403400105

Cool, V. A., & Keith, T. Z. (1991). Testing a model of school learning: Direct and indirect effects on academic achievement. *Contemporary Educational Psychology*, *16*(1), 28–44. https://doi.org/10.1016/0361-476X(91)90004-5

Cooper, H., Lindsay, J. J., Nye, B., & Greathouse, S. (1998). Relationships among attitudes about homework, amount of homework assigned and completed, and student achievement. *Journal of Educational Psychology*, *90*(1), 70–83. https://doi.org/10.1037/0022-0663.90.1.70

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does Homework Improve Academic Achievement? A Synthesis of Research, 1987-2003. *Review of Educational Research*, *76*(1), 1–62. https://doi.org/10.3102/00346543076001001

Cooperation Council for the Arab States of the Gulf (1981). The Charter. Retrieved from http://www.gcc-sg.org/en-us/AboutGCC/Pages/Primarylaw.aspx

Cotton, K. (1995). Effective Schooling Practices: A Research Synthesis, 1995 Update. *Northwest Regional Educational Laboratory*. Retrieved from http://www.kean.edu/~lelovitz/docs/EDD6005/Effective%20School%20Prac.pdf

Creemers, B. P. M., & Kyriakides, L. (2010). School Factors Explaining Achievement on Cognitive and Affective Outcomes: Establishing a Dynamic Model of Educational Effectiveness. *Scandinavian Journal of Educational Research*, *54*(3), 263–294. https://doi.org/10.1080/00313831003764529

Creemers, B. P. M. (1994). *The effective classroom*. *School development series*. London, New York, NY: Cassell.

Creemers, B. P. M., & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement*, *17*(3), 347–366. https://doi.org/10.1080/09243450600697242

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. *Contexts of learning*. London, New York: Routledge.

Creemers, B. P. M., Kyriakides, L., & Sammons, P. (2010). *Methodological advances in educational effectiveness research* (1st ed). *Quantitative methodology series*. Milton Park, Abingdon, Oxon, New York: Routledge.

Cronbach, L. J., Deken, J. E., & Webb, N. (1976). *Research on Classrooms and Schools: Formulation of Questions, Design and Analysis*. Calif. Stanford Evaluation.

Dar, Y., & Resh, N. (1994). Separating and mixing students for learning: concepts and research. *Pedagogisch Tijdschrift, 19(2)*, 109–126. https://doi.org/10.13140/2.1.2606.1769

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement. *Education Policy Analysis Archives*, *8*(0), 1. https://doi.org/10.14507/epaa.v8n1.2000

De Boer, H., Donker-Bergstra, A. S., & Kostons, Danny D. N. M. (2012). Effective strategies for self-regulated learning: A meta-analysis. Retrieved from https://www.nro.nl/wp-content/uploads/2014/05/PROO_Effective+strategies+for+self-regulated+learning.pdf

De Jong, R., Westerhof, K. J., & Kruiter, J.H. (2004). Empirical Evidence of a Comprehensive Model of School Effectiveness: A Multilevel Study in Mathematics in the 1st Year of Junior General Education in The Netherlands. *School Effectiveness and School Improvement*, *15*(1), 3–31. https://doi.org/10.1076/sesi.15.1.3.27490

Delors, J. (1996). *Learning: The treasure within: report to UNESCO of the International Commision for Education*: Paris: UNESCO.

Demirjian, H. (2015). Teacher Shortage in the Arab World: Policy Implications. Retrieved from http://english.dohainstitute.org/file/Get/af026c26-6f9d-4aae-97c9-4de4d844f1ba

Dettmers, S., Trautwein, U., & Lüdtke, O. (2009). The relationship between homework time and achievement is not universal: evidence from multilevel analyses in 40 countries. *School Effectiveness and School Improvement*, *20*(4), 375–405. https://doi.org/10.1080/09243450902904601

Donaldson, L. (2001). *The Contingency Theory of Organizations*: SAGE Publications. Retrieved from https://books.google.de/books?id=_mdHDAAAQBAJ

Doyle, W. (1985). Effective secondary classroom practices. In R. M. J. Kyle (Ed.), *Reaching for Excellence: An Effective Schools Sourcebook.* White (E.H.) Co,, San Francisco California.

Driessen, G., & Sleegers, P. (2000). Consistency of Teaching Approach and Student Achievement: An Empirical Test, School Effectiveness and School Improvement. *An International Journal of Research, Policy and Practice*, *11*(1), 57–79.

Ebbs, D., & Korsnakova, P. (2016). Translation and Translation Verification for TIMSS 2015. In M. O. Martin, I. V.S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (7.1-7.16). Boston.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational leadership*, *37*(1), 15–24.

Egbert, A. (2012). A clearer picture : national and international testing in the UAE. *International Developments*, *2*. Retrieved from http://research.acer.edu.au/cgi/viewcontent.cgi?article=1007&context=intdev

Ehmke, T., & Siegle, T. (2005). ISEI, ISCED, HOMEPOS, ESCS. *Zeitschrift für Erziehungswissenschaft*, *8*(4), 521–539. https://doi.org/10.1007/s11618-005-0157-7

Ehrenberg, R. G., Goldhaber, D. D., & Brewer, D. J. (1995). Do Teachers' Race, Gender, and Ethnicity Matter? Evidence from the National Educational Longitudinal Study of 1988. *ILR Review*, *48*(3), 547–561. https://doi.org/10.1177/001979399504800312

Elley, W. B. (1992). *How in the World do Students Read? IEA Study of Reading Literacy*. Hamburg: International Association for the Evaluation of Educational Achievement (IEA).

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Engman, M. (2009). Half a century of exporting educational services: Assessing Egypt's role in educating the Arab world. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/citations?doi=10.1.1.537.2789

Evertson, C. M., Hawley, W. D., & Zlotnik, M. (1985). Making a Difference in Educational Quality Through Teacher Education. *Journal of Teacher Education*, *36*(3), 2–12. https://doi.org/10.1177/002248718503600302

Fargues, P. (2011). Immigration without Inclusion: Non-Nationals in Nation-Building in the Gulf States. *Asian and Pacific Migration Journal*, *20*(3-4), 273–292.

Farsoun, S. K. (1997). Class Structure and Social Change in the Arab World. In N. S. Hopkins & S. E. Ibrahim (Eds.), *Arab Society: Class, Gender, Power, and Development.* American University in Cairo Press.

Fend, H. (2004). Was stimmt mit den deutschen Bildungssystemen nicht? Wege zur Erklärung von Leistungsunterschieden zwischen Bildungssystemen. In G. Schümer, K. J. Tillmann, & M. Weiss (Eds.), *Die Institution Schule und die Lebenswelt der Schüler: vertiefende Analysen der PISA-2000-Daten zum Kontext von Schülerleistungen.* VS Verlag für Sozialwissenschaften.

Fend, H. (2006). *Neue Theorie der Schule: Einführung in das Verstehen von Bildungssystemen : [Lehrbuch]* (1. Aufl.). *Lehrbuch*. Wiesbaden: VS, Verl. für Sozialwiss.

Field, A. (2004). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner* (Repr). *Introducing statistical methods*. London [u.a.]: Sage Publs.

Fishbein, M., & Ajzen, I. (1975). *Belief attitude, intention, behavior: An introduction to theory and research*: Reading, MA: Addison-Wesley.

Foy, P. (2017). *TIMSS 2015 user guide for the international database*. Chestnut Hill MA: TIMSS and PIRLS International Study Center, Lynch School of Education.

Foy, P., & Yin, L. (2016). Scaling the TIMSS 2015 Achievement Data. In M. O. Martin, I. V.S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (13.1-13.62). Boston.

Fraser, B. J. (1994). Research classroom and school climate. In D. Gabel (Ed.), *Handbook of Research on Science Teaching and Learning* (pp. 493–541). Macmillan.

Gage, N. L. (1978). *The Scientific Basis of the Art of Teaching*: Teachers College Press. Retrieved from https://books.google.de/books?id=paRiQgAACAAJ

Galal, A. (2008). *The Road Not Traveled*: World Bank. Retrieved from http://siteresources.worldbank.org/INTMENA/Resources/EDU_Flagship_Full_ENG.pdf

Ganzeboom, H. B.G., Graaf, P. M. D., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1–56. https://doi.org/10.1016/0049-089X(92)90017-B

GCC: Total population and percentage of nationals and non-nationals in GCC countries (latest national statistics, 2010-2015) (2015). Retrieved from http://gulfmigration.eu/total-population-and-percentage-of-nationals-and-non-nationals-in-gcc-countries-latest-national-statistics-2010-2015/

GCC: Total population and percentage of nationals GCC: Total population and percentages of nationals and foreign nationals in GCC countries (2017). Retrieved from http://gulfmigration.eu/gcc-total-population-percentage-nationals-foreign-nationals-gcc-countries-national-statistics-2010-2016-numbers/

Goldhaber, D. D., & Brewer, D. J. (2000). Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement. *Educational Evaluation and Policy Analysis*, *22*(2), 129–145. https://doi.org/10.3102/01623737022002129

Good, T. L., & Brophy, J. (1986). School effects. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 570–602). New York, London: Macmillan; Collier Macmillan.

Good, T. L., & Grouws, D. A. (1979). The Missouri Mathematics Effectiveness Project: An experimental study in fourth-grade classrooms. *Journal of Educational Psychology*, *71*(3), 355–362. https://doi.org/10.1037/0022-0663.71.3.355

Good, T. L., & Weinstein, R. S. (1986). Schools make a difference: Evidence, criticism, and new directions. *American Psychologist*, *41*(10), 1090–1097. https://doi.org/10.1037/0003-066X.41.10.1090

Gorard, S. (2006). Is there a school mix effect? *Educational Review*, *58*(1), 87–94. https://doi.org/10.1080/00131910500352739

Graaf, N. D. D., Graaf, P. M. D., & Kraaykamp, G. (2000). Parental Cultural Capital and Educational Attainment in the Netherlands: A Refinement of the Cultural Capital Perspective. *Sociology of Education*, *73*(2), 92. https://doi.org/10.2307/2673239

Gray, J. (2004). School effectiveness and the 'other outcomes' of secondary schooling: A reassessment of three decades of British research. *Improving Schools*, *7*(2), 185–198. https://doi.org/10.1177/1365480204047348

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The Effect of School Resources on Student Achievement. *Review of Educational Research*, *66*(3), 361–396. https://doi.org/10.3102/00346543066003361

Gregory, K. D., & Martin, M. O. (2001). *Technical Standards for IEA Studies: An Annotated Bibliography.*

Gruehn, S. (2000). *Unterricht und schulisches Lernen: Schüler als Quellen der Unterrichtsbeschreibung*: Waxmann.

Haertel, G. D., Walberg, H. J., & Haertel, E. H. (1981). Socio-Psychological Environments and Learning: A Quantitative Synthesis. *British Educational Research Journal,*, *7*(1), 27–36. Retrieved from http://www.jstor.org/stable/1501325

Hallinger, P., & Murphy, J. F. (1986). The Social Context of Effective Schools. *American journal of education*. (328-355). Retrieved from http://www.jstor.org/stable/1085156

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of economic literature*, *24*, 1141–1177. Retrieved from http://www.worldcat.org/oclc/78838907

Hanushek, E. A. (1995). Interpreting Recent Research on Schooling in Developing Countries. *The World Bank Research Observer*, *10*(2), 227–246. https://doi.org/10.1093/wbro/10.2.227

Hanushek, E. A., & Woessmann, L. (2008). The Role of Cognitive Skills in Economic Development. *Journal of economic literature*, *46*(3), 607–668.

Harber, C., & Muthukrishna, N. (2000). School Effectiveness and School Improvement in Context: The Case of South Africa. *School Effectiveness and School Improvement*, *11*(4), 421–434. https://doi.org/10.1076/sesi.11.4.421.3559

Harker, R., & Tymms, P. (2004). The Effects of Student Composition on School Outcomes. *School Effectiveness and School Improvement*, *15*(2), 177–199. https://doi.org/10.1076/sesi.15.2.177.30432

Harris, A., Chapman, C., Muijs, D., Russ, J., & Stoll, L. (2007). Improving schools in challenging contexts: Exploring the possible. *School Effectiveness and School Improvement*, *17*(4), 409–424. https://doi.org/10.1080/09243450600743483

Harvey, L., & Green, D. (1993). Defining Quality. *Assessment & Evaluation in Higher Education*, *18*(1), 9–34. https://doi.org/10.1080/0260293930180102

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hauser, R. M. (1970). Context and Consex: A Cautionary Tale. *American Journal of Sociology*, *75*(4), 645–664. Retrieved from http://www.jstor.org/stable/2775907?seq=1#page_scan_tab_contents

Heckhausen, H. (1981). Chancengleichheit. In H. Schiefele & A. Krapp (Eds.), *Handlexikon zur pädagogischen Psychologie.* München: Ehrenwirth.

Hegarty-Hazel, E., & Prosser, M. (1991a). Relationship between students' conceptual knowledge and study strategies -- part 2: Student learning in biology. *International Journal of Science Education*, *13*(4), 421–429. https://doi.org/10.1080/0950069910130405

Hegarty-Hazel, E., & Prosser, M. (1991b). Relationship between students' conceptual knowledge and study strategies-part 1: Student learning in physics. *International Journal of Science Education*, *13*(3), 303–312. https://doi.org/10.1080/0950069910130308

Heid, H. (2000). Qualität. Überlegungen zur Begründung einer pädagogischen Beurteilungskategorie, 41–51. Retrieved from http://www.pedocs.de/volltexte/2014/8484/pdf/Heid_2000_Qualitaet_Ueberlegungen_zur_Begruendung.pdf

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts*: Klett/Kallmeyer. Retrieved from https://books.google.de/books?id=MHdxOwAACAAJ

Herman, J. L., Klein, D. C. D., & Abedi, J. (2000). Assessing Students' Opportunity to Learn: Teacher and Student Perspectives. *Educational Measurement: Issues and Practice*, *19*(4), 16–24. https://doi.org/10.1111/j.1745-3992.2000.tb00042.x

Herrnson, P. S. (1995). Replication, Verification, Secondary Analysis, and Data Collection in Political Science. *Political Science and Politics*, *28*(3), 452–455.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking Pedagogical Content Knowledge: Conceptualizing and Measuring Teachers' Topic-Specific Knowledge of Students. *Journal for Research in Mathematics Education*, *39*(4), 372–400.

Hill, P. W., & Rowe, K. J. (1996). Multilevel Modelling in School Effectiveness Research. *School Effectiveness and School Improvement*, *7*(1), 1–34. https://doi.org/10.1080/0924345960070101

Hincapie, D. (2016). Do Longer School Days Improve Student Achievement? Evidence from Colombia. *IDB Working Paper Series*, *IDB-WP-679*.

Hopkins, D. (2001). *School improvement for real*. *Educational change and development*. London, New York: RoutledgeFalmer.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Mulitilevel Analysis: Techniques and applications, Third Edition [Kindle]*: Taylor & Francis.

Hvidt, M. (2016). Challenges to implementing 'Knowledge based economies' in the Gulf region. *Center for Mellemøtstudier, Syddanks Universitet, Sønderborg*. Retrieved from http://www.sdu.dk/-/media/files/om_sdu/centre/c_mellemoest/videncenter/artikler/2016/hvidt+article+(sept+16).pdf

IBM Corp. (2011). IBM SPSS Statistics for Windows, Version 20. Armonk, NY.

International Labour Office. (2012). *International Standard Classification of Occupations 2008: ISCO-08*. Genève: International Labour Office.

International Monetary Fund (2017). World Economic and Financial Surveys: World Economic Outlook Database. Retrieved from http://www.imf.org/external/pubs/ft/weo/2017/01/weodata/index.aspx

Isac, M. M. (2015). *Effective civic and citizenship: A cross-cultural perspective*, [Groningen].

Jarrar, H., & Alharqan, A. (2016). TIMSS 2015 Encyclopedia | Qatar. In *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science.* Retrieved from http://timssandpirls.bc.edu/timss2015/encyclopedia/countries/qatar/

Jencks, C. (1972). *Inequality: A reassessment of the effect of family and schooling in America*. New York: Basic Books.

Johansone, I. (2016). Survey Operations Procedures in TIMSS 2015. In M. O. Martin, I. V.S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (6.1-6.22). Boston.

Jones, L. V., Davenport, E. C., Bryson, A., Bekhuis, T., & Zwick, R. (1986). Mathematics and Science Test Scores as Related to Courses Taken in High School and Other Factors. *Journal of Educational Measurement*, *23*(3), 197–208. Retrieved from http://www.jstor.org/stable/1434607

Jungbauer-Gans, M. (2004). Einfluss des sozialen und kulturellen Kapitals auf die Lesekompetenz: Ein Vergleich der PISA 2000-Daten aus Deutschland, Frankreich und der Schweiz. *Zeitschrift für Soziologie*, *33*(5), 375–397.

Kane, T., & Cantrell, S. (2012). *Gathering feedback for teaching. Combining high-quality observations with student surveys and achievement gains: METProject Research Paper*. Bill & Melinda Gates Foundation. Seattle, WA. Retrieved from http://files.eric.ed.gov/fulltext/ED540960.pdf

Kapiszewski, A. (2006). Arab Versus Asian Migrant Workers in the GCC Countries. *UN/POP/EGM/2006/02*, 1–20.

Khan, F. (2015). *Seminar on Mathematics Learning Outcomes: TIMSS Results in the Gulf Cooperation Council (GCC): Seminar Report*. Doha.

Kirk, D. (2011). The "knowledge society" in the Middle East: Education and the development of knowledge societies. In *Second annual GCES symposium conference proceedings: Intersections of the public and private education in the GCC.* Symposium conducted at the meeting of Gulf Comparative Education Society, Ras Al Khaimah, United Arab Emirates.

Klieme, E., & Baumert, J. (Eds.). (2001). *TIMSS-Impulse für Schule und Unterricht: Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumentation. Bonn: Bundesministerium für Bildung und Forschung.*

Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In D. PISA-Konsortium & J. Baumert (Eds.), *PISA 2000 — Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland.* VS Verlag für Sozialwissenschaften.

Knuver, A. W.M., & Brandsma, H. P. (1993). Cognitive and Affective Outcomes in School Effectiveness Research. *School Effectiveness and School Improvement*, *4*(3), 189–204. https://doi.org/10.1080/0924345930040302

Koballa, T. R. (1988). Attitude and related concepts in science education. *Science Education*, *72.* https://doi.org/10.1002/sce.3730720202

Kyriakides, L. (2004). Differential School Effectiveness in Relation to Sex and Social Class: Some Implications for Policy Evaluation. *Educational Research and Evaluation*, *10*(2), 141–161. https://doi.org/10.1076/edre.10.2.141.27907

Kyriakides, L. (2005). Extending the Comprehensive Model of Educational Effectiveness by an Empirical Investigation. *School Effectiveness and School Improvement*, *16*(2), 103–152. https://doi.org/10.1080/09243450500113936

Kyriakides, L. (2006). Using international comparative studies to develop the theoretical framework of educational effectiveness research: A secondary analysis of TIMSS 1999 data. *Educational Research and Evaluation*, *12*(6), 513–534. https://doi.org/10.1080/13803610600873986

Kyriakides, L., Campbell, R. J., & Gagatsis, A. (2000). The Significance of the Classroom Effect in Primary Schools: An Application of Creemers' Comprehensive Model of Educational Effectiveness. *School Effectiveness and School Improvement*, *11*(4), 501–529. https://doi.org/10.1076/sesi.11.4.501.3560

Kyriakides, L., & Charalambous, C. (2005). Using educational effectiveness research to design international comparative studies: Turning limitations into new perspectives. *Research Papers in Education*, *20*(4), 391–412. https://doi.org/10.1080/02671520500335816

Kyriakides, L., & Creemers, B. P.M. (2008). Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: a study testing the validity of the dynamic model. *School Effectiveness and School Improvement*, *19*(2), 183–205. https://doi.org/10.1080/09243450802047873

Kyriakides, L., & Creemers, B. P.M. (2009). The effects of teacher factors on different outcomes: two studies testing the validity of the dynamic model. *Effective Education*, *1*(1), 61–85. https://doi.org/10.1080/19415530903043680

Lamb, S., & Fullarton, S. (2001). Classroom And School Factors Affecting Mathematics Achievement: A Comparative Study of the US and Australia Using TIMSS. *TIMSS Australia Monograph Series (10)*. Retrieved from http://research.acer.edu.au/timss_monographs/10

Lankes, E.-M., Bos, W., Mohr, I., Plaßmeier, N., & Schwippert, K. (2003). Lehr- und Lernbedingungen in den Teilnehmerländern. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, G. Walther, & R. Valtin (Eds.), *Erste Ergebnisse aus IGLU: Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich.* Münster: Waxmann Verlag.

LaRoche, S., Joncas, M., & Foy, P. (2016). Sample Design in TIMSS 2015. In M. O. Martin, I. V.S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (3.1-3.37).

Boston. Retrieved from http://timss.bc.edu/publications/timss/2015-methods/chapter-3.html

Lauder, H., Jamieson, I., & Wikely, F. (2003). Models of Effective Schools: Limits and Capabilities. In R. Slee, S. Tomlinson, & G. Weiner (Eds.), *School Effectiveness for Whom?*. Taylor & Francis.

Lavy, V. (2010). Do differences in school's instruction time explain international achievement gaps in math, science, and reading? Evidence from developed and developing countries. (Working Paper 16227). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.379.978&rep=rep1&type=pdf

Lee, J.-W., & Barro, R. (2001). Schooling Quality in a Cross-Section of Countries. *Economica*, *68*(272), 465–488. https://doi.org/10.1111/1468-0335.d01-12

Lenkeit, J. (2012). *Achievement status and growth as predictors of educational outcomes and effectiveness* (Doctoral dissertation). Retrieved from http://ediss.sub.uni-hamburg.de/volltexte/2013/6041/pdf/Dissertation.pdf

Leung, F. K.S. (2002). Behind the High Achievement of East Asian Students. *Educational Research and Evaluation*, *8*(1), 87–108. https://doi.org/10.1076/edre.8.1.87.6920

Levine, D. U., & Lezotte, L. W. (1990). *Unusually Effective Schools: A Review and Analysis of Research and Practice*. Madison, WI: National Center for Effective Schools Research and Development. Retrieved from http://collections.lakeforest.edu/files/original/40080ec6e13f64e8260b40a79d14bd61.pdf

Lietz, P., Wagemaker, H., Neuschmidt, O., & Hencke, J. (Eds.). (2008). *Educational Issues in the Middle East North Africa Region: Outcomes of the IEA Arab Region Training Seminar Series 2006/2007*. Amsterdam.

Little, R. J. A., & Rubin, D. B. (1989). The Analysis of Social Science Data with Missing Values. *Sociological Methods & Research*, *18*(2-3), 292–326. https://doi.org/10.1177/0049124189018002004

Louis, K. S., Wahlstrom, K. L., Michlin, M., Gordon, M., Thomas, E., Leithwood, K., . . . Moore, S. (2010). Learning from Leadership: Investigating the Links to Improved Student Learning: Final Report of Research to the Wallace Foundation.

Low, L., & Salazar, L. C. (2011). *The Gulf Cooperation Council: A Rising Power and Lessons for ASEAN*: Institute of Southeast Asian Studies. Retrieved from https://books.google.de/books?id=YtE9LiyqttQC

Luyten, H. (1994). Stability of school effects in dutch secondary education: The impact of variance across subjects and years. *International Journal of Educational Research*, *21*(2), 197–216. https://doi.org/10.1016/0883-0355(94)90032-9

Maktab (2007). *Encyclopedia Britannica Online*. Retrieved from https://www.britannica.com/topic/maktab

Mandeville, G. K., & Anderson, L. W. (1987). The Stability of School Effectiveness Indices across Grade Levels and Subject Areas. *Journal of Educational Measurement*, *24*(3), 203–216. Retrieved from http://www.jstor.org/stable/pdf/1434631.pdf?refreqid=excelsior%3A4df92955e9bd8ca577b9f0342f540705

Mansour, N., & Al-Shamrani, S. (2015). *Science education in the Arab Gulf States: Visions, sociocultural contexts and challenges*. *Cultural and historical perspectives on science education. Distinguished contributors: volume 4*. Rotterdam: SensePublishers.

Marks, G. N., Cresswell, J., & Ainley, J. (2006). Explaining socioeconomic inequalities in student achievement: The role of home and school factors. *Educational Research and Evaluation*, *12*(2), 105–128. https://doi.org/10.1080/13803610600587040

Martin, M. O., & Mullis, I. V.S. (Eds.). (2013). *TIMSS and PIRLS 2011: Relationships Among Reading, Mathematics, and Science Achievement at the Fourth Grade—Implications for Early Learning.*: Boston: International Study Center, Lynch School of Education, Boston College. Retrieved from https://timssandpirls.bc.edu/timsspirls2011/downloads/TP11_Relationship_Report.pdf

Martin, M. O., Mullis, I. V.S., Gregory, K. D., Hoyle, C., & Shen, C. (Eds.). (2000). *Effective schools in science and mathematics: IEA's Third International Mathematics and Science Study*. Boston. Retrieved from https://isc.bc.edu/timss1995i/TIMSSPDF/T95_Eff-School.pdf

Martin, M. O., Mullis, I. V.S., Beaton, A. E., Gonzalez, E. J., Smith, T. A., & & Kelly, D. L. (1997). *Science Achievement in the Primary School Years. IEA's Third International Mathematics and Science Study (TIMSS)*. Campion Hall, Chestnut Hill, MA 02167.

Martin, M. O., Mullis, I. V.S., & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Boston: TIMSS & PIRLS International Study Center. Retrieved from https://timss.bc.edu/timss2007/mathreport.html

Martin, M. O., Mullis, I. V.S., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Science*. Boston: TIMSS & PIRLS International Study Center. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/wp-content/uploads/filebase/full%20pdfs/T15-International-Results-in-Science-Grade-4.pdf

Martin, M. O., Mullis, I. V.S., Foy, P., & Stanco, G. M. (2012). *TIMSS 2011 international results in science*. Chestnut Hill MA: IEA TIMSS & PIRLS International Study Center Lynch School of Education Boston College. Retrieved from https://www.bc.edu/content/dam/files/research_sites/timssandpirls/timss2011/downloads/T11_IR_Science_FullBook.pdf

Martin, M. O., Mullis, I. V.S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1999). School contexts for learning and instruction: lEA's Third International Mathematics and Science Study (TlMSS).

Martin, M. O., Mullis, I. V.S., & Hooper, M. (Eds.). (2016). *Methods and Procedures in TIMSS 2015*. Boston. Retrieved from http://timss.bc.edu/publications/timss/2015-methods.html

Marx, K. (2012). *Das Kapital*: Jazzybee Verlag. Retrieved from https://books.google.de/books?id=N7aLMVPEMMQC

Marzano, R. J., & Kendall, J. S. (2006). *The New Taxonomy of Educational Objectives*: SAGE Publications. Retrieved from https://books.google.de/books?id=JT4KAgAAQBAJ

Marzano, R. J. (2003). *What works in schools: Translating research into action*. Alexandria: Association for Supervision and Curriculum Development.

Maslowski, R. (2001). *School culture and school performance: An explorative study into the organizational culture of secondary schools and their effects*. Retrieved from http://doc.utwente.nl/36122/1/t0000012.pdf

May, H. (2006). A multilevel Bayesian item response theory method for scaling. *Journal of Educational and Behavioral Statistics, 31(1),*, 63–79.

Mayer, D. P., Mullens, J. E., & Moore, M. T. (2000). *Monitoring school quality: An indicators report*. Washington, D.C.: U.S. Dept. of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

McConney, A., & Perry, L. B. (2010). Socioeconomic status, self-efficacy, and mathematics achievement in Australia: a secondary analysis. *Educational Research for Policy and Practice*, *9*(2), 77–91. https://doi.org/10.1007/s10671-010-9083-4

McDonnell, L. M. (1995). Opportunity to Learn as a Research Concept and a Policy Instrument. *Educational Evaluation and Policy Analysis*, *17*(3), 305–322. https://doi.org/10.3102/01623737017003305

McLeod, D. B. (1992). Research on affect in mathematics education: A reconceptualization. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics.* New York [u.a.]: Macmillan.

Metz, H. C. (1993). Persian Gulf States: A country Study.

Ministry of Education Oman & World Bank. (2012). *Education in Oman: The drive for quality*. Oman: Ministry of Education.

Moghadam, V., & Decker, T. (2010). Social Change in the Middle East. In E. M. Lust-Okar (Ed.), *The Middle East* (12th ed., pp. 73–106). Washington, DC: CQ Press. Retrieved from http://nuweb8.neu.edu/advance/legacy/wp-content/uploads/Moghadam-and-Decker.pdf

Monk, D. H. (1994). Subject area preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review*, *13*(2), 125–145. https://doi.org/10.1016/0272-7757(94)90003-5

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters: The junior years*. Wells: Open books.

Mueller, C. W., & Parcel, T. L. (1981). Measures of socioeconomic status: Alternatives and recommendations. *Child Development*, *52(1)*, 13–30.

Muijs, D., Campbell, J., Kyriakides, L., & Robinson, W. (2005). Making the Case for Differentiated Teacher Effectiveness: An Overview of Research in Four Key Areas. *School Effectiveness and School Improvement*, *16*(1), 51–70. https://doi.org/10.1080/09243450500113985

Muijs, D., Kyriakides, L., van der Werf, Greetje, Creemers, B. P. M., Timperley, H., & Earl, L. (2014). State of the art – teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, *25*(2), 231–256. https://doi.org/10.1080/09243453.2014.885451

Muijs, D., & Reynolds, D. (2000). School Effectiveness and Teacher Effectiveness in Mathematics: Some Preliminary Findings from the Evaluation of the Mathematics Enhancement Programme (Primary). *School Effectiveness and School Improvement*, *11*(3), 273–303. https://doi.org/10.1076/0924-3453(200009)11:3;1-G;FT273

Muijs, D., & Reynolds, D. (2003). Student Background and Teacher Effects on Achievement and Attainment in Mathematics: A Longitudinal Study. *Educational Research and Evaluation*, *9*(3), 289–314. https://doi.org/10.1076/edre.9.3.289.15571

Mullis, I. V.S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 International Results in Reading*: TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/pirls2011/downloads/P11_IR_FullBook.pdf

Mullis, I. V.S. (2012). *Timss 2011 encyclopedia: Education policy and curriculum in mathematics and science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/timss2011/encyclopedia-timss.html

Mullis, I. V.S., Cotter, K. E., Fishbein, B. G., & Centurino, V. A. S. (2016). Developing the TIMSS 2015 achievement items. In M. O. Martin, I. V.S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (1.1-1.22). Boston.

Mullis, I. V.S., & Martin, M. O. (Eds.). (2016). *TIMSS 2015 Assessment Frameworks*. Boston: TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/timss2015/frameworks.html

Mullis, I. V.S., Martin, M. O., Beaton, A. E., Gonzalez, E., Kelly, D. L., & Smith, T. A. (1997). *Mathematics Achievement in the Primary School Years. IEA's Third International Mathematics and Science Study (TIMSS)*. Campion Hall, School of Education, Boston College, Chestnut Hill, MA 02167.

Mullis, I. V.S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Boston: TIMSS & PIRLS International Study Center. Retrieved from https://timssandpirls.bc.edu/timss2011/downloads/T11_IR_Mathematics_FullBook.pdf

Mullis, I. V.S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International Results in Mathematics*. Boston: TIMSS & PIRLS International Study Center. Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/wp-content/uploads/filebase/full%20pdfs/T15-International-Results-in-Mathematics.pdf

Mullis, I. V.S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 International Report*. Chestnut Hill, MA. Retrieved from https://timss.bc.edu/pirls2006/intl_rpt.html

Nash, R. (2003). Is the School Composition Effect Real? A Discussion With Evidence From the UK PISA Data. *School Effectiveness and School Improvement*, *14*(4), 441–457. https://doi.org/10.1076/sesi.14.4.441.17153

National Center for Education Development. (2016). Kuwait. In *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science* (pp. 1–6). Retrieved from http://timssandpirls.bc.edu/timss2015/international-results/encyclopedia/

Neuschmidt, O., Hencke, J., Rutkowski, L., & Rutkowski, D. (2011). *Effective Schools in Arab Countries: An Analysis of Teacher Level Variables Using TIMSS 2007*. Paper presented at the Annual Meeting of the American Educational Research Association in New Orleans, 8-12 April.

Neuschmidt, O. (2016). *Gender Differences in the Gulf States: An Analysis of differential effectiveness in mathematics and science based on Creemers' comprehensive model of educational effectiveness*. Paper presented at the International Congress for School Effectiveness and Improvement in Glasgow, 3-6 January 2016.

Neuschmidt, O., & Aghakasiri, P. (2015). *Differential Effectiveness in Relation to Gender: An Analysis in the Gulf States based on Creemers' comprehensive model of educational effectiveness*. Paper presented at the IEA Research Conference (IRC) in Capetown, 24-26 June, 2015.

Neuschmidt, O., Hencke, J., Rutkowski, L., & Rutkowski, D. (2010). Effective Schools in Arab Educational Systems. A Multi-level Approach Using TIMSS 2003 Data. In D. K. Sharpes (Ed.), *Handbook on international studies in education.* Charlotte, NC: Information Age Pub.

Neuschmidt, O., & Tölle, J. (2017). *Differential Effectiveness for Residents and Immigrants in the Gulf States: An Analysis Based on Creemers' Comprehensive Model of Educational Effectiveness Using TIMSS 2015*. Paper presented at the IEA Research Conference (IRC) in Prague, 28-30 June, 2017.

Nilsen, T., & Gustafsson, J.-E. (Eds.). (2016). *Teacher Quality, Instructional Quality and Student Outcomes: Relationships Across Countries, Cohorts and Time*: Springer Verlag.

Nilsen, T., Gustafsson, J.-E., & Blömeke, S. (2016). Conceptual Framework and Methodology of This report. In T. Nilsen & J.-E. Gustafsson (Eds.), *Teacher Quality, Instructional Quality and Student Outcomes: Relationships Across Countries, Cohorts and Time.* Springer Verlag.

Nixon, L. A., & Robinson, Michael, D. (1999). The Educational Attainment of Young Women: Role Model Effects of Female High School Faculty. *Demography*, *36*(2), 185–1999.

Nuttall, D. L., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research*, *13*(7), 769–776. https://doi.org/10.1016/0883-0355(89)90027-X

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. https://doi.org/10.3102/01623737026003237

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science*. Santa Monica, CA: Rand Corp.

O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel Modeling of Educational Data*: Information Age Publishing, Incorporated. Retrieved from https://books.google.de/books?id=GfonDwAAQBAJ

O'Dwyer, L. M. (2005). Examining the Variability of Mathematics Performance and its Correlates Using Data From TIMSS '95 and TIMSS '99. *Educational Research and Evaluation*, *11*(2), 155–177. https://doi.org/10.1080/13803610500110802

OECD. (2008). *Measuring Improvements in Learning Outcomes: Best Practices to Assess the Value-Added of Schools*. Paris: OECD.

OECD. (2009). *Creating effective teaching and learning environments: First results from TALIS* (2009 ed.): OECD Pub.

OECD (2010a). PISA 2009 Ergebnisse: Zusammenfassung.

OECD. (2010b). *PISA 2009 Results: Overcoming Social Background - Equity in Learning Opportunities and Outcomes (Volume II)*. *Programme for International Student Assessment*. Paris: OECD.

OECD. (2012). *Equity and quality in education - supporting disadvantaged students and schools*. Paris: OECD. Retrieved from https://www.oecd.org/education/school/50293148.pdf

OECD. (2013). *PISA 2012 Results: What Makes Schools Successful (Volume IV)*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-IV.pdf

OECD (2014a). Does Homework Perpetuate Inequities in Education? Advance online publication. https://doi.org/10.1787/5jxrhqhtx2xt-en

OECD. (2014b). *PISA Technical Report.*

OECD. (2016a). *PISA 2015 Results (Volume I)*: OECD Publishing. Retrieved from http://www.oecd.org/education/pisa-2015-results-volume-i-9789264266490-en.htm

OECD. (2016b). *PISA 2015 Results in Focus*. Retrieved from http://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Opdenakker, M.-C., & van Damme, J. (2000). Effects of Schools, Teaching Staff and Classes on Achievement and Well-Being in Secondary Education: Similarities and Differences Between School Outcomes. *School Effectiveness and School Improvement*, *11*(2), 165–196. https://doi.org/10.1076/0924-3453(200006)11:2;1-Q;FT165

Opdenakker, M.-C., van Damme, J., Fraine, F. de, van Landeghem, G., & Onghena, P. (2002). The Effect of Schools and Classes on Mathematics Achievement. *School Effectiveness and School Improvement*, *13*(4), 399–427. https://doi.org/10.1076/sesi.13.4.399.10283

Oxford Business Group. (2012). *The Report: Bahrain 2012*: Oxford Business Group. Retrieved from https://books.google.de/books?id=bZlnZ4j_fJYC

Paccagnella, O. (2006). Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Evaluation Review*, *30*(1), 66–85. https://doi.org/10.1177/0193841X05275649

Pandey, S., & Elliott, W. (2010). Suppressor Variables in Social Work Research: Ways to Identify in Multiple Regression Models. *Journal of the Society for Social Work and Research*, *1*(1), 28–40. https://doi.org/10.5243/jsswr.2010.2

Papanastasiou, C. (2000). Effects of attitudes and beliefs on mathematics achievement. *Studies in Educational Evaluation*, *26*(1), 27.

Papanastasiou, E. C., & Zembylas, M. (2002). The Effect of Attitudes on Science Achievement: A Study Conducted Among High School Pupils in Cyprus. *International Review of Education/ Internationale Zeitschrift fr Erziehungswissenschaft/ Revue inter*, *48*(6), 469–484. https://doi.org/10.1023/A:1021334424571

Pintrich, P. R. (2005). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 451–502). Elsevier Science.

Postlethwaite, N. T., & Ross, K. N. (1992). Effective schools in reading: Implications for educational planners.

Postlethwaite, T. N., & Wiley, D. E. (1992). *The IEA study of science II: Science achievement in twenty-three countries (Vol. 2)*. Oxford: Pergamon Press.

Principal Component Analysis | SPSS Annotated Output. Retrieved from https://stats.idre.ucla.edu/spss/output/principal_components/

Purkey, S. C., & Smith, M. S. (1983). Effective Schools: A Review. *The Elementary School Journal*, *83*(4), 426–452. Retrieved from http://www.jstor.org/stable/1001168

Randeree, K. (2012). Workforce Nationalization in the Gulf Cooperation Council States. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.2825910

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*: SAGE Publications. Retrieved from https://books.google.de/books?id=uyCV0CNGDLQC

Raudenbush, S. W., & Willms, J. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307–335. https://doi.org/10.3102/10769986020004307

Raudenbush, S., & Bryk, A. S. (1986). A Hierarchical Model for Studying School Effects. *Sociology of Education*, *59*(1), 1–17. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=13008288&site=ehost-live

Raven, J. (1991). The Wider Goals of Education: Beyond the 3 Rs. *The Educational Forum*, *55*(4), 343–363. https://doi.org/10.1080/00131729109335666

Reezigt, G. J., Guldemond, H., & Creemers, B. P.M. (1999). Empirical Validity for a Comprehensive Model on Educational Effectiveness. *School Effectiveness and School Improvement*, *10*(2), 193–216. https://doi.org/10.1076/sesi.10.2.193.3503

Reyes, L. H. (1984). Affective Variables and Mathematics Education. *The Elementary School Journal*, *84*(5), 558–581.

Reynolds, A. J. (1991). The middle schooling process: Influences on science and mathematics achievement from the longitudinal study of american youth. *Adolescence*, *26*(101), 133–158.

Reynolds, A. J., & Walberg, H. J. (1991). A structural model of science achievement. *Journal of Educational Psychology*, *83*(1), 97–107. https://doi.org/10.1037/0022-0663.83.1.97

Reynolds, D. (2006). World Class Schools: Some methodological and substantive findings and implications of the International School Effectiveness Research Project (ISERP). *Educational Research and Evaluation*, *12*(6), 535–560. https://doi.org/10.1080/13803610600874026

Reynolds, D., Sammons, P., Fraine, B. de, Townsend, T., & van Damme, J. (2011). Educational Effectiveness Research (EER): A State of the Art Review.

Reynolds, D., Sammons, P., Fraine, B. de, van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): a state-of-the-art review. *School Effectiveness and School Improvement*, *25*(2), 197–230. https://doi.org/10.1080/09243453.2014.885450

Reynolds, D., Teddlie, C., Stringfield, S., & Creemers, B. P. M. (Eds.). (2002). *World Class Schools: International Perspective on School Effectiveness*: Taylor & Francis.

Ridge, N., Farah, S., & Shami, S. (2013). Patterns and perceptions in male secondary school dropouts in the United Arab Emirates: (Working Paper No. 3). Retrieved from http://www.alqasimifoundation.com/en/Publications/Publications/PublicationsDetail.aspx?UrlId=5b7010ff-6e67-48e7-9723-25e0c26e6799

Ridge, N. (2014). *Education and the reverse gender divide in the Gulf States: Embracing the global, ignoring the local*. *International Perspectives on Education Reform*.

Robinson, F. (1993). Technology and Religious Change: Islam and the Impact of Print. *Modern Asian Studies*, *27*(1), 229–251.

Rosenthal, R. (1968). Self-fulfilling prophecies in behavioral research and everyday life. *Claremont Reading Conference Yearbook*, *32*, 15–33.

Rosenthal, R., & Jacobson, L. F. (1968). Pygmalion in the classroom. *The Urban Review*, *3*(1), 16–20.

Rowan, B., & Denk, C. E. (1984). Management Succession, School Socioeconomic Context, and Basic Skills Achievement. *American Educational Research Journal*, *21*(3), 517–537. https://doi.org/10.3102/00028312021003517

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rutkowski, D., & Rutkowski, L. (2008). Private and Public Education: A Cross-National Exploration With TIMSS 2003. *Paper presented at the annual conference of the American Educational Research*.

Rutter, M. (1983). School Effects on Pupil Progress: Research Findings and Policy Implications. *Child Development*, *54*(1), 1. https://doi.org/10.2307/1129857

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. Cambridge Mass.: Harvard University Press.

Saif, A. (n.d.). Class Analysis and the State in The Middle Eastern Countries. Retrieved from http://www.shebacss.com/docs/poedt002-12.pdf

Sakr, A. (2008). GCC States Competing in Educational Reform. Retrieved from http://carnegieendowment.org/sada/20501

Sammons, P. (1999). *School effectiveness: Coming of age in the twenty-first century*. *Contexts of learning*. Lisse, Exton, PA: Swets & Zeitlinger Publishers.

Sammons, P., Davis, S., & Gray, J. (2015). Methodological and scientific properties of school effectiveness research. In C. Chapman, D. Muijs, D. Reynolds, P. Sammons, & C. Teddlie (Eds.), *The Routledge international handbook series. The international handbook of educational effectiveness and improvement: Research, policy and practice* (pp. 25–76).

Sammons, P., Hillman, J., & Mortimore, P. (1995). Key characteristics of effective schools: A review of school effectiveness research. Retrieved from http://www.mp.gov.rs/resursi/dokumenti/dok132-eng-SESI_Key_characteristics_of_effective_schools.pdf

Sammons, P., Mortimore, P., & Thomas, S. (1996). Do schools perform consistently across outcomes and areas? In J. Gray (Ed.), *School development series. Merging traditions: The future of research on school effectiveness and school improvement.* London, New York: Cassell.

Sandoval-Hernández, A., Aghakasiri, P., Wild, J., & Rutkowski, D. (2013). Does increasing hours of schooling lead toimprovements in student learning? Retrieved from http://www.iea.nl/policy_briefs.html.

SAS Institute Inc. (2015). *SAS/STAT® 14.1 User's Guide.* Cary, NC.

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological methods*, *7*(2).

Scheerens, J. (1992). *Effective schooling: Research, theory and practice*. *School development series*. London: Cassell.

Scheerens, J. (2000). School effectiveness in developed and developing countries; a review of the research evidence. Retrieved from http://www.mp.gov.rs/resursi/dokumenti/dok20-eng-IIEP_school_effectiveness.pdf

Scheerens, J. (2004a). Perspectives on Education Quality, Education Indicators and Benchmarking. *European Educational Research Journal*, *3*(1), 115–138. https://doi.org/10.2304/eerj.2004.3.1.3

Scheerens, J. (2004b). Review of school and instructional effectiveness research. Background paper prepared for the Education for All Global Monitoring Report 2005: The Quality Imperative. Paris, UNESCO.

Scheerens, J. (2013). The use of theory in school effectiveness research revisited. *School Effectiveness and School Improvement*, *24*(1), 1–38. https://doi.org/10.1080/09243453.2012.691100

Scheerens, J. (2016). *Educational effectiveness and ineffectiveness: A critical review of the knowledge base*. Dordrecht: Springer.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.

Scheerens, J., Luyten, H., Steen, R., & Luyten-de Thouars, Y. (2007). Review and meta-analyses of school and teaching effectiveness.

Scherer, R., & Gustafsson, J.-E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: An application of multilevel bifactor structural equation modeling. *Frontiers in Psychology*, *6.* https://doi.org/10.3389/fpsyg.2015.01550

Scherer, R., & Nilsen, T. (2016). The Relations Among School Climate, Instructional Quality, and Achievement Motivation in Mathematics. In T. Nilsen & J.-E. Gustafsson (Eds.), *Teacher Quality, Instructional Quality and Student Outcomes: Relationships Across Countries, Cohorts and Time.* Springer Verlag.

Schoenfeld, A. H. (1989). Explorations of Students' Mathematical Beliefs and Behavior. *Journal for Research in Mathematics Education*, *20*(4), 338. https://doi.org/10.2307/749440

Schofer, E., Ramirez, F. O., & Meyer, J. W. (2000). The Effects of Science on National Economic Development, 1970 to 1990. *American Sociological Review*, *65*(6), 866–887. Retrieved from http://www.jstor.org/stable/pdf/2657517.pdf

Schulz-Heidorf, K. (2016). *Individuelle Förderung im Unterricht: Eine Möglichkeit, soziale Herkunft und Schulerfolg zu entkoppeln? Eine Re-Analyse aus IGLU-E 2011* (2. Auflage). Berlin: epubli.

Schümer, G. (2001). Institutionelle Bedingungen schulischen Lernens im Internationalen Vergleich. In J. Baumert (Ed.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 411–427). Opladen: Leske + Budrich.

Schwippert, K. (2001). *Optimalklassen: mehrebenenanalytische Untersuchungen: Eine Analyse hierarchisch strukturierter Daten am Beispiel des Leseverständnisses*. *Pädagogische Psychologie und Entwicklungspsychologie: Vol. 27*. Münster, München [u.a.]: Waxmann.

Seidel, T., & Steen, R. (2005). The indicators on the teaching and learning compared to the review of recent research. In J. Scheerens, T. Seidel, B. Witziers, M. Hendriks, & G. Doornekamp (Eds.), *Positioning the supervision frameworks for primary and secondary education of the Dutch Educational Inspectorate in current educational discourse and validating core indicators against the knowledge base of educational effectiveness research.* Enschede/ Kiel: University of Twente/ Institute for Science Education (IPN).

Shavelson, R. J., McDonnell, L., & Oakes, J. (1989). *Indicators for monitoring mathematics and science education: A sourcebook*. Santa Monica, CA: Rand Corp.

Shulman, L. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, *15*, 4–14.

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417

Smits, J., & Huisman, J. (2012). Determinants of educational participation and gender differences in education in six Arab countries. *NiCE Working Paper 12-102*. Retrieved from www.ru.nl/publish/pages/516298/nice_12102.pdf

Stacey, R. D. (2007). *Strategic management and organisational dynamics: The challenge of complexity to ways of thinking about organisations* (5th ed.). Harlow: Financial Times Prentice Hall.

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using Video Surveys to Compare Classrooms and Teaching Across Cultures: Examples and Lessons From the TIMSS Video Studies. *Educational Psychologist*, *35*(2), 87–100. https://doi.org/10.1207/S15326985EP3502_3

Strand, S. (2010). Do some schools narrow the gap? Differential school effectiveness by ethnicity, gender, poverty, and prior achievement. *School Effectiveness and School Improvement*, *21*(3), 289–314. https://doi.org/10.1080/09243451003732651

Stubbe, T. C. (2003). *Die Messung von kulturellem und ökonomischem Kapital in der internationalen Schulleistungsforschung mit Hilfe von latenten Analyseverfahren* (Diplom). Institut für Soziologie, Hamburg.

Supovitz, J. A., Mayer, D. P., & Kahle, J. B. (2000). Promoting Inquiry-Based Instructional Practice: The Longitudinal Impact of Professional Development in the Context of Systemic Reform. *Educational Policy*, *14*(3), 331–356. https://doi.org/10.1177/0895904800014003001

Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching*, *37*(9), 963–980. https://doi.org/10.1002/1098-2736(200011)37:9<963::AID-TEA6>3.0.CO;2-0

Tatto, M. T., Schwille, J., Senk, S. L., Ingvarson, L., Rowley, G., Peck, R., Bankov, K., Rodriguez, M., & Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M)*. Amsterdam: International Association for the

Evaluation of Educational Achievement (IEA). Retrieved from http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/TEDS-M_International_Report.pdf

Teddlie, C., Kirby, P. C., & Stringfield, S. (1989). Effective versus Ineffective Schools: Observable Differences in the Classroom. *American journal of education*, *97*(3), 221–236. Retrieved from http://www.jstor.org/stable/1085165?seq=1#page_scan_tab_contents

Teddlie, C., & Reynolds, D. (Eds.). (2000). *The international handbook of school effectiveness research*. London, New York: Falmer Press.

Terhart, E. (2000). Qualität und Qualitätssicherung im Schulsystem. Hintergründe - Konzepte - Probleme. *Zeitschrift für Pädagigik*, *46*(6).

The International Labour Organization (2018). ILOSTAT - ILO database of labour statistics. Retrieved from http://www.ilo.org/ilostat/faces/ilostat-home/home?_adf.ctrl-state=9xofseqn4_251&_afrLoop=2156501972136256#!

Thomas, S. (2001). Dimensions of Secondary School Effectiveness: Comparative Analyses Across Regions. *School Effectiveness and School Improvement*, *12*(3), 285–322. https://doi.org/10.1076/sesi.12.3.285.3448

Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997). Stability and Consistency in Secondary Schools' Effects on Students' GCSE Outcomes over Three Years. *School Effectiveness and School Improvement*, *8*(2), 169–197. https://doi.org/10.1080/0924345970080201

TIMSS & PIRLS International Study Center, Boston College. (2016a). *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science*. Retrieved from http://timssandpirls.bc.edu/timss2015/encyclopedia/

TIMSS & PIRLS International Study Center, Boston College (2016b). TIMSS 2015 International Database. Retrieved from https://timssandpirls.bc.edu/timss2015/international-database/

Trautwein, U. (2007). The homework–achievement relation reconsidered: Differentiating homework time, homework frequency, and homework effort. *Learning and Instruction*, *17*(3), 372–388. https://doi.org/10.1016/j.learninstruc.2007.02.009

Trautwein, U., Köller, O., Schmitz, B., & Baumert, J. (2002). Do Homework Assignments Enhance Achievement? A Multilevel Analysis in 7th-Grade Mathematics. *Contemporary Educational Psychology*, *27*(1), 26–50. https://doi.org/10.1006/ceps.2001.1084

Travers, K. J., & Weinzweig, A. I. (1999). The Second International Mathematics Study. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *Studies in mathematics education series: Vol. 11. International comparisons in mathematics education.* London, Philadelphia: Falmer Press.

UNDP. (2003). *Arab Human Development Report 2003: Building a knowledge society*. New York: United Nations Publications.

UNESCO (2000). The Dakar Framework for Action, Education for All: meeting our collective commitments. Retrieved from http://unesdoc.unesco.org/images/0012/001211/121147e.pdf

UNESCO (2012a). Education for All Regional Report 2012 for Arab States: Global Education for All Meeting. Retrieved from http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/ED/ED_new/pdf/ARB_EN.pdf

UNESCO. (2012b). *International Standard Classification of Education (ISCED) 2011*. Paris: Unesco.

UNESCO Institute for Statistics (UIS) (2017). ISCED Mappings. Retrieved from http://uis.unesco.org/en/isced-mappings

United Nations (n.d.a). United Nations Millennium Development Goals. Retrieved from http://www.un.org/millenniumgoals/bkgd.shtml

United Nations (n.d.b). World Population Prospects 2017. Retrieved from https://esa.un.org/unpd/wpp/Download/Standard/Population/

United Nations (2015). Transforming our World: the 2030 Agenda for Sustainable Development. Retrieved from http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

Van de Gaer, E., Fraine, B. de, Pustjens, H., van Damme, J., Munter, A. D., & Onghena, P. (2009). School effects on the development of motivation toward learning tasks and the development of academic self-concept in secondary education: A multivariate latent growth curve approach. *School Effectiveness and School Improvement*, *20*(2), 235–253. https://doi.org/10.1080/09243450902883920

Van der Wal, M., & Waslander, S. (2007). Traditional and Non-Traditional Educational Outcomes: Trade-off or complementarity? *School Effectiveness and School Improvement*, *18*(4), 409–428. https://doi.org/10.1080/09243450701712502

Veenman, M. V. J., Van Hout-Wolters, Bernadette H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3–14. https://doi.org/10.1007/s11409-006-6893-0

Verhaeghe, J. P., van Damme, J., & Knipprath, H. (2011). *Differences in value added between primary schools with high proportions of minority students: a longitudinal study*. Presented at the meeting of the European Association for Research on Learning and Instruction, Leuven.

Walberg, H. J. (1971). Models for optimizing and individualizing school learning. *Interchange*, *2*(3), 15–27. https://doi.org/10.1007/BF02282467

Walberg, H. J. (1986). Syntheses of research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York, London: Macmillan; Collier Macmillan.

Walberg, H. J., Fraser, B. J., & Welch, W. W. (1986). A Test of a Model of Educational Productivity among Senior High School Students. *The Journal of Educational Research*, *79*(3), 133–139.

Wang, J. (1998). Opportunity to Learn: The Impacts and Policy Implications. *Educational Evaluation and Policy Analysis*, *20*(3), 137–156. https://doi.org/10.3102/01623737020003137

Wang, J., & Staver, J. R. (1996). An empirical approach toward the prediction of students' science achievement in the United States and Hubei, China. *Journal of Research in Science Teaching*, *33.* https://doi.org/10.1002/(SICI)1098-2736(199603)33:3<283::AID-TEA3>3.0.CO;2-P

Weber, A. S. (2011). The role of education in knowledge economies in developing countries. *Procedia - Social and Behavioral Sciences*, *15*, 2589–2594. https://doi.org/10.1016/j.sbspro.2011.04.151

Webster, B. J., & Fisher, D. L. (2000). Accounting for Variation in Science and Mathematics Achievement: A Multilevel Analysis of Australian Data. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, *11*(3), 339–360.

Wigfield, A., & Eccles, J. (2000). Expectancy-Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. Washington D.C.: Falmer.

Wimpelberg, R. K., Teddlie, C., & Stringfield, S. (1989). Sensitivity to Context: The Past and Future of Effective Schools Research. *Educational Administration Quarterly*, *25*(1), 82–107. https://doi.org/10.1177/0013161X89025001005

Wiseman, A., Al Sadaawi, A., & Alromi, N. H. (2008). Educational Indicators and National Development in Saudi Arabia.

Wiseman, A. W., Alromi, N. H., & Alshumrani, S. (2013). Science Education Impacts on Labor Market and University Expectations of Students by Citizenship Status in the Kingdom of Saudi Arabia: A Comparative Analysis Using TIMSS 2007 Data. *Citizenship, Social and Economics Education*, *12*(3), 216–229. https://doi.org/10.2304/csee.2013.12.3.216

Won, S. J., & Han, S. (2010). Out-of-School Activities and Achievement among Middle School Students in the U.S. and South Korea. *Journal of Advanced Academics*, *21*(4), 628–661. Retrieved from http://files.eric.ed.gov/fulltext/EJ906117.pdf

World Bank (n. d.). Percentage of enrolment in primary education in private institutions (%). Retrieved from https://data.worldbank.org/indicator/SE.PRM.PRIV.ZS

World Bank (2016). Labor force participation rate, female. Retrieved from https://data.worldbank.org/indicator/SL.TLF.CACT.FE.ZS?view=map

Wößmann, L. (2003). Schooling Resources, Educational Institutions and Student Performance: The International Evidence. *Oxford Bulletin of Economics and Statistics*, *65*(2), 117–170. https://doi.org/10.1111/1468-0084.00045

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achievement. Issues & Answers. REL 2007-No. 033. Regional Educational Laboratory Southwest (NJ1). Retrieved from http://files.eric.ed.gov/fulltext/ED498548.pdf

Yuan, Y. C. (2000). Multiple imputation for missing data: Concepts and new development. *Proceedings of the Twenty - Fifth Annual SAS Users Group International Conference (Paper No. 267). Cary, NC: SAS Institute*.

Zhu, M. (2014). Analyzing Multilevel Models with the GLIMMIX Procedure. *Paper SAS026-2014*, 1–18.

APPENDIX A : THE TIMSS 2015 QUESTIONNAIRES

The following section presents summarized information of the four TIMSS grade four questionnaires (Foy, 2017) which served as a basis for the variable categorization and subsequent analyses.

Student Questionnaire

Table A-1: Content of the TIMSS 2015 grade 4 student questionnaire

Question	Item Content	Description	Number of Items	Assigned to Model Factor	Factor - Details	Comments
G1	Gender	Student's gender	1	Student characteristics	Student background	
G2	Age	Student's birth date	2			
G3	Language of test	Frequency student speaks language of test at home	1			
G4	Books at home	Number of books at home	1	Student characteristics	Student background	
G5	Home possessions	Educational resources and items at home	6+5 country-specific	Student characteristics	Student background	
G6	Parents born in country	Father and mother born in country of test	2	Student characteristics	Student background	
G7	Child born in country	Child born in country of test	1	Student characteristics	Student background	
G8	Absence from school	Frequency of absence	1	Time on task		
G9	Breakfast on school days	Frequency of eating breakfast	1			
G10	Computer usage	Frequency of using computer at home, school, and other places	3			
G11	General attitude towards school	Thoughts about feeling save, belonging, being fairly treated, etc.	7			
G12	School safety	Experiences of problematic behavior by other students	8			
MS1	Liking mathematics	How much the student likes and enjoys mathematics	9	Student characteristics	Subject motivation	
MS2	Learning activities in mathematics	Student perception of mathematics instruction	10	Quality	Clear and structured teaching (2a,b,e,f,i)	
MS3	Confidence in mathematics	How confident students feel with mathematics	9			
MS4	Liking science	How much the student likes and enjoys science	9	Student characteristics	Subject motivation	
MS5	Learning activities in science	Student perception of science instruction	10	Quality	Clear and structured teaching (5a,b,e,f,i)	
MS6	Confidence in science	How confident students feel with science	7			

Note. Column "Question": G = general question/ M = mathematics question/ S = science question

Early Learning Survey (Parent questionnaire)

Table A-2: Content of the TIMSS 2015 grade 4 parent questionnaire

Question	Item Content	Description	Number of Items	Assigned to Model Factor	Factor - Details	Comments
1	Completion of questionnaire	Person completing the questionnaire	3			
2	Early numeracy activities	Activities undertaken by parents in regard to learning before primary education	16			
3	Child born in country	Child born in country of test	A1/B1	Student characteristics	Student background	
4	Languages spoken at home	Language spoken at home before child went to school	6 country-specific options			
5	Attendance of pre-primary education	Educational program attended and duration	A2/B1			
6	Age of school entry	How old was the child when beginning primary education	1			
7	Student's reading abilities before school	Letter, word recognition and reading abilities before primary	6			
8	Student's numeracy abilities before school	Frequency of absence (a-c)	7	Student characteristics	Proxy for Aptitude	
9	Homework	Frequency of homework and assistance by parents	A1/B3	Opportunity (9bb)		
10	Extra lessons or tutoring	Extra lessons or tutoring in the last year in math and science and duration	A2/B2			Not used as it has diff. meaning for diff. students
11	Parents view of child's school	Parents thoughts about the contribution to their child's academic succes.	8			
12	Time spent for reading	Weekly reading activities of parents	1			
13	Books at home	Number of books at home	1	Student characteristics	Student background	
14	Children's books at home	Number of children's books at home	1			
15	Digital information devices at home	Number of digital devices at home	1			
16	Valuing science and mathematics	In how far parents value mathematics and science	8			
17	Parents born in country	Parents born in country of test	A1/B1	Student characteristics	Student background	
18	Language spoken at home	Language used at home by father and mother	A1/B1			
19	Language of test	Frequency of the language of test spoken at home by the child	1			
20	Highest level of parental education	Highest level of education completed by father and mother	A1/B1	Student characteristics	Student background	
21	Educational expectations	Educational expectations of parents for their child	1			
22	Parental employment situation	Employment situation for father and mother	A1/B1			
23	Parental occupation	Father's and mother's occupation	A1/B1	Student characteristics	Student background	

Notes. Column "Question": G = general question/ M = mathematics question/ S = science question
Column "Number of Items": The number of items is given for each item part separately (A2 for example denotes two items for part A of the question)

Teacher questionnaire

Table A-3: Content of the TIMSS 2015 grade 4 teacher questionnaire

Question	Item Content	Description	Number of Items	Assigned to Model Factor	Factor - Details	Comments
G1	Teaching experience	Numbers of years as a teacher	1	Input	Teacher Background	
G2	Gender	Teacher's gender	1	Input	Teacher Background	
G3	Age	Teacher's age	1			
G4	Formal education	Teacher's highest education level	1	Input	Teacher Background	
G5	Major area of study	Teacher's main area of study and specialization	A6/B4	Input	Teacher Background	
G6	Emphasis on academic success	Teacher's perception of items related to emphasis on academic success and parental involvement	17	Quality	Climate	
G7	Save and orderly school environment	Teacher's perception of the school environment	8	Quality (d-h)	Climate	
G8	Shortage of resources	Severity in terms of shortages of basic facilities and resources	7			Model regards ed. resources only on school level
G9	Collaboration among teachers	Interaction and collaboration with colleagues	7			
G10	Job satisfaction	Teacher's job satisfaction	A2/B2			
G11	Problems faced by teachers	Problems faced in terms of time pressure, etc.	8			
G12	Number of students per class	Total number and fourth graders in class	A1/B1			
G13	Difficulties in the language of test	Number of students facing difficulties understanding spoken language of test	1			
G14	Teaching activities	Frequency of a range of diverse teaching activities	8	Quality	Cognitive activation	
G15	Limitation of teaching	Student characteristics limiting teaching the class	7	Quality (15d)	Classroom managment	
M1	Mathematics teaching time	Minutes per week of mathematics teaching	1	Time		
M2	Confidence in teaching mathematics	Teacher's confidence in general competencies needed	9	Input	Teacher Background	
M3	Teaching activities in math lessons	Teacher's judgement of different activities used in mathematics	9			
M4	Use of calculators	Restriction of the use of calculators	1			
M5	Use of computers	Availability, access, and use of computers in mathematics	A1/B3/C3			
M6	Mathematic topics taught	Asks in how far main topics of the TIMSS test already have been taught	A8/B7/C2	Opportunity		
M7	Homework assignment	Frequency, duration, and control of homework assignment	A1/B1/C3	Quality (7ca-cc)/Time (7a/b)	Assessment	
M8	Monitoring progress	Emphasis on monitoring student's progress	3	Quality (8a)	Assessment	
M9	Professional development	Areas of professional development	7			
M10	Time spent in professional development	Total amount of hours spent for development in the last two years	1	Time		
M11	Preparedness to teach mathematics	Preparedness for different content domains	A8/B7/C2	Input	Characteristics	
S1	Type of science teaching and teaching time	Separate or integrated science teaching and minutes of science teaching per week	A1/B1	Time		
S2	Confidence in teaching science	Teacher's confidence in general competencies needed	9	Input	Characteristics	
S3	Teaching activities in science lessons	Teacher's judgement of different activities used in science	14			Options related to the concept of cognitive activation were not included for the sake of comparability with the math analyses
S4	Use of computers	Availability, access, and use of computers in science	A1/B3/C4			
S5	Science topics taught	Asks in how far main topics of the TIMSS test already have been taught	A6/B9/C7	Opportunity		
S6	Homework assignment	Frequency, duration, and control of homework assignment	A1/B1/C3	Quality (6ca-cc)/Time (7a/b)	Assessment	
S7	Monitoring progress	Emphasis on monitoring student's progress	3	Quality	Assessment	
S8	Professional development	Areas of professional development	7			
S9	Time spent in professional development	Total amount of hours spent for development in the last two years	1	Time		
S10	Preparedness to teach mathematics	Preparedness for different content domains	A7/B9/C7	Input	Teacher Background	

Notes. Column "Question": G = general question/ M = mathematics question/ S = science question
Column "Number of Items": The number of items is given for each item part separately (A2 for example denotes two items for part A of the question)

School Questionnaire

Table A-4: Content of the TIMSS 2015 grade 4 school questionnaire

Question	Item Content	Description	Number of Items	Assigned to Model Factor	Factor - Details	Comments
1	Enrollment	Total student enrollment	1			
2	4th grade enrollment	Enrollment of 4th graders	1			
3	Student's background	Percentage of students from economically disadvantaged or affluent homes	A1/B1			
4	Native language of students	Percentage of students whose native language is the language of test	1			
5	Community characteristics	Community size and location where the school is located	A1/B1			
6	Provision of free meals	Provision of free breakfast and lunch by school	A1/B1			
7	Emphasis on health topics	Emphasis on different health topics by school	4			
8	Instructional time	Number of days per year, school days per week, and length of typical school day	A1/B1/C1	Time		
9	Space and assistance for school work	Provision of space and assistants for schoolwork before/after school	A1/B1			
10	Policies in terms of tracking/streaming	School policies related to tracking/streaming in mathematics and science	A1/B1	Opportunity		
11	Number of computers	Number of computers available for students	1	Input	Resources	
12	Availability of science laboratory and assistants	Schools' availability of a laboratory and assistants to help with experiments	A1/B1			Not used for the sake of comparable math and science models
13	School library	Availability of school library and number of books and periodicals	A1/B1	Input	Resources	
14	Shortage in school resources	Affected by shortage in general, mathematics, and science related school resources	A6/B5/C4	Input	Resources	
15	Emphasis on academic success	Principal's perspective on emphasis on academic success and parental support	13	Quality (15a-e,k-m)	Environment (SLE)	Parent-related items not used (not part of the theoretical framework)
16	School discipline and safety	Problems related to school discipline and safety	10	Time (16a-b)/Quality (16d-j)	Environment (SLE)	
17	Absenteeism	Problems with absenteeism and late arrival	2	Time (incl. 16a-b)		
18	School readiness	Percentage of students with certain literacy and numeracy skills	11			
19	Experience as a principal	Principal's years of experience	1			
20	Experience in the selected school	Principal's years of experience in the selected school	1			
21	Formal education	Principal's highest level of education	1			
22	Degrees in leadership	Degrees in educational leadership held by the principal	2			

Notes. Column "Question": G = general question/ M = mathematics question/ S = science question
Column "Number of Items": The number of items is given for each item part separately (A2 for example denotes two items for part A of the question)

APPENDIX B : INDICATORS AND VARIABLE RECODING

Creation of indicators: The following indicators have been created during data preparation:

Nationality of the Student

In the GCC countries, children born to a father with citizenship status are citizens of that country, irrespective of their place of birth. In certain countries, children have to apply for citizenship at the age of 18 if their mother is born in the respective country but not the father. In Bahrain, Oman, Saudi Arabia, and the United Arab Emirates children born to a stateless father and a mother of the respective country are also automatically citizens. Foreigners may be granted citizenship by 'naturalization', but requirements are stringent (e.g., residing in the country for more than 25 years, as is the case in Qatar and other countries) and citizenship is only rarely granted. While the international TIMSS questionnaires ask the parents and child where they are born, the data do not contain any information on whether the child is legally considered to be a *national* (a resident of the respective Gulf State). This question was only asked as a national question in the United Arab Emirates.

The national data from the United Arab Emirates were compared with different definitions about nationality and also only considering the variable *father born in country*. The minimum deviation from the related national indicator variable in the United Arab Emirates was obtained using the PISA definition of *immigration* (in PISA, national students are defined as those having at least one parent born in the country; OECD, 2014b, p. 307) and with the *father born in country* variable on its own. In both cases, about 12.3% of the answers did not match the student's answer of the national question. As more valid data were available when only the variable *father born in country* was used and as this definition is more applicable to the nationality laws in the region, it was decided to use *father was born in country* as nationality indicator for the current study. This question was asked in both the home and the student questionnaires. Data from the home questionnaire was assumed to be more valid, but had relatively high missing rates. The final indicator, therefore, was created by replacing missing parent answers by valid student answers, if available. This procedure, for example in the United Arab Emirates, reduced the missing rate from over 12% to less than 3%.

In summary, nationality of the student was derived from a combination of variable ASBH17A (Was the child's father born in country) of the home questionnaire and ASBG06B (Was your father born in country) from the student's questionnaire. Valid values from ASBG06B were used to reduce the missing rate of the default variable ASBH17.

National = 0: National student

National = 1: Non-national student

Number of Books RBOOKS (HQ-13/SQ-4)

The number of books were based on variable ASBH13 from the home questionnaire but in case of missing values complemented by variable ASBG04 from the student questionnaire.

Preparedness to teach: ITBM11Z/ITBS10Z (TQ-M11/TQ-S10)

The preparedness to teach for math and science was calculated as the average preparedness over all items. The categories “not-applicable” and “not well prepared” were coded as “0” for each item, while “somewhat prepared” was coded “1” and “very well prepared” was coded to “2”. A maximum of two missing options were allowed.

Emphasis on Monitoring: ITBM08Z/ITBS07Z (TQ-M08/S07)

After reverse-recoding and setting the lowest category to “0”, the different sources to monitoring students’ progress were added up. Thus, the index can take values between 0 and 6. All answers must have valid values.

Homework time: ITDM07Z/ITDS06S (TQ-M07/S06)

The overall time spent on homework was calculated by multiplying the number of assignments (ATBM07A/ATBS06A) with the average assigned length (ATBM07B/ATBS06B). As the minutes per assignment were categorized, the average value of each interval was used to determine the average duration of the homework assignments.

Verification of homework assignment: ITBM07Z/ITBS06Z (TQ-M07/S06)

After reverse-recoding and setting the lowest category to “0”, the frequency of different activities related to monitoring students’ homework were added up. Hence, the index can take values between 0 and 6. All answers must have valid values.

Topics covered: ITDM06Z/ITDS05Z (TQ-M06/S05)

The index for topics already taught was calculated as the average over all items. The category “not yet taught or just introduced” was coded to “0” for each item, while the categories

"mostly taught this year" and "mostly taught before this year" were coded to "1". A maximum of two missing options were allowed. For an easier interpretation of the multilevel results the obtained averages were multiplied by 10. One unit then can be interpreted as 10% of the topics covered.

Number of books in the school library: ICBG13A (SCQ-13/13A)

Based on question 13 of the school questionnaire asking about the availability of a school library and question 13A asking about the number of books with different titles available in the library, an index of the total number of print books available in the school library was created according to the following recoding rules: If the availability of a library was answered with "no", the created index was set to zero. If a library was available, the index was set to the average of the ranges of print books listed for each option in question 13A. For better interpretation of the association with achievement, values were divided by 100 which mean that one unit represents 100 books.

Reliability Analyses for indices

All indices consisting of more than two variables were submitted to a reliability analysis as described in section 8.3.6.2. Table B-1 shows the internal scale consistency measured by Cronbach's alpha. The reliability for all indices except ITBM07Z & ITBM06Z can be judged as at least "acceptable". As conceptually the homework verification style was regarded as an important indicator for the model factor assessment, it was nevertheless decided to keep the index even with a "poor" internal consistency.

Table B-1: Results from reliability analyses for the created indices

Index	Description	Cronbach's Alpha	Number of Items
ITBM11Z	Preparedness to teach (M)	0.90	17
ITBS10Z	Preparedness to teach (S)	0.95	23
ITBM07Z	Verification of homework completion (M)	0.56	3
ITBS06Z	Verification of homework completion (S)	0.64	3
ITBM06Z	Amount of topics covered (M)	0.78	17
ITBS05Z	Amount of topics covered (S)	0.83	23

Further recodings of variables as listed in Table B-2 were performed to allow for better comparability and/or better interpretation of the final results.

Table B-2: Further recodings

Question Location	Label	Original values	Recoded	Comment
TQM04	Teachers highest education level	1 (Did not complete ISCED 3) … 7 (doctor or equivalent)	6 yrs … 21 yrs	Recoded to years of schooling, according to the recoding of parental education
TQ-M01/ TQ-S01B	Time spent on teaching	Minutes per week	Hours per week	Divided by 60
TQ-M10/ TQ-S09	Professional development	1 (none) 2 (< 6 hrs) 3 (6-15 hrs) 4 (16-35 hrs) 5 (> 35 hrs)	0 0.38 1.31 3.19 5.63	Class means of 0/3/ 10.5/25.5/45 hrs devided by 8 to be interpreted as working days
SQ-8	Absenteeism	1 (>= once a week) 2 (every two weeks) 3 (once a month) 4 (never or almost never)	4 2 1 0	Converted into absences per month
HQ-9BB	Help with homework	1 (every day) 2 (3-4 times a week) 3 (1-2 times a week) 4 (less than once a week) 5 (never or almost never)	5 3.5 1.5 0.5 0	Times per week

APPENDIX C : ADDITIONAL ANALYSES

Non-nationals/Nationals by highest education level of their parents:

The following tables list achievement results by highest educational level of their parents (ASDHEDUP). Results for the category "I don't know" were not included into the analyses presented here. Results marked bold are significant.

Results show that non-national students in all countries, except in Bahrain and Oman, achieve higher results than nationals in both mathematics and science on each of the educational levels. However, standard errors are sometimes relatively high due to low number of students in some of the cells and in consequence, while there are often relatively high differences, not all of them are statistically significant.

Table C-1: Mathematics results by parental education level

Country	Some Primary or lower		Lower Secondary		Upper Secondary		Post-Secondary		University or higher	
	Nationals	Non-Nationals	Nationals	Non-Nationals	Nationals	Non-Nationals	Nationals	Non-Nationals	Nationals	Non-Nationals
Bahrain	409 (7.1)	398 (9.5)	404 (6.5)	408 (7.6)	437 (2.2)	439 (5.3)	455 (3.1)	450 (9.1)	474 (2.6)	**490** (4.1)
Kuwait	272 (16.3)	**362** (27.9)	292 (6.8)	**346** (16.1)	315 (5.8)	**379** (12.9)	330 (4.1)	**391** (10.1)	365 (5.0)	**420** (6.8)
Oman	395 (3.8)	389 (15.6)	417 (3.9)	**392** (10.7)	431 (3.9)	**405** (12.2)	453 (5.7)	**420** (12.2)	471 (3.7)	**438** (6.5)
Qatar	374 (17.8)	402 (15.1)	371 (12.8)	**413** (10.5)	374 (6.1)	**424** (6.5)	398 (9.1)	**462** (6.4)	418 (5.1)	**490** (4.5)
Saudi Arabia	365 (7.2)	374 (15.6)	371 (9.3)	387 (19.3)	375 (5.9)	393 (9.6)	394 (9.6)	422 (12.4)	391 (4.9)	**448** (7.6)
United Arab Emirates	353 (5.1)	**388** (9.8)	368 (4.8)	**411** (8.1)	385 (4.1)	**424** (4.8)	402 (4.4)	**467** (4.5)	429 (3.9)	**508** (2.6)
Gulf Average	**361 (11.0)**	**386 (16.7)**	**371 (8.0)**	**393 (12.8)**	**386 (4.9)**	**411 (9.1)**	**405 (6.5)**	**436 (9.6)**	**425 (4.3)**	**466 (5.6)**

Note. Significant differences between nationals and non-nationals (0.05 level [2-tailed]) are marked in bold

Table C-2: Science results by parental education level

Country	Some Primary or lower		Lower Secondary		Upper Secondary		Post-Secondary		University or higher	
	Nationals	Non-Nationals	Nationals	Non-Nationals	Nationals	Non-Nationals	Nationals	Non-Nationals	Nationals	Non-Nationals
Bahrain	419 (12.1)	393 (15.3)	412 (10.3)	390 (18.3)	447 (4.1)	452 (8.6)	471 (5.0)	460 (18.8)	478 (5.1)	**509** (5.9)
Kuwait	250 (22.4)	**339** (38.5)	280 (12.0)	**310** (29.2)	294 (8.4)	**340** (13.8)	314 (7.0)	**400** (14.1)	357 (7.6)	**413** (10.4)
Oman	393 (5.0)	388 (18.2)	424 (5.0)	**387** (11.4)	438 (4.8)	**409** (12.9)	464 (6.2)	**435** (11.9)	485 (4.1)	**454** (6.0)
Qatar	364 (18.1)	391 (15.5)	358 (14.2)	**405** (13.9)	374 (7.0)	**418** (9.6)	393 (10.4)	**467** (7.8)	418 (5.1)	**494** (4.4)
Saudi Arabia	358 (9.4)	391 (16.3)	362 (9.4)	389 (21.7)	385 (6.7)	406 (14.2)	403 (7.6)	431 (12.6)	405 (5.8)	**464** (7.7)
United Arab Emirates	334 (6.5)	**379** (11.1)	354 (5.4)	**399** (9.3)	370 (4.6)	**426** (5.1)	391 (5.2)	**476** (4.5)	421 (4.6)	**522** (2.5)
Gulf Average	**353 (13.7)**	**380 (21.1)**	**365 (10.0)**	**380 (18.6)**	**385 (6.1)**	**409 (11.2)**	**406 (7.1)**	**445 (12.5)**	**427 (5.5)**	**476 (6.6)**

Note. Significant differences between nationals and non-nationals (0.05 level [2-tailed]) are marked in bold

APPENDIX D : VARIANCE COMPONENTS FOR GROUP-CENTERED APPROACH

Explained variance components for the different multilevel models using a group-centered approach for all level 1 predictors:

Table D-1: Variance components for the mathematics models

Country	% of variance in mathematics achievement that is between courses	% of between-course variance attributable to level 1 + 2 home background	% of between-course variance attributable to home background and full model
Bahrain (BHR)	24	38	59
Kuwait (KWT)	30	54	57
Oman (OMN)	28	24	41
Qatar (QAT)	41	64	74
Saudi Arabia (SAU)	36	16	32
United Arab Emirates (ARE)	59	63	72

Table D-2: Variance components for the science models

Country	% of variance in mathematics achievement that is between courses	% of between-course variance attributable to level 1 + 2 home background	% of between-course variance attributable to home background and full model
Bahrain (BHR)	29	35	45
Kuwait (KWT)	32	55	63
Oman (OMN)	29	23	38
Qatar (QAT)	35	58	73
Saudi Arabia (SAU)	34	39	49
United Arab Emirates (ARE)	55	68	75

Eidesstattliche Erklärung

Ich versichere, dass ich die Arbeit selbständig angefertigt, nicht anderweitig für Prüfungszwecke vorgelegt, alle benutzten Quellen und Hilfsmittel angegeben, sowie wörtliche und sinngemäße Zitate gekennzeichnet habe. Ich habe keine kommerzielle Promotionsberatung in Anspruch genommen.

Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

---------- Oliver Neuschmidt

Ort, Datum Unterschrift