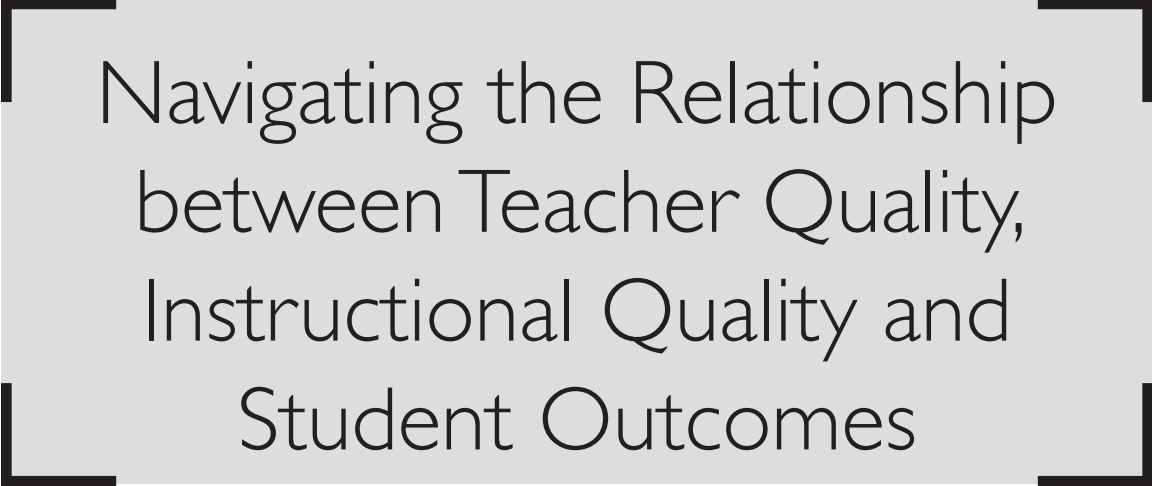


Navigating the Relationship Between Teacher Quality, Instructional Quality and Student Outcomes

Lokendra Chakrabarti



Navigating the Relationship
between Teacher Quality,
Instructional Quality and
Student Outcomes

Navigating the Relationship between Teacher Quality, Instructional Quality and Student Outcomes

Edited by
Lokendra Chakrabarti



Published by Vidya Books,
305, Ajit Bhawan,
21 Ansari Road,
Daryaganj, Delhi 110002

Edited by Lokendra Chakrabarti
ISBN: 978-93-5431-683-8

© 2022 Vidya Books

This book contains information obtained from authentic and highly regarded sources. All chapters are published with permission under the Creative Commons Attribution Share Alike License or equivalent. A wide variety of references are listed. Permissions and sources are indicated; for detailed attributions, please refer to the permissions page. Reasonable efforts have been made to publish reliable data and information, but the authors, editors and publisher cannot assume any responsibility for the validity of all materials or the consequences of their use.

Trademark Notice: All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy. Furthermore, the publisher ensures that the text paper and cover boards used have met acceptable environmental accreditation standards.

Table of Contents

Chapter 1	Conceptual Framework and Methodology of this Report	1
Chapter 2	Relation of Student Achievement to the Quality of their Teachers and Instructional Quality	20
Chapter 3	The Relations Among School Climate, Instructional Quality, and Achievement Motivation in Mathematics	50
Chapter 4	The Impact of School Climate and Teacher Quality on Mathematics Achievement: A Difference-in-Differences Approach	80
Chapter 5	The Importance of Instructional Quality for the Relation Between Achievement in Reading and Mathematics	95
Chapter 6	The Relation Between Students' Perceptions of Instructional Quality and Bullying Victimization	112
Chapter 7	Final Remarks	131



Conceptual Framework and Methodology of This Report

Trude Nilsen, Jan-Eric Gustafsson and Sigrid Blömeke

Abstract In this volume, five separate studies examine differing aspects of relations between teacher quality, instructional quality and learning outcomes across countries, taking into account context characteristics such as school climate. The 2007 and 2011 TIMSS (Trends in Mathematics and Science Study) cycles provided the research data. These five studies cover grade four or grade eight students and their teachers, including cognitive or affective-motivational learning outcomes. This introductory chapter describes the overall conceptual framework and the research questions posed by each chapter, and outlines the general design features of TIMSS. Key constructs, and common methodological issues among the five studies are discussed, and this introduction concludes with an overview of all chapters.

Keywords Instructional quality · Teacher quality · Student outcome · Theoretical framework · Trends in Mathematics and Science Study (TIMSS)

1.1 Introduction

Researchers and practitioners have long known that the quality of teachers and the quality of their instruction are key determinants of student learning outcomes (Klieme et al. 2009; Seidel and Shavelson 2007). However, the relationships have

T. Nilsen (✉)

Department of Teacher Education and School Research, University of Oslo,
Oslo, Norway

e-mail: trude.nilsen@ils.uio.no

J.-E. Gustafsson

Department of Education and Special Education, University of Gothenburg,
Gothenburg, Sweden

e-mail: jan-eric.gustafsson@ped.gu.se

J.-E. Gustafsson · S. Blömeke

Faculty of Educational Sciences, Centre for Educational Measurement at the
University of Oslo (CEMO), Oslo, Norway

e-mail: sigrid.blomeke@cemo.uio.no

often been difficult to quantify and understand empirically. Reviews of previous research have pointed to challenges in measuring teacher and instructional quality (Schlesinger and Jentsch 2016; Kunter et al. 2013). Moreover, the impact of student background often swamps the effects of the other variables, rendering them less visible. Finally, due to teacher selection and rules of certification, these variables often vary only little within a school system, making it difficult to identify effects.

Advancements in psychometrics and quantitative methods, along with the establishment of international large-scale assessments (ILSA), offer researchers new opportunities to study relations between teachers, their instruction and learning outcomes (Chapman et al. 2012). For instance, ILSA data provide the opportunity for multi-level analysis, standardized definitions of variables, trend design and representative samples from a large number of educational systems, in the following also called countries. Perhaps the best known ILSAs are the International Association for the Evaluation of Educational Achievement (IEA) Trends in Mathematics and Science Study (TIMSS), and the Organisation for Economic Cooperation and Development (OECD) Programme for International Student Assessment (PISA) and Teaching and Learning International Survey (TALIS). Out of these, TIMSS is the only one that provides data on the student, class and school levels. TIMSS therefore provides data well suited for an examination of relations between teacher quality, instructional quality and student outcomes across cohorts, time, and countries from all continents.

Using the world as a global educational laboratory may contribute toward an international understanding of teacher quality and instructional quality, and establish their importance for student learning outcomes across and within countries and over time. This demands research that takes into account: (1) the complexity of educational systems with many hierarchical layers and interwoven relationships (Scheerens and Bosker 1997); (2) the complexity of relationships within each layer with direct and indirect effects; (3) the variation of these relationships across countries; and (4) their development over time. Since it is difficult to take all these complexities into account within one study, combining results from different studies investigating subsets of relations may currently be the best way to make progress.

This book presents five studies which have been undertaken in this spirit. The studies complement each other to address the complexities mentioned above. The studies examined the following research questions:

- (1) *Which relations exist between teacher quality, instructional quality and mathematics achievement in grade four across and within countries, and is it possible to identify larger world regions or clusters of countries where similar relational patterns exist? (Chap. 2)*
- (2) *Which relations exist between school climate, instructional quality, and achievement motivation in mathematics in grade eight across and within*

- countries, and is it possible to identify larger world regions or clusters of countries where similar relational patterns exist? (Chap. 3)*
- (3) *To what extent can a causal influence of school climate and teacher quality on mathematics achievement in grade eight be identified in country-level longitudinal analyses? (Chap. 4)*
 - (4) *Which relations exist between instructional quality and reading, and between instructional quality and mathematics achievement in grade four, and to what extent does instructional quality moderate the relations between reading and mathematics achievement? (Chap. 5)*
 - (5) *Which relations exist between bullying and instructional quality in grade four across countries and within countries? (Chap. 6)*

The last chapter of this book summarizes the results obtained in these five studies and discusses conceptual and methodological challenges, as well as possible improvements in both research and practice. In taking this approach, our aim is to contribute to educational effectiveness research, to educational policy and practice, and to the field of educational measurement.

1.2 Conceptual Framework

Our research is situated within the field of educational effectiveness research, and this field has made great progress over the last three decades. This is partly because certain limitations of previous studies have been amended (Creemers and Kyriakides 2008; Chapman et al. 2012). These limitations included models which could only partially account for the nested nature of data, non-random samples, cross-sectional designs, or non-robust software. However, while there were methodological advances within the field of educational effectiveness, Creemers and Kyriakides (2006, p. 348) argued that there was also a need for “rational models from which researchers can build theory.” Over the years, they developed and tested a model for educational effectiveness, which they called the dynamic model of educational effectiveness. This model takes into account the complexity of educational systems, where students are nested within classes that are nested within schools, where variables within and across these levels can be directly and indirectly related, and where changes occur. This model also accounts for a national context level, which refers to the educational system at large, including the educational policy at the regional and/or national level, which should be examined in comparative studies (Kyriakides 2006). The model is well recognized internationally (Sammons 2009).

In this book, a conceptual framework (Fig. 1.1) is used that starts with the dynamic model of educational effectiveness (Creemers and Kyriakides 2008) and operationalizes it with respect to the research questions of this report. In line with

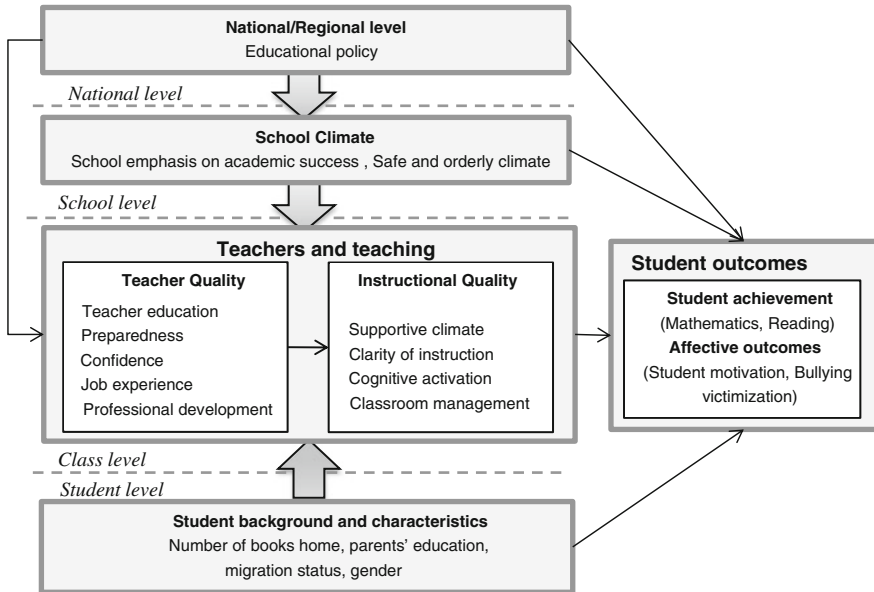


Fig. 1.1 Conceptual framework of determinants of student outcomes examined in this book

Kyriakides et al. (2009) and other studies (for example Baumert et al. 2010; Kane and Cantrell 2010), teacher and teaching variables at the class level are hypothesized to be most important for student learning. The conceptual framework focuses on relations between the national, school, class, and student level. The model shows how the national level is hypothesized to influence the school and teacher levels, as well as student outcomes in the five studies of this report. These relations may be both direct and indirect. Because of differences between educational systems, including different cultural contexts, educational values, educational policies, and structural features of the school system, we hypothesize that the relations of the indicators examined at lower levels, such as schools, classes and students, vary substantially within countries. Based on existing research, we also hypothesize that patterns exist that reflect similarities between *groups* of countries, due to similarities in culture, values, policies or school structure (see for example Blömeke et al. 2013).

School level variables are hypothesized to influence the class and student level (Fig. 1.1). In this book, we examine the school features School emphasis on academic success and Safe and orderly climate. The class level contains two important variables for learning outcomes, namely teacher quality and instructional quality. These constructs are also hypothesized to be interrelated (Fig. 1.1). Finally, in line with existing research (Gustafsson et al. 2013; Hansen and Munk 2012) student characteristics (such as gender and minority status) and home background (for example, parents' education) are hypothesized to be related to student outcomes. Such outcomes may be cognitive or affective.

1.3 Operationalization of School-, Class- and Student-Level Features

This section presents a brief outline of how crucial constructs were operationalized. A detailed presentation is provided in the following chapters.

1.3.1 *Teacher Quality*

Goe (2007) presented a framework for understanding the key components of teacher quality and their relations to student learning outcomes. According to this framework, teacher quality includes both teacher qualifications and characteristics (*inputs*) that influence teachers' instruction (*process*) and student *outcomes* (e.g., achievement and motivation). In this book, teacher quality is operationalized via qualifications such as teacher education level, job experience and participation in professional development activities, as well as by teacher characteristics such as self-efficacy. The Teacher Education and Development Study in Mathematics (TEDS-M) was the first international large-scale assessment that examined these features, with representative samples from a broad range of countries (see for example Blömeke et al. 2011; Tatto et al. 2012). In mathematics, teacher quality has been shown to be of importance for student achievement in a number of within-country studies (Baumert et al. 2010; Blömeke and Delaney 2014). A substantial research gap exists with respect to non-Western countries and comparative research across countries applying the same kind of instruments. This book intends to narrow this research gap.

1.3.2 *Instructional Quality*

Instructional quality is a construct that reflects those features of teachers' instructional practices well known to be positively related to student outcomes, both cognitive and affective ones (Decristan et al. 2015; Fauth et al. 2014; Good et al. 2009; Hattie 2009; Klusmann et al. 2008; Seidel and Shavelson 2007). The construct is understood and operationalized differently across the field but its multidimensionality was revealed in major research projects originating in both Europe (Baumert et al. 2010; Kunter et al. 2008) and the United States (Ferguson 2010; Kane and Cantrell 2012). As with teacher quality, a research gap exists with respect to non-Western countries and calls for comparative research across countries.

The operationalization of instructional quality used in this book is mainly based on the model of three "global dimensions of classroom process quality" (Klieme et al. 2001; Klieme and Rakoczy 2003; Lipowsky et al. 2009). Klieme and colleagues' model was developed based on data from the German extension to TIMSS

Video and subsequently applied to data from PISA 2000; its dimensions include cognitive activation, supportive climate, and classroom management. This model is similar to studies carried out independently in the USA (Kane and Cantrell 2012; Pianta and Hamre 2009; Reyes et al. 2012).

Cognitive activation refers to teachers' ability to challenge students cognitively, and comprises instructional activities in which students have to evaluate, integrate, and apply knowledge in the context of problem solving (Baumert et al. 2010; Fauth et al. 2014; Klieme et al. 2009). Supportive climate is a dimension that refers to classrooms where teachers provide extra help when needed, listen to and respect students' ideas and questions, and care about and encourage the students (Kane and Cantrell 2012; Klieme et al. 2009). Supportive climate may include clear and comprehensive instruction, clear learning goals, connecting new and old topics, and summarizing at the end of the lesson, but some research shows that supportive climate should be discriminated from clarity of instruction (Kane and Cantrell 2010). We therefore consider clarity of instruction as a fourth dimension of instructional quality.

1.3.3 School Climate

While teacher quality and instructional quality may directly influence students' learning and motivation, school climate creates the foundation for instruction and may hence influence learning both directly and indirectly (Kyriakides et al. 2010; Thapa et al. 2013; Wang and Degol 2015; see Fig. 1.1). In a recent review of school climate across several fields, Wang and Degol (2015) observed that school climate is defined differently across studies, but that certain aspects may be key. There seems to be broad consensus that academic climate and a safe and orderly climate are such key aspects and that they are positively related to learning outcomes (Bryk and Schneider 2002; Hoy et al. 2006; Thapa et al. 2013).

Academic climate focuses on the overall quality of the academic atmosphere; the priority and ambition for learning and success (Hoy et al. 2006; Martin et al. 2013; Nilsen and Gustafsson 2014; Wang and Degol 2015). School emphasis on academic success (SEAS) is therefore examined as an indicator of academic climate in this book. SEAS reflects a school's ambition and priority for learning and success. It has been shown to be related to students' learning in a number of countries (Martin et al. 2013; Nilsen and Gustafsson 2014). A second variable examined in this book is a safe and orderly climate, which refers to the degree of physical and emotional security provided by the school, as well as to an orderly climate with disciplinary practices (Goldstein et al. 2008; Gregory et al. 2012; Wang and Degol 2015). Studies have revealed that this variable is also related to student learning outcomes.



1.3.4 *Student Outcomes*

Throughout this book, different types of student outcomes are taken into account to address the multidimensionality of educational objectives of schooling. The main emphasis is on student achievement in mathematics at grade four and eight, but reading achievement using the IEA's Progress in Reading and Literacy Study (PIRLS) data, as well as student motivation and bullying victimization are also examined.

Cognitive outcomes in mathematics and reading

In grade four, students are assessed in TIMSS in the domains Number, Geometric Shapes and Measures, and Data Display, and in grade eight in Number, Algebra, Geometry, and Data and Chance. In addition to covering these content domains, the items also cover the cognitive demands Knowing, Applying and Reasoning (Mullis et al. 2012a). According to Niss (2003), mathematical competence “means the ability to understand, judge, do, and use mathematics in a variety of intra- and extra-mathematical contexts and situations in which mathematics plays or could play a role” (p. 6). In other words, students do not just need knowledge in mathematics, but must also be able to apply knowledge and conceptual understanding in different contexts, and to analyze, and reason to solve problems. The TIMSS framework reflects this notion fairly well (Mullis et al. 2012b) and is also in line with a number of other frameworks in mathematics (e.g. Kilpatrick 2014; Schoenfeld and Kilpatrick 2008).

TIMSS does not capture every aspect of mathematical competence. According to Niss (2003), mathematical competence includes eight different competencies that, for instance, involve mathematical theory like using and understanding theorems, communication in mathematics, handling symbols, including manipulating equations, and making use of aids and tools (including information technology). Although there are some items that reflect such aspects, extra-mathematical contexts and students' communication in mathematics are not measured extensively in TIMSS. In contrast, TIMSS does measure to some extent mathematical theory like using and understanding theorems, and students' ability to handle symbols, including manipulating equations (Hole et al. 2015). Moreover, TIMSS is based on the cores of the curricula of all countries participating, and it includes crucial cognitive demands such as knowing, applying and reasoning. Thus, TIMSS measures the key competencies in mathematics described by Niss (2003) to a satisfying degree.

In Chap. 5 of this book, reading achievement is included as well as mathematics achievement because reading literacy is regarded to be the foundation of most learning processes and an important ability students need to acquire during schooling. The data come from TIMSS and PIRLS 2011, where reading is defined as “the ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers in

school and everyday life, and for enjoyment” (Mullis et al. 2009). This definition has changed over study cycles, but is a good reflection of recent theories of reading literacy (Alexander and Jetton 2000; Ruddell and Unrau 2004; for more details, see Chap. 5).

1.3.5 Student Affective Outcomes

In addition to achievement, a number of studies also include interest, motivation, and self-beliefs as student outcomes (Bandura 1997; Eccles and Wigfield 2002). These constructs reflect students’ motivational states (see Chap. 3 for more theory on this). A substantial research gap exists with respect to studies in which school-, teacher- and class-level features are related to affective student outcomes in Western and non-Western countries, as well as with respect to comparative research across countries applying the same set of instruments. This book intends to reduce this research gap.

Given that learning takes place in social settings (i.e., in classrooms and schools), social interaction with peers must also be taken into account in considering student outcomes and their determinants. One of the constructs reflecting the results of such interactions refers to bullying victimization, which has been shown to be linked with achievement and motivation (Engel et al. 2009; Skues et al. 2005) and has been found to be related to classroom and school factors such as discipline, teacher support, instructional quality and school climate within several countries (Kyriakides et al. 2014; Murray-Harvey and Slee 2010; Richard et al. 2012). This aspect of research is progressed in this book using a comparative approach applied across a large range of countries.

1.4 TIMSS Design

TIMSS is an international large-scale survey of student achievement in mathematics and science. First conducted in 1995, TIMSS assesses students in grade four and eight every fourth year. Most chapters in this book draw on the 2011 TIMSS data, which included over 60 countries. All chapters considered as many countries as possible, but some countries had to be excluded depending on the chapter’s research question; for example due to missing data on a crucial variable.

The TIMSS assessments include so-called trend items, meaning that the exact same items are reused in adjacent cycles (for example repeated for both 2007 and 2011; such data are used in Chap. 4 of this report). There are roughly equal numbers of multiple choice and constructed response (open) items. In order to cover the broad range of content and cognitive domains, approximately 200 items were included in the mathematics assessment. To ease the burden of responding to such a large number of items, TIMSS uses a so-called rotating matrix-sampling design

(for more on this, see Martin and Mullis 2012). Hence, students do not all answer the same set of questions/items.

Because each student only responds to a part of the item pool, the TIMSS scaling approach uses multiple imputation methodology to obtain proficiency scores for all students. This method generates multiple imputed scores or plausible values from the estimated ability distributions (Martin and Mullis 2012). In addition, a conditioning process, in which student responses to the items are combined with information about the student's background, is implemented to increase score reliability. Plausible values hence provide consistent estimates of population characteristics. In 1995, the mean mathematics achievement was set to a score of 500, with a standard deviation of 100. After this, all cycles have been calibrated to the same scale as that of 1995 by means of concurrent calibration, using the trend items and data from countries that participated in adjacent cycles (Martin and Mullis 2012).

In addition to assessment in mathematics, students, parents, teachers and school leaders respond to questionnaires with questions pertaining to background and context (Foy et al. 2013).

TIMSS employs a two-stage random sample design, where schools are drawn as a first stage, and then intact classes of students are selected from each of the sampled schools as a second stage. Hence, students are nested within classes, and classes are nested within schools. Students are representative samples of the entire population of students within a country. Teachers are connected to the sample of classes within each country, which does not necessarily mean that TIMSS includes representative samples of teachers. Hence, results concerning teacher variables, such as teachers with high levels of education, reflect representative samples of students whose teachers have high levels of education. Some classes had more than one mathematics teacher. The percentage of students with more than one mathematics teacher was 1.4 % in grade four, and 1.7 % in grade eight. For students with more than one mathematics teacher, data from only one of them was included at random. The amount of data deleted by this procedure was negligibly small.

1.5 Measuring Key Constructs

The rich data from the large number of participating students, teachers, classrooms, schools and educational systems offer great opportunities to explore and compare different solutions to these measurement challenges, and to investigate characteristics of different measurement models. But as issues of validity and reliability of measurement are present in virtually all empirical research, they also provide challenges in secondary analyses of large-scale data such as TIMSS. Typically, few items are available to measure each of the many complex constructs that are central to educational research. Furthermore, since these items need to reflect conceptualizations of constructs in many different cultural and educational contexts, they may not be perfectly relevant as indicators of the theoretical constructs that a particular researcher wants to investigate.

The researchers involved in the different chapters designed measurement approaches to suit their research problems within the common framework and with the data available from TIMSS (see <http://timssandpirls.bc.edu/timss2011/international-database.html>). Below we present the measurement solutions adopted for the constructs used in more than one chapter.

1.5.1 Instructional Quality

Instructional quality is a key construct, central to most of the chapters of this volume. As is described above, there is converging evidence from within-country studies that four dimensions (clarity of instruction, cognitive activation, classroom management, and supportive climate) may be needed to adequately measure instructional quality. In TIMSS, both the student and the teacher questionnaires include items covering some of these aspects. However, some construct under-representation exists in both cases. Furthermore, concerns have been raised about the reliability and validity of both teacher and student assessments of instructional quality. Social desirability bias in teachers' assessments is often mentioned as a threat to validity, as is lack of competence and stability in younger students' assessments of instructional quality. So, both approaches may have benefits and limits. Recent research suggests in addition that while a single student's assessment is likely to be unreliable, the aggregated assessments of a classroom of students may be both reliable and valid (Marsh et al. 2012; Scherer and Gustafsson 2015). All chapters where students' ratings were used therefore identified the construct both at the student and the class level (Marsh et al. 2012; Wagner et al. 2015).

Four chapters investigated instructional quality. Blömeke, Olsen and Suhl (Chap. 2, grade four) used teacher data due to the young age of grade four students. They created three indicators of instructional quality (clarity of instruction, cognitive activation, and supportive climate) from six items included in the teacher questionnaire and used these item parcels as indicators of a latent variable representing instructional quality. They were thus able to deal with the inherent multidimensionality of the construct. Scherer and Nilsen (Chap. 3, grade eight) used four items from the student questionnaire aimed to assess clarity of instruction and supportive climate. They employed a two-level confirmatory factor analysis model with latent variables representing perceived instructional quality at the class- and student-levels. Nortvedt, Gustafsson and Lehre (Chap. 5, grade four) used a similar two-level approach to measure class-level instructional quality, but they took advantage of student assessments of both teaching of mathematics and of reading. Rutkowski and Rutkowski (Chap. 6, grade four) also used student assessments of instructional quality in mathematics with four items in the class- and student-level models to represent instructional quality.

Thus there is considerable overlap between the approaches used in the different chapters, but there also are differences both in the actual items included in the

models and in whether teacher or student responses are relied upon. In the last chapter, we discuss this further, and assess the results obtained from the different analyses.

1.5.2 Teacher Quality

As is described in greater detail in the theoretical section and in Chap. 2, teacher quality may analytically be differentiated into teacher qualifications, such as education, experience and professional development, and teacher characteristics, such as motivation and self-efficacy.

Formal qualifications are indicated by the number of years of education, the level of the teaching license, years of teaching experience, major academic discipline studied, and professional development. These features can be assessed with good reliability. However, formal qualifications are sometimes found to be weakly related to measures of instructional quality or student achievement across educational systems or content areas because a major qualification in mathematics in a program on ISCED level 5 may mean something different than in a program on ISCED level 6 or 7, because recruitment to the more advanced program is more selective. This problem has led to attempts to measure teacher efficiency with value-added techniques, an approach that is approximated in this book by combining the variables available from the TIMSS data set in one model. In other lines of research, teacher knowledge and skills, such as pedagogical content knowledge and content knowledge, are measured directly (see Baumert et al. 2010), but this is not possible to implement in large-scale international studies, unless this is the aim of the study, as was the case with the TEDS-M study (Blömeke et al. 2011, 2013).

Two chapters included teacher quality variables. Blömeke, Olsen and Suhl (Chap. 2, grade four) investigated number of years of experience, level of formal education completed, and major (in this book and the TIMSS framework defined as the main academic discipline studied) in either mathematics or mathematics education, professional development in mathematics instruction, with attention to both broad activities and specific challenges, as well as collaborative school-based professional development with peers. They also measured teacher self-efficacy with items asking about preparedness to teach numbers, geometry and data. Gustafsson and Nilsen (Chap. 4, grade eight) investigated number of years of experience, level of formal education completed, whether teachers had a major qualification in mathematics or not, professional development in five different areas, and teacher self-efficacy in teaching number, algebra, geometry and data and chance. Thus, similar variables were investigated, the differences being due to the fact that different grade levels were investigated.

1.5.3 School Climate

School climate is often regarded as a foundation for instructional quality. Scherer and Nilsen (Chap. 3, grade four) investigated empirically whether this is the case or not across a broad range of countries. Gustafsson and Nilsen (Chap. 4, grade eight) asked if there is a causal relation between school climate and achievement. As a well-established measure of academic climate, SEAS was used in both chapters. In addition, Scherer and Nilsen (Chap. 3) created a safety scale from three items and an order scale from two items of the TIMSS student survey.

1.5.4 Socioeconomic Status

In educational research, socioeconomic status (SES) is often used to control for selection bias, but may also be a variable which is of interest in its own right. In the IEA study frameworks, an item asking about number of books at home (Books) has a long tradition as an indicator of SES. In TIMSS 2011, further SES indicators were introduced: parents' highest level of education and level of home study supports, such as students having their own room or internet connection. The TIMSS Home Educational Resources (HER) index (Martin and Mullis 2012) was created from these indicators.

SES was included as a control variable in the analyses presented in three chapters. Blömeke, Olsen and Suhl (Chap. 2, grade four), and Rutkowski and Rutkowski (Chap. 6, grade four) used Books as an indicator, while Scherer and Nilsen (Chap. 3, grade four) relied on the HER index. A case can be made for both choices. While the HER index has better measurement properties than Books, the latter indicator has remained unaltered for a long time and similar indicators of home background are used in the other international large-scale studies, allowing for easy comparisons with previous research.

1.6 Challenges in Analyzing the Data

In addition to measuring the intended constructs appropriately, data analysis also presented challenges. Those that were common across the chapters in this book are briefly discussed below.

1.6.1 Causality

Many of the research questions asked in this report concern issues of causality. Basically, two types of causal questions can be identified. The first type concerns

causal effects, or whether a certain factor (for example instructional quality) influences an outcome variable, such as mathematics achievement. If there is a causal relation, increasing instructional quality will cause mathematics achievement to improve. However, TIMSS data are cross-sectional by nature and can mostly only provide correlations between instructional quality and achievement. There is insufficient evidence to conclude that a causal relation exists because third-variable explanations or reversed causality cannot be excluded.

If, for example, students receiving better instructional quality also have higher SES, an alternative explanation could be that the correlation arises because SES is related both to achievement and to instructional quality. If information about SES is available, this hypothesis can be tested by statistically controlling for the effect of SES on the relation between instructional quality and achievement. However, given that there are many unobserved variables that potentially may account for an observed correlation between instructional quality and achievement, it is unlikely that data on all of them exists. Cross-sectional studies therefore cannot rule out the possibility that omitted variables are causing an observed correlation. A way to strengthen causal inference is to use a longitudinal approach (Gustafsson 2013). Gustafsson and Nilsen (Chap. 4) present the idea behind such an approach and apply it to analyses of effects of teacher quality and school climate on mathematics achievement using data from TIMSS 2007 and 2011.

The other type of causal question concerns causal mechanisms, or how sequences of variables influence one another. Reversed causality is a well-known problem in educational research using cross-sectional data in this context. An example would be that the relation between teacher quality and student achievement is negative although longitudinal studies show the opposite. An explanation could be that a country may have taken specific actions to compensate for weak student achievement, perhaps by placing the best teachers in the weakest classes. The correlation between teacher quality and student achievement based on cross-sectional data would then be negative, although, in this case, longitudinal data would reveal that classes with better teachers develop better than other classes provided the starting achievement level is taken into consideration.

Illustrating how sequences of variables may influence one another, is Blömeke, Olsen and Suhl's (Chap. 2) study, which tested the hypothesis that teacher quality influences instructional quality, which in turn influences mathematics achievement. The question is whether instructional quality partly mediates the relation between teacher quality and mathematics achievement. A similar question is asked by Scherer and Nilsen (Chap. 3), who examined relations between school climate, instructional quality, and achievement motivation in mathematics, asking if instructional quality mediates the relation between school climate and achievement motivation. Informed by strong theory, application of structural equation modeling can provide insights into the mechanisms through which causal effects occur. However, this kind of study also assumes that the relations among variables are causal, and that there may be omitted variables that would change the patterns of results if they were introduced to the model.

1.6.2 *Multilevel Data*

The sampling design of TIMSS generates data where the observations of students are nested within classes that are nested within schools. Analytical techniques for dealing with such multilevel data are available, and the studies reported here have relied on the procedures implemented in Mplus (Muthén and Muthén 1998–2012). Two levels were included in the analyses because there are few educational systems where the sample includes more than one classroom from each school, making it necessary to combine the school- and class-levels into one class level.

1.6.3 *Measurement Invariance*

Most of the studies presented here took advantage of measurement models with latent variables. While such models offer great possibilities for summarizing several indicators of a construct that is not directly observable while dealing with problems of measurement error, they also offer challenges, because they are based on assumptions that should not be violated. Thus, when data from multiple groups are analyzed, such as different educational systems, the latent variables must have the same meaning across groups. This can be investigated empirically through analyses of measurement invariance of the latent variables across groups.

To answer the research questions posed by this book, so-called “metric invariance” must be established because relations between variables are to be compared across countries. This is tested through comparing the loadings of the observed indicators on the latent variables to see if they are the same; if that is the case, metric invariance is established, and relations between constructs across countries can be meaningfully compared. To be able to compare means of latent variables across countries, an added requirement would be that the means of the observed indicators, given the latent variable, are invariant across groups (“scalar invariance”).

In the analyses here, the measurement invariance of the latent constructs used was investigated. In only one case was scalar invariance supported by the data (the bullying scale in Chap. 6), but in most cases metric invariance was supported; in exceptions, separate models were fitted for each group.

1.7 Overview of Chapters

Chapter 2 examines the relations between teacher quality, instructional quality and mathematics achievement. Chapter 3 investigates the relations between school climate, instructional quality and student motivation in mathematics. Chapters 2 and 3 conducted cross-sectional secondary analysis of TIMSS 2011 data, using the



Table 1.1 Overview of the chapters

Chapter	Objective	Data and sample	Method of analysis
1	Describe conceptual framework and methodological challenges of the book	–	
2	Investigate relations between instructional quality, teacher quality and student achievement	TIMSS 2011, grade 4	Multi-group, multilevel (students and classes) SEM, mediation model
3	Investigate the relations between school climate, instructional quality and student motivation in mathematics	TIMSS 2011, grade 8	Multi-group, multi-level (students and classes) SEM, mediation models
4	Investigate the influence of teacher quality and school climate on achievement	TIMSS 2007 and 2011, grade 8	Longitudinal analyses of within-country change, difference in differences
5	Investigate if instructional quality can weaken the relation between reading and mathematics achievement	TIMSS and PIRLS 2011, grade 4	Multilevel (students and classes) SEM, random slopes models
6	Determine the degree to which instructional quality serves as a protective factor against school bullying victimization	TIMSS 2011, grade 4	Zero-inflated Poisson regression
7	Summary, discussion and concluding remarks	–	

Note: SEM structural equation modelling. IEA TIMSS and PIRLS 2011 data are available at <http://timssandpirls.bc.edu/>

grade four data set in Chap. 2 and the grade eight data set in Chap. 3, applying multi-group multilevel structural equation modeling (MG-MSEM). Chapter 4 investigates a similar research question to Chap. 3, taking advantage of TIMSS 2007 and 2011 data that are longitudinal at the country-level (Gustafsson 2013). Chapter 5 goes deeper into mathematics education, and investigates the role instructional quality plays in the relation between reading and mathematics achievement in grade four by drawing on both TIMSS 2011 and PIRLS 2011 data. In Chap. 6, instructional quality is investigated in the context of bullying experienced in grade four. Finally, in Chap. 7, we summarize the findings of the five studies, discussing both their contribution to the state of research, and limitations and further research needs (Table 1.1).

References

- Alexander, P. A., & Jetton, T. L. (2000). Learning from text: A multidimensional and developmental perspective. In M. L. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 285–310). Mahwah, NJ: Lawrence Erlbaum Associates.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: WH Freeman and Co.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133–180.
- Blömeke, S., & Delaney, S. (2014). Assessment of teacher knowledge across countries: A review of the state of research. In S. Blömeke, F.-J. Hsieh, G. Kaiser, & W. H. Schmidt (Eds.), *International perspectives on teacher knowledge, beliefs and opportunities to learn* (pp. 541–585). Dordrecht: Springer.
- Blömeke, S., Suhl, U., & Döhrmann, M. (2013). Assessing strengths and weaknesses of teacher knowledge in Asia, Eastern Europe and Western countries: Differential item functioning in TEDS-M. *International Journal of Science and Mathematics Education*, *11*, 795–817.
- Blömeke, S., Suhl, U., & Kaiser, G. (2011). Teacher education effectiveness: Quality and equity of future primary teachers' mathematics and mathematics pedagogical content knowledge. *Journal of Teacher Education*, *62*, 154–171.
- Bryk, A. S., & Schneider, B. L. (2002). *Trust in schools: A core resource for improvement*. New York: Russell Sage Foundation Publications.
- Chapman, C., Armstrong, P., Harris, A., Muijs, D., Reynolds, D., & Sammons, P. (Eds.). (2012). *School effectiveness and improvement research, policy and practice: Challenging the orthodoxy?* Abingdon, Oxon: Routledge.
- Creemers, B. P., & 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.
- Creemers, B., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. Abingdon, Oxon: Routledge.
- Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., & Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal*, *52*, 1133–1159.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132.
- Engel, L. C., Rutkowski, D., & Rutkowski, L. (2009). The harsher side of globalisation: Violent conflict and academic achievement. *Globalisation, Societies and Education*, *7*(4), 433–456.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9.
- Ferguson, R. (2010). *Student perceptions of teaching effectiveness*. Retrieved from http://www.gse.harvard.edu/ncte/news/Using_Student_Perceptions_Ferguson.pdf.
- Foy, P., Arora, A., & Stanco, G. A. (Eds.). (2013). TIMSS 2011 user guide for the international database. *Supplement 1: International version of the TIMSS 2011 background and curriculum*

- questionnaires*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center and International Association for the Evaluation of Educational Achievement.
- Goe, L. (2007). The link between teacher quality and student outcomes: A research synthesis. *National comprehensive center for teacher quality*.
- Goldstein, S. E., Young, A., & Boyd, C. (2008). Relational aggression at school: Associations with school safety and social climate. *Journal of Youth and Adolescence*, 37(6), 641–654.
- Good, T. L., Wiley, C. R., & Florez, I. R. (Eds.). (2009). Effective teaching: An emerging synthesis. *International handbook of research on teachers and teaching* (pp. 803–816). New York: Springer.
- Gregory, A., Cornell, D., & Fan, X. (2012). Teacher safety and authoritative school climate in high schools. *American Journal of Education*, 118(4), 401–425.
- Gustafsson, J.-E. (2013). Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement I. *School Effectiveness and School Improvement*, 24(3), 275–295.
- Gustafsson, J. E., Hansen, K. Y., & Rosèn, M. (2013). Effects of home background on student achievement in reading, mathematics, and science at the fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—implications for early learning*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Hansen, K., & Munk, I. (2012). Exploring the measurement profiles of socioeconomic background indicators and their differences in reading achievement: A two-level latent class analysis. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 5, 49–77.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. USA: Routledge.
- Hole, A., Onstad, T., Grønmo, L. S., Nilsen, T., Nortvedt, G. A., & Braeken, J. (2015). Investigating mathematical theory needed to solve TIMSS and PISA mathematics test items. *Paper presented at the 6th IEA International Research Conference, 24–26 June 2015*, Cape Town, South Africa. Retrieved from http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2015/IRC-2015_Program.pdf.
- Hoy, W. K., Tarter, C. J., & Hoy, A. W. (2006). Academic optimism of schools: A force for student achievement. *American Educational Research Journal*, 43(3), 425–446.
- Kane, T., & Cantrell, S. (Eds.). (2010). Learning about teaching: Initial findings from the measures of effective teaching project. *MET Project Research Paper, Bill & Melinda Gates Foundation, Seattle, WA*. Retrieved from <https://docs.gatesfoundation.org/Documents/preliminary-findings-research-paper.pdf>.
- Kane, T., & Cantrell, S. (2012). Gathering feedback for teaching. Combining high-quality observations with student surveys and achievement gains. *MET Project Research Paper, Bill & Melinda Gates Foundation, Seattle, WA*. Retrieved from <http://files.eric.ed.gov/fulltext/ED540960.pdf>.
- Kilpatrick, J. (2014). Competency frameworks in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 85–87). Dordrecht: Springer.
- Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, & K.J. Tillmann (Eds.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 334–359). Opladen, Germany: Leske & Budrich.
- Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). New York: Waxmann Publishing Co.
- Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: “Aufgabenkultur” und Unterrichtsgestaltung im internationalen Vergleich. In E. Klieme & J. Baumert (Eds.), *TIMSS: Impulse für Schule und Unterricht* (pp. 43–47). Bonn: Bundesministerium für Bildung und Forschung.

- Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Teachers' occupational well-being and quality of instruction: The important role of self-regulatory patterns. *Journal of Educational Psychology, 100*(3), 702.
- Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820. doi:10.1037/a0032583.
- Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction, 18*(5), 468–482.
- 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.
- Kyriakides, L., Creemers, B., & Antoniou, P. (2009). Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25*(1), 12–23.
- Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2010). A synthesis of studies searching for school factors: Implications for theory and research. *British Educational Research Journal, 36*(5), 807–830. doi:10.1080/01411920903165603.
- Kyriakides, L., Creemers, B., Muijs, D., Rekers-Mombarg, L., Papastyliaou, D., Van Petegem, P., & Pearson, D. (2014). Using the dynamic model of educational effectiveness to design strategies and actions to face bullying. *School Effectiveness and School Improvement, 25*(1), 83–104. doi:10.1080/09243453.2013.771686.
- Lipowsky, F., Rakoczy, K., Pauli, Ch., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean theorem. *Learning and instruction, 19*, 527–537.
- Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106–124.
- Martin, M. O., Foy, P., Mullis, I. V. S., & O'Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade-Implications for early learning* (pp. 109–178). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012a). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2012b). *TIMSS 2011 Assessment Frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Murray-Harvey, R., & Slee, P. T. (2010). School and home relationships and their impact on school bullying. *School Psychology International, 31*(3), 271–295.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (Seventh edition). Los Angeles, CA: Muthén & Muthén.
- Nilsen, T., & Gustafsson, J. E. (2014). School emphasis on academic success: Exploring changes in science performance in Norway between 2007 and 2011 employing two-level SEM. *Educational Research and Evaluation, 20*(4), 308–327.
- Niss, M. (2003). Mathematical competencies and the learning of mathematics: the Danish KOM project. *Paper presented at the Proceedings of the 3rd Mediterranean Conference on*

- Mathematical Education*. Atenas: Hellenic Mathematical Society. Retrieved from <http://www.math.chalmers.se/Math/Grundutb/CTH/mve375/1213/docs/KOMkompetenser.pdf>.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38, 109–119.
- Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology*, 104, 700–712.
- Richard, J. F., Schneider, B. H., & Mallet, P. (2012). Revisiting the whole-school approach to bullying: Really looking at the whole school. *School Psychology International*, 33(3), 263–284.
- Ruddell, R. B., & Unrau, N. J. (Eds.). (2004). *Theoretical models and processes of reading* (5th ed.). Newark, DE: International Reading Association.
- Sammons, P. (2009). The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools. *School Effectiveness and School Improvement*, 20(1), 123–129. doi:10.1080/09243450802664321.
- Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.
- 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, 1550. doi:10.3389/fpsyg.2015.01550.
- Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM*, 1–12. doi:10.1007/s11858-016-0765-0.
- Schoenfeld, A. H., & Kilpatrick, J. (2008). Toward a theory of proficiency in teaching mathematics. In D. Tirosh & T. Wood (Eds.), *International handbook of mathematics teacher education* (Vol. 2, pp. 321–354). Rotterdam: Sense Publishers.
- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499.
- Skues, J. L., Cunningham, E. G., & Pokharel, T. (2005). The influence of bullying behaviours on sense of school connectedness, motivation and self-esteem. *Australian Journal of Guidance and Counselling*, 15(1), 17–26.
- Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research*, 83(3), 357–385.
- Tatto, M. T., Peck, R., Schwille, J., Bankov, K., Senk, S. L., Rodriguez, M., & Rowley, G. (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-MM)*. Amsterdam: The International Association for the Evaluation of Academic Achievement. Retrieved from http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/TEDS-M_International_Report.pdf.
- Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2015). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, 3, 52. doi:10.1037/edu0000075.
- Wang, M.-T., & Degol, J. L. (2015). School climate: A review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review*, 1–38. doi:10.1007/s10648-015-9319-1.

Relation of Student Achievement to the Quality of Their Teachers and Instructional Quality

Sigrid Blömeke, Rolf Vegar Olsen and Ute Suhl

Abstract This chapter examines how crucial input and process characteristics of schooling are related to cognitive student outcomes. It was hypothesized that teacher quality predicts instructional quality and student achievement, and that instructional quality in turn predicts student achievement. The strengths of these relations may vary across countries, making it impossible to draw universal conclusions. However, similar relational patterns could be evident within regions of the world. These hypotheses were investigated by applying multi-level structural equation modeling to grade four student and teacher data from TIMSS 2011. The sample included 205,515 students from 47 countries nested in 10,059 classrooms. Results revealed that teacher quality was significantly related to instructional quality and student achievement, whereas student achievement was not well predicted by instructional quality. Certain characteristics were more strongly related to each other in some world regions than in others, indicating regional patterns. Participation in professional development activities and teachers' sense of preparedness were, on average, the strongest predictors of instructional quality across all countries. Professional development was of particular relevance in Europe and Western Asian/Arabian countries, whereas preparedness played an important role in instructional quality in South-East Asia and Latin America. The ISCED level of teacher education was on average the strongest predictor of student achievement across all countries; this characteristic mattered most in the Western Asia/Arabia region.

S. Blömeke (✉) · R.V. Olsen
Faculty of Educational Sciences, Centre for Educational Measurement at the University of Oslo (CEMO), Oslo, Norway
e-mail: sigrid.blomeke@cemo.uio.no

R.V. Olsen
e-mail: r.v.olsen@cemo.uio.no

U. Suhl
Institut Für Erziehungswissenschaften, Humboldt-Universität Zu Berlin, Berlin, Germany
e-mail: ute.suhl@staff.hu-berlin.de

Keywords Instructional quality · Teacher quality · Student achievement · Two-level structural equation modeling mediation models · Trends in Mathematics and Science Study (TIMSS) 2011

2.1 Rationale

The framework of the TIMSS study describes policy malleable features at the system, school, classroom and student level that are known to influence selected desired outcomes of education, such as achievement in the core curricular domain of mathematics (Mullis et al. 2009). Without going into details of the multi-stage sampling procedure applied in TIMSS, a distinguishing feature is that it produces a sample of intact classrooms, including their mathematics teacher(s), representing the 4th grade students in the participating countries (Joncas and Foy 2012). In other words, the data set from TIMSS provides a unique opportunity to link responses from students in a classroom with those from their teacher(s) for a large number of world regions, educational cultures and systems (in the following also called “countries”).

It is well known from previous research that classroom matters. First and foremost, teachers matter (for a summary of the state of research see, for example, Kyriakides et al. 2009). Teachers’ experience, teacher education background, beliefs and motivations, as well as their content knowledge, pedagogical content knowledge, and general pedagogical knowledge (actual and perceived), are characteristics that, to varying degrees, have been shown to have effects on student outcomes. Secondly, teaching or instruction matters for student outcomes (for a summary of research see, for example, Seidel and Shavelson 2007). Educational effectiveness studies and qualitatively oriented classroom observational studies seem to converge on some key features of high quality instruction. In short, high quality teaching consists of instructional practices leading to students being dedicated to cognitively active time on task.

However, there are not many studies seeking to model how teacher quality is related to student achievement, and how teacher quality is put into action by what teachers actually do in the classrooms. This research gap applies particularly to international comparative research. Most of the reported studies of these relationships, although valuable (for example Baumert et al. 2010), took place in one country only, and usually in a Western country. Comparative research that tries to extend the findings from these studies to other educational cultures and systems is lacking. The generalizability of the findings is therefore an open question.

From most definitions of learning it follows that learning occurs as a result of an interaction between the individual learner and his or her surroundings. In the school setting these are, such interactions that most often are generally planned and staged by the teacher. Teacher quality should thus matter, but the degree of its influence may vary by depending on teacher quality indicators or among educational systems. Furthermore, although some aspects of teacher quality have been shown to be

directly positively related to student outcomes, they are also resources for the instructional processes in classrooms, and hence teacher quality may be a predictor of instructional quality. As pointed out above, we know for instance that stronger pedagogical content knowledge of mathematics teachers (one possible indicator of teacher quality) is positively related to student achievement in mathematics (Baumert et al. 2010). This may be a direct effect, where teachers influence individual students by diagnosing their (mis)conceptions and addressing these directly, or it may influence the teachers to create classroom conditions for learning where students are cognitively challenged and activated.

In line with this reasoning, we hypothesized that teacher quality is partly mediated by instructional quality. Although the capacity of TIMSS to address this issue is limited because of its design and instruments, the study has collected a lot of information from the teachers about their background and dispositions. The study has also collected rudimentary information, from both the teachers and the students, about the degree to which the classroom is characterized by instructional activities known from other research to be beneficial for student learning.

Against this background, the following research questions led this study:

- (1) *Which teacher characteristics are significantly related to instructional quality?*
- (2) *To what extent do the relations between teacher quality and instructional quality vary by country? Is it possible to identify regions or clusters of countries where similar relational patterns exist?*
- (3) *Is instructional quality significantly related to student achievement? Does this relation vary by country, and, does a pattern exist that applies to countries from larger regions or cultures?*
- (4) *If teacher quality is significantly related to instructional quality and if instructional quality is significantly related to achievement, does instructional quality partially mediate the relation between teacher quality and student outcomes?*

2.2 Theory

2.2.1 Educational Effectiveness Research as the Point of Reference

The studies presented in this book are rooted in the tradition of educational effectiveness research (Sammons 2009; Scheerens and Bosker 1997). The analysis in this chapter seeks to establish the structural relationship between aspects of teacher quality, instructional quality and student outcome with the hypotheses that teacher quality matters significantly positively for instructional quality and student outcomes, that instructional quality matters significantly positively for student

outcomes, and that instructional quality partly mediates the influence of teacher quality on student outcomes. Several models for effective schools have been proposed, all of which to some degree include teacher quality and instructional quality. Our model employed a section of the dynamic model proposed by Creemers and Kyriakides (2008). However, this is a “static” model used to analyze cross-sectional data, and thus should accordingly be seen as a pragmatic conceptualization of the relationship between these core concepts of teaching and learning, reflecting the design and data available from the TIMSS study.

Educational effectiveness research (Nordenbo et al. 2008; Scheerens 2013) relates to an explicit notion of input-process-output logic, usually represented by regression models, where an educational outcome, in our case grade four students’ mathematics achievement, is modelled as a function of one or more independent variables, in our case teacher quality and instructional quality. In most of these models one or more intervening concepts are included, in our case instructional quality, to conceptually relate the modelled variables. In other words, this is empirical research that tries to open up the educational system as a “black-box”, where the input is the amount of resources, conditions or other antecedents hypothesized to be related to variation in the outcome. The complexities of studying the degree to which possible inputs affect an outcome involves variables that relate to one or more of the levels in the education system. TIMSS is designed to provide data where these complexities are represented by data at both the student and the class/teacher level.

Scheerens (2013, pp. 10–12) suggested that the lack of a unifying theoretical model for school research may well reflect that “[t]he complexity of educational ‘production’ may be such that different units and levels are addressed by different theories,” and he concluded his systematic review of the theoretical underpinning of educational effectiveness research by stating “[a]s it comes to furthering educational effectiveness research, the piecemeal improvement of conceptual maps and multi-level structural equation models may be at least as important as a continued effort to make studies more theory driven.” This chapter and the other chapters in this book are intended to provide improvements in the conceptual understanding of what characterizes effective instructional practice. By the inclusion of multiple educational systems, these chapters will also contribute to address questions regarding the degree to which educational effectiveness research can provide models and theories which are sensitive also to the wider social, political and cultural context in which education is embedded.

2.2.2 *Teacher Quality*

Teacher quality (TQ) includes different indicators of teacher qualifications, in particular characteristics of teachers’ educational background, amount of experience in teaching, and participation in professional development (PD), as well as personality characteristics such as teachers’ self-efficacy. A number of previous

studies were able to relate measures of such teacher characteristics to student educational outcomes (see for instance the review by Wayne and Youngs 2003).

Evidence suggests that the quality of teacher education does have an impact on teachers' educational outcomes in terms of teacher knowledge and skills (Blömeke et al. 2012; Boyd et al. 2009; Tatto et al. 2012); these, in turn, are significantly related to instructional quality and student achievement (Baumert et al. 2010; Hill et al. 2005; Kersting et al. 2012). The degree and major academic disciplines studied can be regarded as indicators of teachers' education, although they are only rough approximations of specific opportunities to learn. In the case of mathematics teachers, a major in mathematics delivers the body of content knowledge necessary to present mathematics to learners in a meaningful way and to connect mathematical ideas and topics to one another, as well as to the learner's prior knowledge and future learning objectives (Wilson et al. 2001; Cochran-Smith and Zeichner 2005). However, knowing the content provides only a foundation for teaching; student achievement is higher if a strong subject-matter background is combined with strong educational credentials (Clotfelter et al. 2007). Correspondingly, teachers' pedagogical content knowledge and content knowledge of mathematics are of great importance for instructional quality and student achievement in mathematics, with the former exerting a greater effect than the latter (Baumert et al. 2010; Blömeke and Delaney 2012). Whether teachers had an education where mathematics or mathematics education were a major focus and the type of degree are proxy variables available in TIMSS. This makes it possible to study how teachers' educational background may affect teaching and students' achievement across countries.

An almost universal characteristic seems to be that teachers do not feel sufficiently prepared for their complex tasks, in particular during the first years on the job (Kee 2012). TIMSS developed three constructs reflecting teachers' preparedness to teach numbers, geometry and data, respectively. The constructs were developed within the context of Bandura's social-cognitive theory, and the measures of teachers' preparedness for teaching may reasonably be assumed to reflect a concept which is similar to teacher self-efficacy (Bandura 1986; Pajares 1996). Self-efficacy beliefs influence thought patterns and emotions, which in turn enable or inhibit actions. Teachers with strong self-efficacy are typically more persistent and make stronger efforts to overcome classroom challenges than others (Tschannen-Moran et al. 1998). TIMSS provides data about teachers' sense of preparedness so that the relation of this dimension of self-efficacy can be examined across countries.

In almost all countries, a variety of professional development activities exist, from very short classes to comprehensive programs (Goldsmith et al. 2014; Guskey 2000). These include school-based programs, and coaching, seminars, or other types of out- and in-service training with the aim of supporting the development of teacher competencies. Overall, meta-analyses support the hypothesis that professional development is positively related to instructional quality and student achievement if the activities meet certain quality characteristics (Timperley et al. 2007). Desimone (2011) classified these quality features into a focus on content,

active learning, coherence, and a certain minimum length of the professional development course to be sustainable and collaborative activities. Collaboration in terms of joint work on cases and practicing under supervision of colleagues seems to be particularly relevant (Boyle et al. 2005). Discussions, reflection and continuous feedback seem to stimulate real changes in beliefs and routines (Goldsmith et al. 2014). TIMSS included several scales that assessed both teachers' participation in formal professional development activities and their involvement in continuous and collaborative professional development activities with colleagues in the school.

2.2.3 *Instructional Quality*

Several studies have established a relationship between measures of instructional quality (InQua) and student achievement, student motivation or other outcomes of schooling. Even though the concept of instructional quality is understood differently by different researchers in the field of educational effectiveness research, there is agreement that it is a multidimensional construct (Baumert et al. 2010; Creemers and Kyriakides 2008). Besides classroom management, three instructional characteristics, namely cognitive activation, clarity of instruction, and a supportive climate, are regarded as essential features (Rakoczy et al. 2010; Decristan et al. 2015). TIMSS includes several measures relating to different aspects of instructional quality, with responses both from teachers and students. For more about the theoretical framework of this construct see Chap. 1.

2.2.4 *Universal, Cultural or Country-Specific Models?*

National specifications of degrees and licenses, foci of programs in terms of majors, amount of in-service training and length and level of teacher education reflect partly overlapping and partly differing visions of the knowledge and skills that teachers are expected to have in a country (Schwille et al. 2013). These specifications of what is required of mathematics teachers before they are allowed to teach mathematics to students at grade four can be assumed to be intentionally developed by national educational policy makers and teacher education institutions (Stark and Lattuca 1997). The same applies to professional development activities provided to teachers or to characteristics regarded as high quality teaching in a country.

In his study of primary school education in England, France, India, Russia, and the United States, Alexander (2001) illustrated the subtle and long-term relationship between culture and pedagogy. Based on videotaped lessons and interviews with teachers, he demonstrated that opportunities to learn provided during schooling reflected a country's educational philosophy transmitted and mediated through the classroom talk between teachers and students. Leung (2006) confirmed similar

cultural differences, specifically with respect to mathematics education in the East and the West. Although mathematics can be regarded as a fairly global construct (Bishop 2004), the curricula of school mathematics, as well as of mathematics teacher education, differ across countries, and are influenced by the context in which they are implemented (Blömeke and Kaiser 2012; Schmidt et al. 1997). With this as a backdrop, it is interesting that a study like TIMSS permits examination of the extent to which the relationship between teacher quality, instructional quality and student achievement can be generalized across the world, or across regions of the world.

2.2.5 Control Variables

Current research indicates that in some countries gender differences in students' mathematics achievement still exist, but that these vary in their direction (Mullis et al. 2012). There is an even stronger relationship between students' socioeconomic background and achievement (Mullis et al. 2012). In order to estimate the relation of teacher quality and instructional quality to mathematics achievement of students at grade four, the background characteristics of students need to be controlled for in the analysis.

2.3 Methods

2.3.1 Sample

This study is based on grade four student and teacher data from the majority of countries participating in TIMSS 2011. Five countries were excluded because there were no data on one or more predictors (Austria, Belgium, Kazakhstan and Russia) or there were very high levels of missing values for most of the variables included in the analysis (Australia). For students with more than one mathematics teacher, data from only one of the teachers was included at random, resulting in a data set with a simple hierarchical structure, where students were nested in one specific class with one specific teacher. The amount of data excluded by this procedure was negligibly small (for details see Chap. 1). The final sample included 205,515 students from 47 countries nested in 10,059 classrooms/teachers with an average classroom size of 20 students. Student sample sizes per country varied between 1423 and 11,228, with the number of classrooms/teachers ranging from 67 to 538, and an average classroom size between 12 and 34 students. The school level was neglected in the analyses to avoid overly complex hierarchical models. Furthermore, the choice of omitting the school level in the analysis is based on the fact that for many countries the classroom and school level cannot be analyzed separately, since only one grade four classroom was drawn per school.

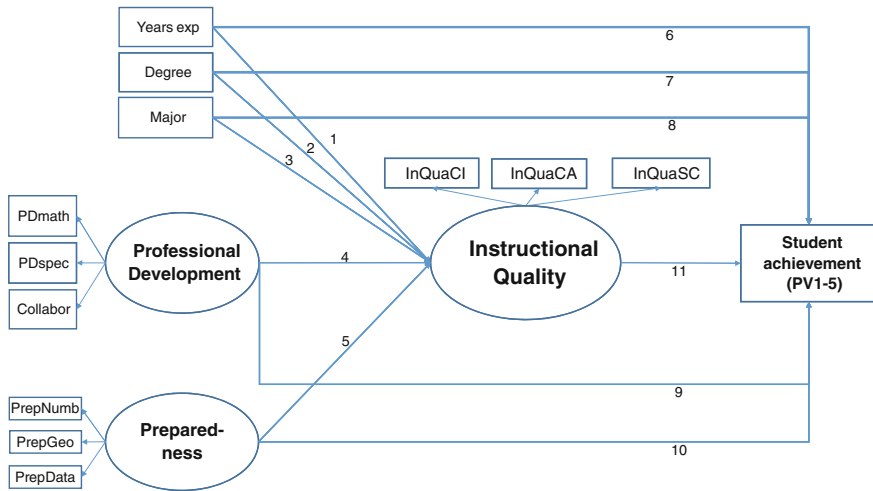


Fig. 2.1 Model of the hypothesized relations of teacher quality (*left hand side* of the figure) in terms of years of teaching experience (*Years exp*), teacher education degree (*Degree*), major focus of teacher education (*Major*), professional development represented by three indicators (*PDmath*, *PDspec* and *Collabor*), and sense of preparedness represented by three indicators (*PrepNumb*, *PrepGeo* and *PrepData*), to instructional quality (*InQuaCI*, *InQuaCA*, and *InQuaSC*), and to student achievement represented by five plausible values (*PV1-5*; *right hand side* of the figure); all abbreviations are explained in Table 2.1, and the numbers linking the relations hypothesized correspond to columns in Table 2.2, where the actual estimates can be found

2.3.2 Variables

A structural model was developed to reflect the hypothesized relations between teacher quality, instructional quality and student achievement (Fig. 2.1). Furthermore, the internationally-pooled descriptives of all variables, including their range across countries were inspected (Table 2.1).¹

Teacher quality measures

Teacher quality is represented by three central dimensions in our model, namely teacher education background, participation in professional development (PD) activities, and teachers’ sense of preparedness. Teacher education background is described by teachers’ years of experience and their formal initial education. These characteristics were included as separate categorical and manifest variables because they do not reflect a joint and theoretically derived latent construct. Instead they represent different and not necessarily related dimensions of teacher quality.

¹For country-specific descriptives including information about their distribution in terms of skewness and kurtosis see Appendices A and B; for more details about the item format see the TIMSS data analysis manual (Foy et al. 2013).

Table 2.1 Descriptives of the variables used in the model

Description of item or item parcel	Variable label in Fig. 2.1	Label in the international database ^a	Mean (SD) and [range of means across countries] ^b	Reliability (coefficient alpha) for item parcels	Percentage missing data and [range across countries]
Number of years of experience	Years exp	ATBG01	“More than 20 years” [“Less than 5 years”–“More than 20 years”] ^c		7 [1–22]
Level of formal education completed	Degree	ATBG04	“Finished ISCED level 5A, first degree” [“ISCED 3”–ISCED 5A, second”] ^c		5 [0–21]
Focus on either mathematics OR mathematics education	Major	ATBG05AC ATBG05BA	0.39 (0.40) [0.04–0.97]		7 [0–41]
PD in mathematics instruction (broad activities)	PDmath	ATBM11A to C	0.43 (0.37) [0.11–0.78]	0.79	7 [0–26]
PD in mathematics instruction (specific challenges)	PDspec	ATBM11D to F	0.37 (0.34) [0.13–0.66]	0.65	7 [0–26]
Collaborative school-based PD with peers	Collabor	ATBG10A to ATBG10E	0.64 (0.30) [0.37–0.94]	0.80	4 [0–19]
Preparedness to teach numbers	PrepNumb	ATBM12AA to ATBM12AH	0.93 (0.13) [0.74–0.99]	0.89	7 [0–27]
Preparedness to teach geometry	PrepGeo	ATBM12BA to ATBM12BG	0.90 (0.15) [0.72–0.97]	0.87	7 [0–27]
Preparedness to teach data	PrepData	ATBM12CA to ATBM12CC	0.90 (0.18) [0.70–0.98]	0.92	14 [1–60]
Instructional quality: Clarity of instruction	InQuaCI	ATBG15A and ATBG15C	0.88 (0.15) [0.68–0.96]	– ^d	4 [0–19]

(continued)



Table 2.1 (continued)

Description of item or item parcel	Variable label in Fig. 2.1	Label in the international database ^a	Mean (SD) and [range of means across countries] ^b	Reliability (coefficient alpha) for item parcels	Percentage missing data and [range across countries]
Instructional quality: Cognitive activation	InQuaCA	ATBG15B and ATBG15F	0.73 (0.19) [0.55–0.87]	– ^d	4 [0–19]
Instructional quality: Supportive climate	InQuaSC	ATBG15D and ATBG15E	0.94 (0.12) [0.78–0.99]	– ^d	4 [0–19]
Student achievement: five plausible values		ASMMAT01 to ASMMAT05	500 (100) [248–606]		

International mean values were computed by averaging country means

Note PD = professional development, SD = standard deviation

^aRefers to the labels in the TIMSS 2011 user guide for the international database (Foy et al. 2013)

^bAll scales transformed to a 0–1 scale representing proportion of maximum score for the scale

^cModal category across countries

^dFor parcels with only two items coefficient alpha is not meaningful

The variation between countries for these variables was remarkably large. Across all countries, the modal category of number of years of experience (“By the end of this school year, how many years will you have been teaching altogether?”) was more than 20 years. The Eastern European countries were particularly pronounced in having many teachers with extensive teaching experience, indicating an older teaching force than elsewhere (see Appendix A, Table A.1). But there were also countries in the data set where the largest group of teachers that taught mathematics at grade four had less than 10 years of experience, and, in some countries, less than 5 years of experience. The Arabian countries were most pronounced in having a relatively young teaching force.

Teachers provided information about their degree from teacher education (“What is the highest level of formal education you have completed?”) out of six options from “did not complete ISCED level 3” to “finished ISCED level 5A, second degree or higher”. Across all countries, the modal category was “ISCED level 5A, first degree”, indicating that many countries had a large proportion of teachers with a bachelor degree. But there were also some countries where the largest group of teachers did not have university degrees, but had completed practically-based programs at ISCED level 3. Italy and the African countries were most pronounced in this respect (see Appendix A, Table A.2). In contrast, there were countries where the largest group of teachers held a university degree at least equivalent to a master

degree (“ISCED level 5A, second degree or higher”). The Eastern European countries were most pronounced in this respect.

A dichotomous variable was created by combining teachers’ responses to two questions regarding their specialization in mathematics. This variable identifies teachers with a major in mathematics or in mathematics education (“During your <post-secondary> education, what was your major or main area(s) of study?” and “If your major or main area of study was education, did you have a <specialization> in any of the following?”). On average, slightly fewer than 40 % of all teachers across all countries had a major with a specialization in mathematics. However, in some countries the proportion was below 10 % (for example in some of the Eastern European countries), whereas in other countries the proportion was more than 80 % (for example in several Arabian countries) (see Appendix A, Table A3).

Furthermore, there were measures of teachers’ participation in PD activities. One set of questions asked the teachers whether or not they had participated in PD during the last two years. These questions are represented in the model by two item parcels reflecting either broad PD activities covering, for example, “mathematics content” in general, or reflecting PD activities preparing for specific challenges, for example “integrating information technology into mathematics”. Across all countries, approximately 40 % of the teachers had participated in broad or specific PD activities, respectively. However, the between-country variation was large, from countries having as few as 10 % the teachers taking part in broad or specific PD, to countries where more than two-thirds of the teachers had taken part in one or both forms of PD activities. It is difficult to discern any systematic cultural pattern in these differences (see Appendix A, Table A.4).

In addition, there was a set of questions regarding whether teachers had taken part in collaborative activities representing continuous, collaborative and school-based PD (“How often do you have the following types of interactions with other teachers?”, with “Visit another classroom to learn more about teaching” as an “exemplary” form of interaction). Across all countries, teachers commonly participated in these types of activities two to three times each month. However, in some countries the largest group of teachers participated in collaborative PD daily or almost daily. These questions were included as the third item parcel defining the latent construct of PD.²

The third teacher quality dimension included in the model reflects teachers’ self-efficacy. The indicator used was their self-reported sense of preparedness to teach specific topics in mathematics within the three domains of number, geometric shapes and measures, as well as data display (“How well prepared do you feel you are to teach the following mathematics topics?”, with “Adding and subtracting with

²The TIMSS data set includes an IRT-based construct composed of these items, labelled as Collaborate to Improve Teaching (CIT). For the purpose of being able to interpret the mean and range in country comparisons in the same way as the other two parcels, we therefore opted for a classical mean raw score used as a third item parcel, each representing different aspects of PD. Furthermore, we were able to confirm measurement invariance of the latent construct PD with this indicator.

decimals” included as an exemplary topic). For each domain, teachers were asked to rate these topics on a three-point Likert scale from “Not well prepared” (0) to “Very well prepared” (2). Teachers were also invited to use a “not applicable” response category if the topic was not covered in their curriculum. In our analysis, the items marked as not applicable were treated as missing. To simplify the final model, the three domains were represented as item-parcel indicators of the latent construct of preparedness. Across all countries, the mean of the three item parcels was each time around 1.8 and, thus, close to the maximum category of the Likert scale. This suggests that there was little discrimination evident in the items. The international variation was also more limited within this dimension than in others included in the model. The lowest means were around 1.5 and, thus, straddled the categories “Somewhat prepared” and “Very well prepared”. Interestingly, slightly lower self-efficacy was most evident in Japan and Thailand (see Appendix A, Table A.5).

Instructional quality measures

The measure of InQua applied in this chapter is based on the teacher questionnaire in TIMSS where six questions asked teachers to report how often they perform various activities in this class (“How often do you do the following in teaching this class?”). This measure was preferred over other measures available (see Sect. 2.5) since it has a more explicit relation to three of the four characteristics of high quality instruction (Table 2.1). Teachers were asked to rate these activities on a four-point Likert scale from “Never” (0) to “Every or almost every lesson” (3). These items are represented by three item parcels with two items in each parcel covering different aspects of the latent construct InQua. The first parcel reflected teaching characteristics that were intended to deepening students’ understanding through clear instruction (such as “Use questioning to elicit reasons and explanations”). The second parcel pursued this objective through cognitive activation (through questions such as “Relate the lesson to students’ daily lives”). The final parcel covered a supportive climate (for example “Praise students for good effort”). Across all countries, the indicators for a supportive climate appeared to be widely present, as the mean was close to the maximum of the scale. The mean of the other two parcels was slightly lower. Interestingly, Scandinavian countries had the lowest means on the cognitive-activation item-parcel (see Appendix A, Table A.6). Some international variation existed on all three item parcels.

Outcome measure

We selected student achievement in mathematics represented by five plausible values as our outcome measure. The scale was defined by setting the international mean to 500 and the standard deviation to 100. Country means varied between 248 and 606 points, which is a difference of more than 3.5 standard deviations (for more information, see Martin and Mullis 2012).

Control variables

Data about gender and socioeconomic background were gathered through students' self-reports to the questions "Are you a girl or a boy?" and the frequently used proxy measure of home background "About how many books are there in your home?"³

2.3.3 Analysis

The research questions were examined using multi-level structural equation modeling (MLSEM). The intra-class correlation (ICC) for students' achievement in the pooled international data set (ICC = 0.30) and within countries (ICC = 0.07–0.56) were all above the threshold at which multi-level modeling is recommended (Snijders and Bosker 2012).

Item-parcels were used as indicators, as recommended when structural characteristics of the constructs are the focus of interest (Little et al. 2002), as applies in the present investigation, and when sample size is limited in comparison to the number of parameters to be estimated (Bandalos and Finney 2001). The latter also applies to the present investigation given that there are only about 140 to 260 classrooms in most of the countries. By using parcels as indicators for the latent variables, the number of free parameters to be estimated was significantly reduced. The items were combined into parcels based on theoretical expectations confirmed by initial exploratory analysis of sub-dimensions in the latent variables included in the model.

Data analysis was carried out using the software MPlus 7.4. The clustered data structure was taken into account by using a maximum-likelihood estimator with robust sandwich standard errors to protect against being too liberal (Muthén and Muthén 2008–2012). Missing data were handled by using the full-information-maximum-likelihood (FIML) procedure. The model fit was evaluated with the chi-square deviance and a range of fit indices.⁴

³The TIMSS data set includes an index representing socioeconomic background in terms of Home Educational Resources (HER) that includes also other indicators such as parental income, occupation and education level. Nevertheless, we opted for using "books at home" because in contrast to HER this variable has remained unaltered for many cycles and very similar indicators of home background are used in all other international large-scale studies. This makes it easier to compare results with previous research. Moreover, "books at home" has been and still is a powerful predictor of achievement (compared to parents' education, which is part of HER).

⁴The fit indexes were evaluated to the following commonly recommended criteria: a ratio of chi-square deviance and degrees of freedom of <2 indicates a very good model fit, estimates <3 indicate a good fit. Estimates of the comparative fit index (CFI) and the Tucker Lewis index (TLI) >0.95 indicate a very good model fit, and estimates >0.90 indicate a good model fit (Hu and Bentler 1999). Estimates of the root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMR) <0.05 indicate a very good model fit, and estimates <0.08 indicate a good model fit.

Before the final model was run, measurement invariance (MI) across countries was tested for the latent constructs in the model. Comparing constructs and their relations across countries produces meaningful results only if the instruments measure the same construct in all countries (Van de Vijver and Leung 1997). In order to ascertain such equivalence, MI was established using multiple-group confirmatory factor analysis (MG-CFA; Chen 2008). As instructional quality and the teacher constructs were measured at the classroom level, we tested for measurement invariance at the school level. Firstly, configural invariance was examined, which means that in each country the same items had to be associated with the same latent factors. As a second step, we tested for metric invariance, by studying whether the factor loadings were invariant across countries. Invariance of factor loadings enabled us to compare the relationship between latent variables across groups. It was possible to establish metric invariance for all latent constructs included in the present model (see Appendix B).

To examine our research questions, a single-group model was first applied before country-by-country analyses were carried out. In the multi-group model, factor loadings were constrained to be the same for all countries, reflecting the metric invariance criterion referred to above, in order to ensure comparability. Indirect relations at the between-level were estimated by multiplying the coefficients for the respective direct relation. In the single-group model, the two control variables gender and books at home were grand-mean centered on the international mean, whereas all predictors, the mediator InQua and the dependent variable student achievement in mathematics were group-mean centered on the country means. In the multi-group model the control variables were again grand-mean centered (this meant now on the country mean) whereas the predictors, the mediator and the dependent variable remained unaltered. Relations were regarded significant on the within-level if $p < 0.05$, but given the relative small number of units at the between-level as compared to the number of parameters to be estimated, a more liberal decision rule for the significance testing with $p < 0.10$ was applied for this level.

2.4 Results

2.4.1 Model Fit

The fit of the pooled model to the full data set was very good; both with respect to relative and to absolute fit indices (see Table 2.2). Only the ratio of the chi-square deviance to the degrees of freedom was unsatisfactory which is commonly observed with large samples. Within countries, the model fit varied substantially but given the small sample sizes the fit was sufficient on most indices in the majority of countries. Only in nine out of the 47 countries more than two of the applied indices indicated an unsatisfactory model fit. Typically for these cases, the CFI and TLI estimates were below the threshold of 0.90 and the SRMR estimate on the between-level above the 0.08 criterion.

Table 2.2 Sample size, intra-class correlation (ICC) and model fit of the pooled and the country-by-country models

Country	Number of students	Number of classes	\emptyset cluster	ICC math	χ^2	df	χ^2/df	RMSEA	SRMR_BW	CFI	TLI
Pooled model	205,515	10,059	20.4	0.30	232.71	60	3.9	0.00	0.04	0.97	0.97
<i>Western Asian/Arabian countries</i>											
Armenia	4433	205	21.6	0.27	78.24	60	1.3	0.01	0.07	0.94	0.93
Azerbaijan	3574	226	15.8	0.46	110.94	60	1.9	0.02	0.09	0.82	0.77
Bahrain	3656	150	24.4	0.27	100.29	60	1.7	0.01	0.10	0.85	0.81
Georgia	4322	220	19.6	0.36	121.16	60	2.0	0.02	0.08	0.78	0.72
Iran	5271	229	23.0	0.35	75.45	60	1.3	0.01	0.07	0.96	0.95
Kuwait	3401	177	19.2	0.14	93.72	60	1.6	0.01	0.06	0.85	0.81
Oman	9584	389	24.6	0.20	79.92	60	1.3	0.01	0.07	0.96	0.95
Qatar	3485	169	20.6	0.44	107.24	60	1.8	0.02	0.07	0.91	0.88
Saudi Arabia	4202	189	22.2	0.36	54.65	60	0.9	0.00	0.06	1.00	1.02
Turkey	6896	250	27.6	0.38	116.92	60	2.0	0.01	0.08	0.89	0.85
United Arab Emirates	11,228	467	24.0	0.45	107.86	60	1.8	0.01	0.06	0.90	0.87
Yemen	6357	190	33.5	0.42	68.20	60	1.1	0.01	0.08	0.90	0.87
<i>African countries</i>											
Botswana	3392	126	26.9	0.36	99.93	60	1.7	0.01	0.09	0.90	0.87
Morocco	4186	186	22.5	0.51	108.80	60	1.8	0.01	0.09	0.87	0.83
Tunisia	3090	149	20.7	0.25	90.01	60	1.5	0.01	0.09	0.91	0.88
<i>Latin American countries</i>											
Chile	4695	174	27.0	0.38	92.86	60	1.6	0.01	0.09	0.82	0.77
Honduras	2671	120	22.3	0.46	84.45	60	1.4	0.01	0.09	0.87	0.83
<i>European countries</i>											
Croatia	4004	260	15.4	0.12	99.08	60	1.7	0.01	0.08	0.92	0.90
Czech Republic	4363	228	19.1	0.14	87.00	60	1.5	0.01	0.08	0.96	0.95

(continued)

Table 2.2 (continued)

Country	Number of students	Number of classes	\emptyset cluster	ICC math	χ^2	df	χ^2/df	RMSEA	SRMR_BW	CFI	TLI
Denmark	3355	187	17.9	0.14	106.95	60	1.8	0.02	0.08	0.91	0.88
Finland	4053	241	16.8	0.12	105.21	60	1.8	0.01	0.08	0.91	0.89
Germany	3102	171	18.1	0.19	107.09	60	1.8	0.02	0.08	0.92	0.90
Hungary	4919	243	20.2	0.25	82.03	60	1.4	0.01	0.06	0.97	0.96
Italy	2360	137	17.2	0.28	63.22	60	1.1	0.01	0.07	0.99	0.98
Lithuania	4540	273	16.6	0.19	147.20	60	2.5	0.02	0.06	0.89	0.86
Netherlands	2038	110	18.5	0.14	104.54	60	1.7	0.02	0.08	0.91	0.88
Norway	2841	184	15.4	0.15	68.67	60	1.1	0.01	0.08	0.98	0.97
Poland	4821	249	19.4	0.10	142.93	60	2.4	0.02	0.09	0.86	0.82
Portugal	3674	221	16.6	0.37	110.29	60	1.8	0.02	0.08	0.87	0.84
Romania	3523	193	18.3	0.38	118.40	60	2.0	0.02	0.09	0.83	0.78
Serbia	4219	214	19.7	0.23	73.20	60	1.2	0.01	0.07	0.96	0.95
Slovak Republic	5331	301	17.7	0.27	115.98	60	1.9	0.01	0.07	0.93	0.91
Slovenia	4350	239	18.2	0.07	90.71	60	1.5	0.01	0.07	0.94	0.93
Spain	3715	181	20.5	0.17	109.22	60	1.8	0.02	0.09	0.91	0.88
Sweden	3554	211	16.8	0.09	82.33	60	1.4	0.01	0.09	0.96	0.95
<i>English-speaking countries</i>											
England	2499	134	18.6	0.22	98.29	60	1.6	0.02	0.08	0.91	0.89
Ireland	4303	215	20.0	0.16	117.52	60	2.0	0.02	0.09	0.85	0.81
New Zealand	5189	418	12.4	0.29	162.90	60	2.7	0.02	0.07	0.81	0.76
Northern Ireland	2909	151	19.3	0.11	78.86	60	1.3	0.01	0.07	0.98	0.97
USA	10,460	538	19.4	0.26	100.13	60	1.7	0.01	0.05	0.96	0.95

(continued)

Table 2.2 (continued)

Country	Number of students	Number of classes	\emptyset cluster	ICC math	χ^2	df	χ^2/df	RMSEA	SRMR_BW	CFI	TLI
<i>South-east Asian countries</i>											
Hong Kong	3504	129	27.2	0.39	112.44	60	1.9	0.02	0.08	0.86	0.81
Japan	3651	156	23.4	0.06	104.69	60	1.7	0.01	0.10	0.94	0.92
Korea	4026	142	28.4	0.04	84.29	60	1.4	0.01	0.09	0.97	0.96
Malaysia	1423	67	21.2	0.09	136.79	60	2.3	0.03	0.11	0.65	0.55
Singapore	6044	338	17.9	0.56	152.34	60	2.5	0.02	0.08	0.84	0.79
Chinese Taipei	4218	154	27.4	0.07	98.71	60	1.7	0.01	0.10	0.95	0.93
Thailand	4084	158	25.8	0.43	92.21	60	1.5	0.01	0.07	0.93	0.91

Note ICC = intra-class correlation, df = degrees of freedom, RMSEA = root mean square error of approximation, SRMR_BW = standardized root mean square residual for the between level, CFI = comparative fit index, TLI = Tucker-Lewis index

2.4.2 Relation Between Teacher Quality, Instructional Quality and Mathematics Achievement

The pooled model using the data from all countries reveals that participation in PD activities and teachers' sense of preparedness were the strongest predictors of InQua (see Table 2.2), with relatively large effect sizes given that the directions of relations typically vary across countries. Effect sizes around $\beta = 0.20$ may therefore be a first indication of a widely recognizable, if not universal, pattern. This is supported by the country-by-country results. In almost half of the countries PD activities (23 countries) and preparedness (22) were significantly related to InQua, with moderately strong effect sizes ($\beta = 0.61$ or $\beta = 0.50$ respectively), all of which were uniformly positive. Whereas PD activities were related to InQua particularly in European (11 out of 18) and Western Asian/Arabian (7 out of 12) countries, teachers' sense of preparedness was significantly associated with InQua in South-East Asia (4 out of 7), Latin America (2 out of 2) and the Scandinavian (4 out of 5) countries. The relevance of the predictor preparedness was also evident through its somewhat weaker, but still statistically significant relation to student achievement.

Another predictor that influenced InQua and students' mathematics achievement was teachers' experience. On average, across countries, students with higher mathematics achievement were taught by more experienced teachers, and teachers with more experience also reported higher instructional quality. However, for both of these relationships there were also significant effects in the opposite direction for a number of countries, which contradicts the hypothesized relationship.

Teachers' level of education was not associated with InQua in the pooled data set, but a significant positive relationship was found in nine countries. However, students who were taught by teachers with relatively higher ISCED levels performed somewhat higher in the mathematics achievement test, and this positive relationship was also confirmed for twelve of the countries. This characteristic was most prominent in the Western Asia/Arabia-region, although with moderate effect sizes.

Whether a teacher education program had had a major focus on mathematics or mathematics education did not significantly predict InQua. Still, as with teacher education level, students in classrooms demonstrating stronger mathematics achievement were in the overall international analysis more often taught by a teacher who had majored in one of these fields. Within countries, these relationships were mostly insignificant, but we found also both moderate significant positive and negative coefficients in some countries (Table 2.3).

Across all countries, mathematics achievement of students at grade four was not predicted by InQua, and within countries the predictor had a significant relation to achievement in only three countries. As a result, the mediation effect of InQua was negligible and thus the hypothesized mediation effect of InQua on student achievement is not supported by the data included in this analysis.

Table 2.3 Results for the single-group pooled model and the country-by-country models. (Numbers in column headers refer to relations displayed in Fig. 2.1)

Country	InQua on indicators of TQ										Achievement on TQ and InQua										Indirect effects of TQ × InQua on Student achievement					Level 1 variables	
	1	2	3	4	5	6	7	8	9	10	11	1 × 6	2 × 6	3 × 6	4 × 6	5 × 6	Girl	Books									
Pooled model	0.08	-0.00	-0.00	0.21	0.18	0.08	0.05	0.05	0.00	0.06	0.02	0.00	-0.00	-0.00	0.00	0.00	-0.04	0.21									
<i>Western Asian/Arabian countries</i>																											
Armenia	-0.23	0.12	-0.19	0.21	0.19	0.09	0.27	0.05	0.10	0.09	-0.10	0.02	-0.01	0.02	-0.02	0.03	0.08										
Azerbaijan	0.02	0.49	0.31	-0.12	0.33	0.08	-0.01	-0.13	0.23	-0.06	0.24	0.00	0.12	0.07	-0.03	0.08	0.11										
Bahrain	-0.08	0.16	-0.16	0.28	0.19	0.06	0.22	-0.48	0.12	0.11	-0.01	0.00	0.00	0.00	0.00	0.06	0.05										
Georgia	0.07	0.07	0.06	-0.16	0.07	-0.13	-0.06	0.03	-0.04	0.17	0.15	0.01	0.01	-0.02	0.01	0.03	0.16										
Iran	0.14	0.06	0.16	0.21	0.17	0.23	0.12	0.12	0.02	0.02	0.00	0.00	0.00	0.00	0.00	-0.02	0.14										
Kuwait	0.07	-0.13	0.19	0.43	0.12	0.01	-0.16	-0.08	-0.10	0.07	0.19	0.01	-0.02	0.04	0.08	0.02	-0.01										
Oman	0.02	0.21	-0.19	0.29	0.17	0.11	0.09	0.01	0.13	-0.00	0.29	0.01	0.06	-0.06	0.08	0.05	0.12										
Qatar	0.12	0.13	0.16	0.49	0.01	0.26	0.25	-0.26	-0.08	-0.02	0.09	0.01	0.01	0.01	0.04	0.00	0.02										
Saudi Arabia	-0.30	-0.17	-0.03	0.61	0.22	0.13	-0.00	0.04	-0.01	0.07	0.23	-0.07	-0.04	-0.01	0.14	0.05	0.06										
Turkey	0.06	-0.03	0.07	0.22	0.10	0.45	0.18	0.08	0.17	0.10	0.11	0.01	0.00	0.01	0.02	0.01	-0.03										
United Arab Emirates	0.13	-0.06	0.15	0.14	0.22	-0.06	0.05	-0.37	-0.13	0.05	0.15	0.02	-0.01	0.02	0.02	0.03	0.05										
Yemen	-0.13	-0.03	0.05	0.29	0.38	-0.08	0.02	0.04	0.17	-0.16	0.28	-0.04	-0.01	0.01	0.08	0.11	-0.02										
<i>African countries</i>																											
Botswana	-0.20	-0.25	0.02	0.32	0.16	0.19	0.19	-0.05	0.41	0.04	-0.08	0.02	0.02	0.00	-0.03	-0.01	0.13										
Morocco	0.34	0.18	-0.02	0.08	-0.12	-0.24	0.03	0.18	0.01	0.06	-0.11	-0.04	-0.02	0.00	-0.01	0.01	0.03										
Tunisia	0.29	0.53	-0.06	0.09	0.50	0.13	-0.11	0.07	-0.06	0.05	-0.00	0.00	0.00	0.00	0.00	0.01	0.06										
<i>Latin American countries</i>																											
Chile	0.13	-0.15	-0.10	0.08	0.39	0.04	0.03	0.09	0.25	0.22	-0.11	-0.01	0.02	0.01	-0.01	-0.04	-0.08										
Honduras	0.31	-0.24	0.25	0.20	0.48	-0.09	0.20	0.02	-0.12	-0.08	0.17	0.05	-0.04	0.04	0.03	0.08	-0.11										

(continued)

Table 2.3 (continued)

Country	InQua on indicators of TQ										Achievement on TQ and InQua										Indirect effects of TQ × InQua on Student achievement						Level 1 variables	
	1	2	3	4	5	6	7	8	9	10	11	1 × 6	2 × 6	3 × 6	4 × 6	5 × 6	1 × 6	2 × 6	3 × 6	4 × 6	5 × 6	Girl	Books					
<i>South-east Asian countries</i>																												
Hong Kong	-0.04	-0.05	0.20	0.25	0.33	0.24	0.14	-0.02	-0.08	0.01	0.35	-0.01	-0.02	0.07	0.09	0.12	-0.11	0.15	-0.01	0.00	0.00	0.00	0.00	0.12	-0.11	0.15		
Japan	0.17	-0.10	0.01	0.19	0.15	0.04	0.00	0.04	0.08	0.04	-0.02	0.00	0.00	0.00	0.00	0.00	-0.03	0.29	0.00	0.00	0.00	0.00	0.00	0.00	-0.03	0.29		
Malaysia	-0.25	-0.16	0.47	-0.03	0.08	0.04	0.32	0.07	0.18	0.27	-0.12	0.03	0.02	-0.06	0.00	-0.01	-0.05	0.14	0.03	0.02	-0.06	0.00	0.00	-0.01	-0.05	0.14		
Singapore	-0.01	-0.01	0.11	0.31	0.21	0.00	0.12	0.05	0.11	0.04	-0.01	0.00	0.00	0.00	0.00	0.00	-0.05	0.12	0.00	0.00	0.00	0.00	0.00	0.00	-0.05	0.12		
South Korea	-0.09	-0.08	0.08	0.29	0.06	0.19	0.07	0.25	0.07	0.09	0.07	-0.01	-0.01	0.01	0.02	0.00	-0.05	0.35	-0.01	-0.01	0.01	0.02	0.00	0.00	-0.05	0.35		
Chinese Taipei	0.04	0.06	0.13	0.18	0.16	0.08	-0.00	0.16	-0.03	-0.05	-0.05	0.00	0.00	-0.01	-0.01	-0.01	-0.00	0.33	0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.00	0.33		
Thailand	0.18	0.04	-0.09	0.05	0.36	-0.07	-0.11	0.06	-0.10	0.08	0.01	0.00	0.00	0.00	0.00	0.00	0.11	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.06		
<i>Central and Eastern European countries</i>																												
Croatia	0.03	0.10	-0.11	0.27	0.19	0.13	0.03	-0.12	0.07	-0.02	0.13	0.00	0.01	-0.01	0.04	0.02	-0.10	0.24	0.00	0.01	-0.01	0.04	0.02	0.02	-0.10	0.24		
Czech Republic	0.25	-0.02	-0.08	0.24	0.09	-0.11	0.22	0.03	-0.23	-0.10	0.26	0.07	-0.01	-0.02	0.06	0.02	-0.11	0.32	0.26	-0.01	-0.02	0.06	0.02	0.02	-0.11	0.32		
Hungary	0.02	-0.03	-0.42	0.17	0.08	0.14	0.01	-0.07	0.15	0.01	-0.13	0.00	0.00	0.05	-0.02	-0.01	-0.05	0.37	0.00	0.00	0.00	-0.02	-0.01	-0.01	-0.05	0.37		
Lithuania	-0.08	0.01	-0.10	0.31	0.13	-0.03	0.10	-0.06	-0.19	0.15	0.18	-0.01	0.00	-0.02	0.06	0.02	-0.04	0.23	-0.01	0.00	-0.02	0.06	0.02	0.02	-0.04	0.23		
Poland	0.01	0.11	0.18	0.25	-0.14	0.09	0.07	-0.01	-0.07	0.04	0.01	0.00	0.00	0.00	0.00	0.00	-0.10	0.32	0.00	0.00	0.00	0.00	0.00	0.00	-0.10	0.32		
Romania	-0.05	-0.19	0.09	0.07	0.09	0.09	0.21	-0.12	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00	-0.04	0.28	0.00	0.00	0.00	0.00	0.00	0.00	-0.04	0.28		
Serbia	-0.16	0.09	-0.00	0.22	0.22	-0.02	0.25	0.05	0.13	0.25	-0.08	0.01	-0.01	0.00	-0.02	-0.02	-0.05	0.22	0.01	-0.01	0.00	-0.02	-0.02	-0.02	-0.05	0.22		
Slovak Republic	0.16	-0.01	0.06	0.15	0.10	0.01	0.06	-0.07	0.08	-0.12	0.02	0.00	0.00	0.00	0.00	0.00	-0.08	0.34	0.00	0.00	0.00	0.00	0.00	0.00	-0.08	0.34		
Slovenia	0.25	0.13	0.12	0.39	0.27	0.26	0.13	0.12	-0.09	0.09	-0.25	-0.06	-0.03	-0.10	-0.07	-0.09	-0.09	0.32	-0.03	-0.03	-0.03	-0.10	-0.07	-0.07	-0.09	0.32		
<i>Scandinavian countries</i>																												
Denmark	-0.08	0.13	0.12	0.20	0.19	0.05	0.05	-0.03	-0.03	-0.28	0.08	-0.01	0.01	0.01	0.02	0.02	-0.06	0.31	-0.01	0.01	0.01	0.02	0.02	0.02	-0.06	0.31		
Finland	0.01	-0.08	0.03	0.11	0.39	0.03	0.10	0.06	0.13	0.04	0.16	0.00	-0.01	0.00	0.02	0.06	-0.07	0.25	0.00	-0.01	0.00	0.02	0.06	0.06	-0.07	0.25		
Norway	-0.01	0.22	-0.07	0.08	0.32	0.10	-0.02	0.10	-0.07	-0.02	0.17	0.00	0.04	-0.01	0.01	0.05	-0.07	0.24	0.00	0.04	-0.01	0.01	0.05	0.05	-0.07	0.24		
Sweden	-0.25	-0.09	0.02	0.18	0.22	0.09	0.21	-0.16	-0.05	0.06	0.07	-0.02	-0.01	0.00	0.01	0.02	-0.08	0.32	-0.02	-0.01	0.00	0.01	0.02	0.02	-0.08	0.32		

(continued)

Table 2.3 (continued)

Country	InQua on indicators of TQ					Achievement on TQ and InQua					Indirect effects of TQ × InQua on Student achievement					Level 1 variables		
	1	2	3	4	5	6	7	8	9	10	11	1 × 6	2 × 6	3 × 6	4 × 6	5 × 6	Girl	Books
<i>Western European countries</i>																		
Germany	0.05	-0.01	-0.19	0.34	0.20	0.13	-0.13	0.27	0.11	-0.12	0.03	0.00	0.00	-0.01	0.01	0.01	-0.09	0.35
Italy	0.19	-0.06	-0.00	0.30	0.00	0.05	0.12	0.04	0.14	0.07	-0.07	-0.01	0.00	0.00	-0.02	0.00	-0.09	0.17
Netherlands	-0.03	0.18	0.12	0.13	-0.03	0.03	0.05	-0.01	-0.21	-0.03	0.02	0.00	0.00	0.00	0.00	0.00	-0.07	0.25
Portugal	-0.09	0.13	-0.01	0.26	0.31	0.09	-0.10	-0.09	-0.21	0.27	0.05	0.00	0.01	0.00	0.01	0.02	-0.08	0.25
Spain	-0.07	-0.06	-0.05	0.54	0.31	0.32	0.09	-0.12	0.31	0.05	-0.01	0.00	0.00	0.00	-0.01	0.00	-0.10	0.25
<i>English-speaking countries</i>																		
England	0.09	-0.04	-0.16	-0.11	0.12	0.31	0.12	-0.12	-0.26	0.14	-0.13	-0.01	0.01	0.02	0.01	-0.02	-0.05	0.35
Ireland	-0.06	0.25	-0.11	0.10	0.10	0.11	-0.02	0.05	-0.05	0.19	-0.27	0.02	-0.07	0.03	-0.03	-0.03	-0.10	0.36
Northern Ireland	-0.19	0.13	0.06	0.20	0.29	0.11	0.00	-0.04	0.18	-0.05	0.13	-0.02	0.02	0.01	0.03	0.04	-0.03	0.39
New Zealand	-0.11	-0.09	0.12	0.06	0.37	-0.01	0.07	-0.08	0.17	0.14	0.02	0.00	0.00	0.00	0.00	0.01	-0.02	0.27
USA	0.05	0.01	-0.09	0.19	0.16	0.03	0.01	0.05	-0.02	0.07	-0.12	-0.01	0.00	0.01	-0.02	-0.02	-0.08	0.26

Note Standardized regression coefficients are reported at the between-level with significant relations ($p < .10$) indicated in bold. Instructional quality predicted by 1 = number of years of teaching experience, 2 = level of formal education (degree), 3 = major focus of teacher education, 4 = professional development, and 5 = preparedness; mathematics achievement of students at grade four aggregated at the classroom level predicted by 6 = number of years of teaching experience, 7 = ISCED level of formal education (teacher education degree), 8 = major, 9 = professional development, and 10 = preparedness; 11 = relation of instructional quality to mathematics achievement of students at grade four aggregated on the classroom level. At level 1, relations of gender and books at home to mathematics achievement of students at grade four were controlled for; here significant relations ($p < .05$) are indicated in bold

The importance of controlling for students' socioeconomic background was demonstrated by the strong relationship between the number of books at home and student achievement. In 39 out of the 47 countries, students who reported more books also had a higher mathematics score. This applied to all European, English-speaking and South-East Asian countries. In contrast, socioeconomic background was not significant in the African countries. Gender differences were evident in 28 countries, particularly in European (17 out of 18) and Latin America (2 out of 2) countries, and these differences unanimously favored boys. In contrast, Western Asian/Arabian (2 out of 12) and African (1 out of 3) countries were much less affected by gender inequalities, and when these were present in these countries, the differences favored girls.

2.5 Discussion

TIMSS data provide a unique opportunity to link student outcomes with teacher and instructional characteristics because they collect data from intact classrooms. The good fit of our model to the data within countries and across countries can be regarded as evidence that the model was well specified and that important teacher predictors of student achievement were selected. However, it seems to be important to distinguish between predictors that can be characterized as being more proximal or distal, respectively, to instructional quality or student achievement. Initial teacher education may have happened decades ago in case of experienced teachers, and programs may have been very different at that time compared to current teacher education programs (Wang et al. 2003). Teachers' initial education is in this manner an example of a teacher characteristic which, at least for a large group of teachers, is distal to the other variables included in the model, and moreover, likely confounded with other omitted variables. Taken together this makes it difficult to identify a systematic relationship between features of mathematics teacher education and instructional quality or student achievement.

Professional development activities taken during the past 2 years and teachers' self-efficacy are, in contrast, much more closely related to what happens currently in classrooms. The analysis presented demonstrates that teachers' participation in PD activities and their self-efficacy are both significantly associated with grade four students' mathematics achievement, both in the pooled international model and within a high number of countries. This finding therefore extends research-based knowledge by providing evidence for the generalizability of the influences of self-efficacy (Bandura 1986) and PD (Timperley et al. 2007) across widely different educational contexts.

However, for all other variables in the model, a large variation between the countries was observed and universal relationships with instructional quality and students' achievement were generally not observed. Teachers teach in a context of structures, policies and expectations. Scheerens (2007) separated these conditions into entities that were more or less "given" antecedents (such as population

characteristics or general valuation of education and teachers) and conditions that were more malleable by policy (such as level and type of decentralization or accountability arrangements). These differences in conditions may affect both the between-country and the within-country variability in teacher quality and instructional quality, and also the relationships between these concepts and students' learning outcomes. The TEDS-M study showed that in some countries teacher education is nationally standardized, while in other systems teacher education can be highly decentralized (Ingvarson et al. 2013). Furthermore, in some countries, teachers are trusted by both the public and their employers, who grant them more or less full autonomy in how they implement the curriculum and the instruction. In other countries, teachers will be firmly placed in a hierarchical system, with less freedom to influence the curriculum and instruction, in the extreme case with prescribed and detailed lesson plans.

Correspondingly, for all variables of teacher quality included in this chapter, we observed a noticeably large variation across countries. One potential consequence of such variation is that, in systems where teachers are fully autonomous individuals with responsibility for developing and implementing instruction, a relatively large within-country variation in instructional quality is possible, while systems characterized by teachers being provided with more or less prescribed lesson plans would likely have fewer degrees of freedom for some of the components typically included in instructional quality. In our models, the observed differences in direct relations of several variables describing teacher quality to instructional quality may be a reflection of this wider "ecology" of teaching. Taken together, this variation illustrates how international studies may use systematic differences in conditions and policies for teaching in order to at least provide examples of how alternative policies work in other settings, although, of course, such interpretations should be done with care since the wider cultural context of education represents a range of potentially very influential omitted variables.

In relation to this, it is also worth discussing how the educational system caters for specialized or generalized teachers of mathematics at grade four. It is reasonable to assume that in more or less all countries teachers in secondary schools will have a specialization in one or a few subjects. However, in primary schools, at least in the first years, there will be a larger between-country variation in the degree to which teachers have a general versus a specialized teacher education. Teachers with general qualifications will by default have a broader background with less in depth subject knowledge. This is a variation at the system level, which to a large degree was observed for the two proxy measures of teachers' educational background.

2.6 Limitations of the Study

One limitation of our study was its reliance on cross-sectional data. In order to study the effect of teacher and instructional quality on student achievement, and not the least, in order to study the possible mediation of teacher qualities by

instructional quality, use of data from experimental or longitudinal designs would be preferred. Follow-up studies with improved designs are urgently needed. Since the international studies are repeated at regular intervals, it should be possible to have repeated measures at country level in later surveys.

However, this would imply measures remain unaltered, which we would not recommend given another limitation of our study; the unsatisfactory quality of some of the measures used. This is primarily an issue regarding the measure of instructional quality used in this analysis. This measure was based on items in the general part of the teacher questionnaire. Consequently, the questions did not include explicit references to the subject of mathematics. In several countries, a teacher of grade four mathematics will also be teaching the same class other subjects. It may be that some of the teachers responded to this list of questions without having mathematics instruction in mind, which may cause validity problems (Schlesinger and Jentsch 2016).

There were other related measures which could have been used, and which are used in the analyses in other chapters in this book. A set of questions in the mathematics specific part of the teacher questionnaire also asked teachers to report their instructional activities in mathematics. However, these questions reflect surface characteristics of teaching practices, and did not correspond to the theoretical framework of instructional quality applied in this book, which is based on current research on instructional quality. A measure based on students' responses could also have been used. However, given the low age of the students in grade four, we opted to rely on the teachers' reports. Improvements in the instructional quality measures to better include recent research in this area (in particular the work done by the Klieme group; see for example Decristan et al. 2015; Rakoczy et al. 2010) seem to be urgently needed.

A third feature of our analysis that may be regarded as a minor limitation is, given the limited sample size of teachers and classrooms in many countries, item parcels were applied instead of single items. This leads to some loss of information. Given that the reliability of most parcels was reasonably high, the grouping of items into parcels can be assumed to represent a minor reduction of information with only small consequences for the analysis.⁵ However, given that there are potentially differential relationships between the three indicators and student achievement across countries and within countries, the research questions of this paper may also merit reinvestigation at the item- or indicator level.

There were other dimensions in the TIMSS questionnaire gauging teacher characteristics that were found to be of relevance for students' achievement. These measures were omitted from this analysis for several reasons. Firstly, for some of them it was not possible to confirm metric measurement invariance (this applied, for

⁵This argument is not directly applicable to the parcels representing the three theoretically-based aspects of InQua, since they consist of two variables only. The total internal consistency for the manifest variable using all six variables as compared to the variable using the three item parcels is only a fraction higher, 0.65 as compared to 0.61, which demonstrates that the parcels function almost equivalent to the single items.

example, to teacher motivation) and, secondly, their inclusion would have introduced a risk of multicollinearity. In addition, as a two-level multi-group analysis framework was applied, keeping the model simple was a necessary priority. It should be noted that the final choice of indicators of teacher qualities in our model did not fully match the dimensions cited most often in contemporary teacher effectiveness studies. For example, TIMSS did not include measures of the teachers' actual knowledge and skills to teach mathematics (see for example, Blömeke et al. 2012; Tatto et al. 2012).

2.7 Conclusions and Recommendations

The results of the present study clearly support the relevance of teacher quality for instructional quality and for educational outcomes. Instructional quality and mathematics achievement were significantly related to several teacher characteristics selected on the basis of contemporary research and, their availability within the TIMSS 2011 data. Patterns emerged across countries and cultures, both with respect to the absolute level of some constructs and the relations between teacher quality, instructional quality and outcomes. Some characteristics were more regionally relevant. However, although the model fits the data from the majority of countries, the structural relations represented by this model do not provide a universal model.

The lack of a universally applicable model is obvious: significant research is needed to clarify the generalizability of these results. One particular topic for research concerns the relevance of initial teacher education, which several times was found to be non-significant, replicating previous findings from other cross-sectional surveys (see for instance, Nordenbo et al. 2008). This could be related to the fact that teacher education has changed profoundly in many countries over the last decades (Wang et al. 2003; Darling-Hammond and Lieberman 2012). It is reasonable to assume that characteristics of students recruited into the profession have changed over time. Access to teacher education may historically have been more selective and restricted to students with relatively higher marks from secondary education. Also, the demand and provision of deep mathematical knowledge in the teacher education may have changed as teacher education has been reformed at specific points in time. Teacher experience and formal qualifications as measured in TIMSS are therefore likely confounded with other characteristics not included in our model. Distinguishing between age cohorts would provide important information, but this was not feasible with the current data set given the already rather small sample size. One solution for future surveys could be to include larger samples of teachers and classrooms in countries where changes in some of these confounding characteristics can be described and included in the model from other sources.

We have chosen to focus on cognitive outcomes in this chapter, given that other chapters in this book cover student motivation or bullying as outcomes. It is important to recall that outcomes of education are multi-dimensional and that

cognitive and motivational variables are both important. Evidence suggests that motives are often positively related to cognitive learning outcomes and that motivation supports cognitive learning long term (Benware and Deci 1984; Grolnick and Ryan 1987). Reducing schooling to cognitive outcomes would therefore be a shortcoming. In further studies of how teacher quality and instructional quality relates to outcomes, it would therefore be relevant to include also students' motivation and interest as dependent variables in one and the same model.

Another major recommendation for future studies based on our experience with analyzing the complex relationship between teacher quality, instructional quality and student outcomes, is that future surveys need invest in the development of improved measures of instructional quality. A long-standing controversy exists whether teacher or student ratings describe instructional quality more reliably and/or more validly (Desimone 2011; Schlesinger and Jentsch 2016; Wagner et al. 2015). Current research understanding suggests that the correlation between these two approaches is only moderate and that their relation with student achievement differs. This may reflect not only that students and teachers perceptions differ, but also that the measures represent slightly different aspects of the instructional activities taking place in the classroom. In general, we would therefore recommend that measures of instructional quality, in line with the current practice in the IEA studies, include both types of sources to develop measures of the quality of the instructional activities.

However, the current measures in both the teacher and the student questionnaires fail to fully represent the depth and breadth of the concept of instructional quality. The three core aspects in the measure of InQua that we applied (clarity of instruction, cognitive activation, and supportive climate) are represented by two items only. Each of these aspects represents separate and relatively broad and many-faceted constructs by themselves, which should be reflected in future studies. Furthermore, classroom management is a vital dimension of instructional quality not included in the generic teacher questionnaire. And not the least, as discussed already, the construct used in this chapter is based on generic questions, while it would provide more fidelity to the analysis if a measure specific to the quality of the mathematics lessons had been applied. In future surveys, priority should rather be given to the improvement of context sensitive measures of instructional quality. Frequency of different specific activities may not represent an ideal way to assess the quality with which these activities are carried through. Some actions probably occur relatively often in high quality teaching (for instance, summarizing at the end of the lecture), while others would probably need to be used less often in order to represent an optimal quality (for instance, working on problems with no obvious solution). In summary, new improved measures of InQua should:

- (1) reflect both students' and teachers' experiences,
- (2) have a broader scope, including the four core components, clarity of instruction, cognitive activation, classroom management, and supportive climate,

- (3) cover each of these aspects in depth by including separate, but related, constructs,
- (4) be subject-specific rather than generic, and
- (5) include scales aimed at capturing qualities of various activities.

References

- Alexander, R. (2001). *Culture and pedagogy: International comparisons in primary education*. West Sussex: Wiley-Blackwell.
- Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: New developments and techniques*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*, 133–180.
- Benware, C., & Deci, E. L. (1984). Quality of learning with an active versus passive motivational set. *American Educational Research Journal*, *21*, 755–765.
- Bishop, A. (2004). Mathematics education in its cultural context. In T. P. Carpenter, J. A. Dossey, & J. L. Koehler (Eds.), *Classics in mathematics education research*. National Council of Teachers of Mathematics: Reston, VA.
- Blömeke, S., & Delaney, S. (2012). Assessment of teacher knowledge across countries: A review of the state of research. *ZDM – The International Journal on Mathematics Education*, *44*, 223–247.
- Blömeke, S., & Kaiser, G. (2012). Homogeneity or heterogeneity? Profiles of opportunities to learn in primary teacher education and their relationship to cultural context and outcomes. *ZDM*, *44*, 249–264.
- Blömeke, S., Suhl, U., Kaiser, G., & Döhrmann, M. (2012). Family background, entry selectivity and opportunities to learn: What matters in primary teacher education? An international comparison of fifteen countries. *Teaching and Teacher Education*, *28*, 44–55.
- Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*, 416–440.
- Boyle, B., Lamprianou, I., & Boyle, T. (2005). A longitudinal study of teacher change: What makes professional development effective? Report of the second year of the study. *School Effectiveness and School Improvements*, *16*, 1–27.

- Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005–1018.
- Clotfelter, Ch., Ladd, H., & Vigdor, J. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*, 673–82.
- Cochran-Smith, M., & Zeichner, K. M. (2005). *Studying teacher education: The report of the AERA panel on research and teacher education*. Mahwah, NJ: Erlbaum.
- Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. London: Routledge.
- Darling-Hammond, D. & Lieberman, A. (Eds.). (2012). *Teacher education around the world: changing policies and practices*. Teacher Quality and School Development Series. Abingdon, Oxon: Routledge.
- Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., & Fauth, B., et al. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal, 52*, 1133–1159.
- Desimone, L. M. (2011). A primer on effective professional development. *Phi Delta Kappan, 92*, 68–71.
- Foy, P., Arora, A., & Stanco, G. M. (Eds.). (2013). *TIMSS 2011 user guide for the international database. Supplement 1: International version of the TIMSS 2011 background and curriculum questionnaires*. Chestnut Hill, MA/Amsterdam: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College & International Association for the Evaluation of Educational Achievement (IEA).
- Goldsmith, L., Doerr, H., & Lewis, C. (2014). Mathematics teachers' learning: A conceptual framework and synthesis of research. *Journal of Mathematics Teacher Education, 17*, 5–36.
- Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology, 52*, 890–898.
- Guskey, T. R. (2000). *Evaluating professional development*. Thousand Oaks, CA: Corwin Press.
- Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*, 371–406.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal, 6*(1), 1–55.
- Ingvanson, L., Schulle, J., Tatto, M. T., Rowley, G., Peck, R., & Senk, S. L. (2013). *An analysis of teacher education context, structure, and quality-assurance arrangements in TEDS-M countries: Findings from the IEA teacher education and development study in mathematics (TEDS-M)*. Amsterdam: International Association for the Evaluation of Educational Achievement.
- Joncas, M., & Foy, P. (2012). Sample design in TIMSS and PIRLS. In I. V. S. Mullis & M. O. Martin (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Kee, A. N. (2012). Feelings of preparedness among alternatively certified teachers: What is the role of program features? *Journal of Teacher Education, 63*, 23–38.
- Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal, 49*, 568–589.
- Kyriakides, L., Creemers, B. P. M., & Antoniou, P. (2009). Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25*, 12–23.
- Leung, F. K. S. (2006). Mathematics education in East Asia and the West: Does culture matter? In F. Leung, K.-D. Graf, & F. Lopez-Real (Eds.), *Mathematics education in different cultural traditions: A comparative study of East Asia and the West, 13th ICMI study*. New York: Springer.

- Little, T., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel? Exploring the question, weighing the merits. *Structural Equation Modeling*, 9, 151–173.
- Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschof, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.
- Muthén, L. K., & Muthén, B. O. (2008–2012). *Mplus user's guide* (7th Edn.). Los Angeles, CA: Muthén & Muthén.
- Nordenbo, S. E., Sjøgaard Larsen, M., Tiftikiçi, N., Wendt, R. E., & Østergaard, S. (2008). *Teacher competences and pupil learning in pre-school and school: A systematic review carried out for the ministry of education and research, Oslo*. Copenhagen: Danish Clearinghouse for Educational Research, School of Education, University of Aarhus.
- Pajares, F. (1996). Self-efficacy beliefs in achievement settings. *Review of Educational Research*, 66, 543–578.
- Rakoczy, K., Klieme, E., Lipowsky, F., & Drollinger-Vetter, B. (2010). Strukturierung, kognitive Aktivität und Leistungsentwicklung im Mathematikunterricht. *Unterrichtswissenschaft: Zeitschrift für Lernforschung*, 38, 229–246.
- Sammons, P. (2009). The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools. *School Effectiveness and School Improvement*, 20, 123–129.
- Scheerens, J. (2007). *Conceptual framework for the PISA 2009 background questionnaires*. Twente, The Netherlands: University of Twente.
- Scheerens, J. (2013). The use of theory in school effectiveness research revisited. *School effectiveness and school improvement*, 24, 1–38.
- Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.
- Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM*. doi:10.1007/s11858-016-0765-0.
- Schmidt, W. H., McKnight, C. C., Valverde, G. A., Houang, R. T., & Wiley, D. E. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in school mathematics*. Dordrecht: Kluwer.
- Schwille, J., Ingvarson, L., & Holdgreve-Resendez, R. (Eds.). (2013). *TEDS-M encyclopedia: A guide to teacher education context, structure, and quality assurance 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.
- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the last decade: Role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77, 454–499.
- Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and applied multilevel analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Stark, J. S., & Lattuca, L. R. (1997). *Shaping the college curriculum: Academic plans in action*. Boston, CT: Allyn and Bacon.
- Tatto, M. T., Schwille, J., Senk, Sh, Rodriguez, M., Bankov, K., & Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics: First findings*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement.
- Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration*. Wellington: Ministry of Education.

- Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*, 202–248.
- Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis of comparative research*. Thousand Oaks, CA: Sage.
- Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2015). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*. Retrieved from <http://psycnet.apa.org/psycinfo/2015-43248-001/>.
- Wang, A. H., Coleman, A. B., Coley, R. J., & Phelps, R. P. (2003). *Preparing teachers around the world (Policy Information Report)*. Princeton, NJ: ETS.
- Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*, 89–122.
- Wilson, S., Floden, R. & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations: Research report prepared for the US Department of Education*. Washington, DC: University of Washington Center for the Study of Teaching and Policy, Seattle.



The Relations Among School Climate, Instructional Quality, and Achievement Motivation in Mathematics

Ronny Scherer and Trude Nilsen

Abstract Instructional quality is considered to be an important classroom variable, as it is significantly related to student achievement and motivation in mathematics. Existing studies in educational effectiveness furthermore identified a positive relation between instructional quality and school climate, suggesting that the school environment plays a significant role in teachers' instructional practices. In order to bring together these two core findings, the relations among different aspects of school climate, instructional quality, and students' achievement motivation for the TIMSS 2011 grade eight mathematics data sets comprising 50 countries are investigated. In particular, the role of instructional quality as a potential mediator between school climate and student motivation is examined, thereby focusing on three aspects of school climate (emphasis on academic success, safety, and order in schools) and three aspects of achievement motivation (self-concept, intrinsic value, and extrinsic value). In general, there was a significant positive relation between instructional quality and achievement motivation at the classroom level in mathematics; in some countries, a partial mediation of instructional quality between school climate and achievement motivation was apparent. Four main patterns of relations occurred. These findings are discussed with respect to implications for educational effectiveness research.

Keywords School climate · Instructional quality · Student motivation · Two-level structural equation modeling mediation models · Trends in Mathematics and Science Study (TIMSS) 2011

R. Scherer (✉)

Faculty of Educational Sciences, Centre for Educational Measurement at the University of Oslo (CEMO), Oslo, Norway
e-mail: ronny.scherer@cemo.uio.no

T. Nilsen

Department of Teacher Education and School Research, University of Oslo, Oslo, Norway
e-mail: trude.nilsen@ils.uio.no



3.1 Rationale

Mathematics may be said to be at the heart of all science, technology, engineering, and mathematics (STEM) subjects. Motivating students to study these subjects is vital for a sustainable development within areas such as technology, economy, health, and environment. Yet, at the same time, an international concern for the decline of students' participation in STEM-related studies and careers has been raised (OECD 2014a). This concern seems to be rooted in a decrease of their STEM motivation (OECD 2014a). It is therefore important to motivate students for mathematics and for pursuing a career in science (Simpkins et al. 2006).

According to the widely accepted expectancy-value theory of achievement motivation, at least three motivational aspects are important for students' choices and performance in STEM areas: (a) self-beliefs, (b) intrinsic value, and (c) extrinsic value (Eccles and Wigfield 2002; Wigfield and Eccles 2000). Whereas (a) refers to students' beliefs in their capabilities, thereby reflecting their expectations of academic success, (b) and (c) are related to the subjective value assigned to subjects or tasks. Students may be driven by one or more of these aspects of motivation; either way, it is pertinent to identify factors that may promote these aspects of motivation and that lie within the power of the school.

Along with this challenge of identifying the motivational factors comes the question of how instruction and the school environment may contribute to student motivation in STEM subjects. According to previous research, teachers and their instruction matter more to student learning and motivation than any other school factor (Baumert et al. 2010; Creemers and Kyriakides 2008). The most important classroom variable is likely teachers' instructional quality, which affects both student achievement and motivation (Blömeke et al. 2013; Creemers and Kyriakides 2008; Fauth et al. 2014). Providing high quality instruction necessitates a safe and orderly school climate with a high priority for academic success (Thapa et al. 2013). Effective teaching is therefore challenged under conditions where teachers and students do not feel safe, where no order exists, and where academic success receives low priority. A healthy school climate consequently is important for student learning and motivation (Wang and Degol 2015).

In summary, a review of previous research indicates that, while instructional quality is important for student learning and motivation, school climate may contribute with ideal conditions for high quality instruction and hence promote learning and motivation. There are several aspects of school climate and motivation though, and the question of how instructional quality is related to these different aspects is complex (Good et al. 2009). This gap in research could be due to the extensive focus on achievement as a learning outcome as compared to motivation. Few studies have investigated the relations between school climate, instructional quality, and student motivation; these studies are almost exclusively focused on single-country analyses (Good et al. 2009; Seidel and Shavelson 2007; Wang and Degol 2015).

As a consequence, this chapter aims to address this research gap by investigating the relations among different aspects of school climate, instructional quality, and achievement motivation. Including all countries that participated in TIMSS 2011 provides a unique opportunity to investigate these relations across countries with widely different cultures from all continents.

3.2 Theoretical Framework

In this section, we first review the conceptualization of the three core constructs under investigation: school climate, instructional quality, and achievement motivation. We then present selected previous research on their relations.

3.2.1 *School Climate*

School climate is a broad concept that includes many dimensions (Thapa et al. 2013). Although it is defined somewhat differently across fields, certain key aspects have been found to be important to student learning. One of these aspects refers to academic climate, which is significantly positively related to student achievement and motivation (Wang and Degol 2015). Even though academic climate commonly refers to the extent to which learning and academic success is emphasized, there exists no consensus on its specific conceptualization. Hoy et al. (2006) referred to this aspect as academic optimism, a concept that reflects academic emphasis, collective efficacy, and faculty trust in parents and students. Together and individually, these three constructs have been found to be positively related to student learning (Goddard 2002; Hoy and Tschannen-Moran 1999; Hoy et al. 2006). Although the measurement of these constructs differed in further studies, this relation has been largely confirmed (Kythreotis et al. 2010; Martin et al. 2013; McGuigan and Hoy 2006; Nilsen and Gustafsson 2014),

In the context of TIMSS 2011, academic climate is represented and measured by the school emphasis on academic success (SEAS) scale. The underlying construct of SEAS has been found to be of great importance for students' learning outcomes and changes in performance across a number of countries (Martin et al. 2013; Nilsen and Gustafsson 2014). Conceptually, SEAS reflects the shared beliefs, capabilities, and trust among the members of the school institution (namely students, parents, teachers, and school leaders; Hoy et al. 2006; Nilsen and Gustafsson 2014). Among other aspects, SEAS comprises schools' trust in parents and students on the one hand and teachers' expectations for students' success on the other hand (Martin et al. 2013).

Another key aspect of school climate relates to a safe and orderly climate (Thapa et al. 2013; Wang and Degol 2015). Safety and order in schools refer to the degree of physical and emotional security, along with an orderly disciplinary climate

(Goldstein et al. 2008; Gregory et al. 2012; Wang and Degol 2015; Wilson 2004). Both safety and order are positively associated with student outcomes in a number of countries (Martin et al. 2013).

3.2.2 Instructional Quality

As detailed in Chap. 1, teachers' instructional quality comprise a number of aspects that have been shown to be highly important for student learning outcomes (Baumert et al. 2010; Creemers and Kyriakides 2008; Fauth et al. 2014; Good et al. 2009; Hattie 2009; Kunter et al. 2013; Seidel and Shavelson 2007). In the context of TIMSS 2011, students' ratings of instructional quality refer to aspects of supportive climate and clarity. We realize that this representation limits the rather broad concept of instructional quality to these two dimensions; yet, they are powerful indicators.

It is noteworthy that most studies investigate relations between instructional quality and achievement, while fewer include achievement motivation as an educational outcome (Fauth et al. 2014; Good et al. 2009).

3.2.3 Achievement Motivation

For decades, there has been an increasing concern for students' limited motivation in STEM subjects (NSF [US National Science Foundation] 2012; OECD 2007). Findings from TIMSS show that students' motivation for mathematics declines between grades four and eight. Moreover, previous studies have found significant differences with respect to the influence of gender on motivation in mathematics (Meece et al. 2006a, b), pointing to the necessity of accounting for gender in models of achievement motivation in mathematics. However, these findings vary across the different aspects of motivation (Wigfield et al. 2002).

According to the expectancy-value theory of achievement motivation proposed by Wigfield and Eccles (2000), there are some factors that directly influence student performance: expectation of success, interest-enjoyment value, attainment value, utility value, and cost. Wigfield and Eccles (2000) argued that students' expectations of success refer to their self-concept, reflected by the degree to which students believe they perform well in mathematics. The interest-enjoyment value refers to students' enjoyment and interest in a task or subject; Wigfield and Eccles (2000) claimed that this construct can be considered to be the intrinsic value or motivation of a subject or task (Deci and Ryan 1985; Harter 1981). Utility value refers to students' future career goals and aspirations, while attainment value reflects the personal importance of, for instance, mathematics. The last two factors reflect what is commonly referred to as 'extrinsic motivation'. In the current investigation, we will use the term extrinsic value. Costs reflect the negative aspects of motivation, such as performance anxiety and fear of both failure and success.

A number of studies have confirmed the importance of self-concept, intrinsic, and extrinsic value for students' career choices and performance (see Bandura 1997; Eccles and Wigfield 2002; Pintrich and Schunk 2002). As a consequence, we consider these aspects of achievement motivation as outcome variables in the current study.

3.3 Review of the Appropriate Level of Analysis

One question that arises with the measurement of the mentioned constructs concerns the appropriate level of analysis. In fact, given that both instructional quality and school climate are most often assessed using student or teacher ratings of the classroom or school environment, variation in these ratings may occur at different levels (namely student, classroom, school, or even country level; Klieme 2013). In order to make clear-cut decisions on the analysis level, a thorough review of the specific research questions is needed (Lüdtke et al. 2009). In the context of teacher effectiveness, research studying how the characteristics of the learning environment affect students' educational outcomes, such as their achievement, motivation, and self-beliefs, is the main focus (Klieme 2013; Lüdtke et al. 2009). Marsh et al. (2012) argued that the classroom or school level is the most appropriate in such scenarios. Nevertheless, as individual ratings of the learning environment may still vary and have a distinct meaning at the individual level, controlling for within-level variation is necessary (Lüdtke et al. 2009). One approach that has proven to be effective in modeling such situations refers to multilevel structural equation modeling (Scherer and Gustafsson 2015a).

In the current study, both instructional quality and school climate were assessed by individual ratings of students and teachers. We decided to study the relations among the two constructs and achievement motivation at the classroom level in a cross-country multi-group setting for two main reasons. First, the relation between instructional quality and student motivation clearly refers to a scenario in which the effects of the learning environment on student outcomes is the focus. Second, although teacher ratings of school climate may be considered a school-level construct, individual differences in these ratings still have a distinct meaning. In fact, teachers within a school may differ greatly in their perceptions of school climate, depending on their job satisfaction, well-being, status of professional development, and further individual-level factors (OECD 2014b; Wang and Degol 2015). Moreover, as teachers are the initiators of instructional practices in classrooms, their individual perceptions of the existing school climate are more important for their instruction than teachers' shared perceptions in a school. This finding has been supported by the results of the OECD Teaching and Learning International Survey (TALIS) 2013, which showed significant relations between individually perceived school climate and classroom instruction (OECD 2014b). As a consequence, we report the results at the classroom level.

3.4 Research Model

Current frameworks for school effectiveness, such as the dynamic model of educational effectiveness, suggest that school climate influences both instructional quality and learning outcomes (Creemers and Kyriakides 2010); it creates the premise and foundation for instruction and learning (Thapa et al. 2013).

While only few studies have investigated the relations between school climate, instructional quality, and achievement motivation, several studies have pointed either to the importance of school climate for student motivation (Wang and Degol 2015) or to the importance of instructional quality for motivation (Fauth et al. 2014; Wagner et al. 2015). However, rarely have all aspects of school climate been investigated in concert (Wang and Degol 2015), and rarely have all aspects of motivation been investigated in relation to school climate or instructional quality. Hence, the relations between these concepts remain obscure.

On the basis of our literature review, we hypothesized that school climate, instructional quality, and achievement motivation were related (Fig. 3.1). The proposed relations at the classroom/teacher level are based on the core assumption that a positively perceived school climate is a prerequisite for creating meaningful instruction, which increases students' motivation to learn (Morin et al. 2014). To account for this assumption, we examine the mediating role of instructional quality. In this regard, we notice that instructional quality was measured by students' reports (representing 'perceived instructional quality'), which were aggregated to the

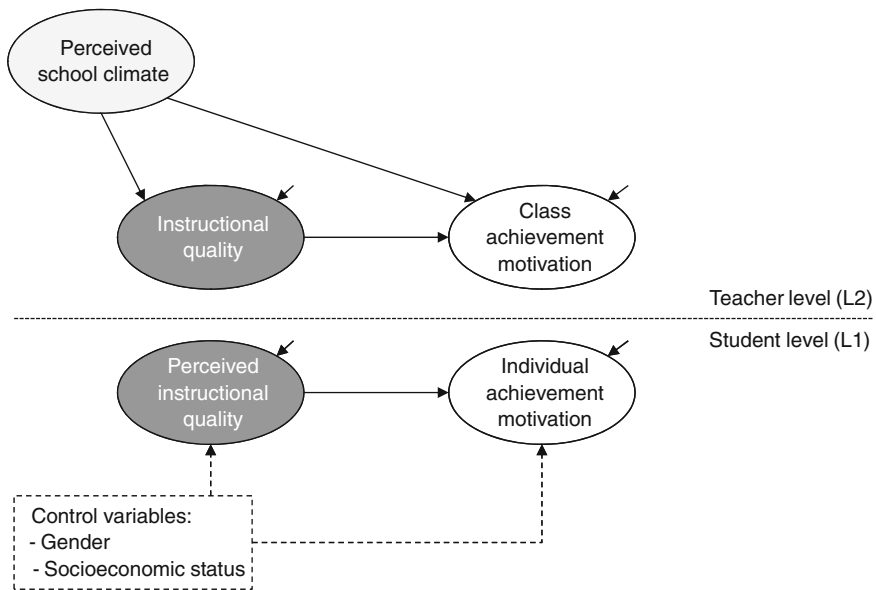


Fig. 3.1 Proposed research model, describing the relations among school climate, instructional quality, and achievement motivation

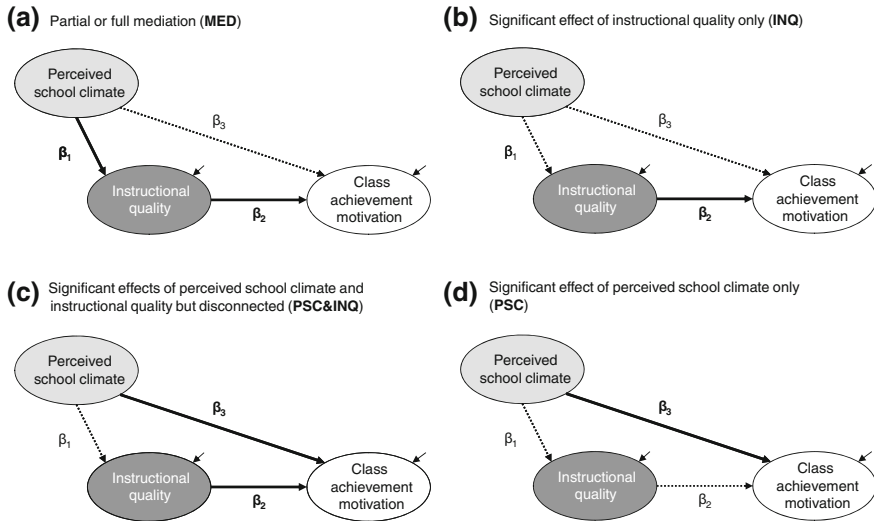


Fig. 3.2 The four scenarios considered in the proposed research model. *Note* Dashed lines represent insignificant regression coefficients ($p > 0.05$). The relation between school climate and instructional quality is named β_1 , the relation between instructional quality and motivation is named β_2 , and the relation between school climate and motivation is named β_3

classroom level ('instructional quality') by means of multilevel structural equation modeling.

However, due to gaps in previous research, the roles played by different aspects of school climate and achievement motivation are unclear. For instance, in some countries, SEAS may predict self-concept to a larger degree than other motivational constructs; at the same time, SEAS may be more important for extrinsic value in other countries. Different combinations of the aspects of school climate and motivation may therefore occur across countries. As a consequence, our research model considered three different aspects of school climate (i.e., SEAS, safety, and order in schools) and three aspects of achievement motivation (i.e., self-concept, intrinsic value, and extrinsic value).

As relationships between these factors and their statistical significance may also differ across countries, differing patterns of relations may result (see Fig. 3.2). These patterns of relations are described by four proposed model scenarios.

1. *Model MED*. The relation between teachers' perceived school climate and achievement motivation may be partially or fully mediated by instructional quality. In this scenario, there is a direct and significant link between school climate and instructional quality perceptions, suggesting that a positive and academically oriented school environment in which there is safety and order may foster a higher quality of instruction that motivates students and strengthens their self-beliefs and beliefs about the importance of mathematics for work and their future. This model proposes a (partial) mediation mechanism between the

three concepts of school climate, instructional quality, and achievement motivation (Deemer 2004).

2. *Model INQ*. There are contexts and situations in which only instructional quality is significantly related to achievement motivation; yet, perceived school climate has no effect. In these scenarios, perceived school climate and instruction are disconnected, indicating that the way teachers organize their instruction in mathematics classrooms is independent of how they perceive, for instance, an orientation toward academic success in the school environment.
3. *Model PSC&INQ*. In some countries, both perceived school climate and instructional quality are significantly related to motivational outcomes, although the two constructs are disconnected. On the one hand, this suggests that teacher perceptions of school climate do not influence the way in which they create learning environments in order to foster students' motivation and self-beliefs. On the other hand, besides the importance of instruction for the motivational outcomes, school climate directly relates to motivational outcomes.
4. *Model PSC*. In some countries and cultures, only perceived school climate, which is an indicator of the actual climate in schools, is significantly related to achievement motivation. In this scenario, school climate seems to be the dominant factor influencing student motivation. In fact, in some countries, a climate of safety and order in schools is considered to be a crucial prerequisite for student learning (Klieme et al. 2009; Mitchell and Bradshaw 2013). Moreover, existing research has suggested that school emphasis on academic success is significantly related to student performance in not only mathematics, but also science achievement tests across almost all TIMSS 2011 participating countries (Mullis et al. 2012; Nilsen and Gustafsson 2014). The relation between perceived school climate and students' motivational outcomes may not necessarily be positive; in fact, negative relations are also likely, particularly with SEAS. More specifically, a strong emphasis on academic success can create highly competitive learning environments that decrease students' motivation and self-beliefs due to a strong performance orientation (Chen and Vazsonyi 2013; Meece et al. 2006a, b). However, this negative relation is not limited to this model.

Our list of scenarios is not exhaustive; further potential scenarios, such as a model in which all relations are insignificant, or a model in which there is only a significant relation between school climate and instructional quality without any connection to achievement motivation, may occur. However, as these last scenarios occurred in only two cases and were of limited substantive relevance, their interpretation was limited.

In light of our considerations, we posed the following research question based on our proposed research model:

To what extent do the different scenarios in the proposed research model, as representatives of different patterns in the relations among school climate, instructional quality, and achievement motivation in mathematics, exist at the teacher/classroom level across the 50 participating TIMSS 2011 grade eight countries (Fig. 3.2)?

3.5 Method

3.5.1 Sample

The total TIMSS 2011 grade eight mathematics student sample, together with their teachers, formed the basis for the present study.¹ This sample comprised $n = 284,899$ students in 12,950 classrooms with an average classroom size of 22 students. All 50 participating TIMSS 2011 countries were included in the analyses. For further details on the sample, please refer to the TIMSS 2011 International Report (Mullis et al. 2012).

3.5.2 Measures

To assess the relations among school climate, instructional quality, and student achievement motivation, we used the existing TIMSS 2011 grade eight student and teacher scales to represent each of the three constructs. These scales and their conceptual underpinnings are briefly described here. For a detailed description of these measures, we refer the reader to the TIMSS 2011 Assessment Frameworks Report (Mullis et al. 2009).

School Climate

In order to capture different aspects of school climate, we followed Wang and Degol's (2015) systematic review and chose three scales as proxies for the construct: School emphasis on academic success, Safety, and Order in schools. Teachers had to rate a number of statements on a four-point agreement scale (0 = I disagree a lot, to 3 = I agree a lot).

SEAS: Teachers' ratings formed the basis for creating a latent variable for SEAS (Martin et al. 2013), because teachers are closer to the classrooms and students than principals. In the teacher questionnaire, teachers were asked to characterize the following within their school: teachers' understanding of and success in implementing the school's curriculum, teachers' expectations for student achievement, parental support for student achievement, and students' desire to do well in school. Hence, the teachers rated SEAS as an aspect of school climate in their schools.

¹In Botswana, Honduras, and South Africa, the TIMSS 2011 study was conducted in grade nine.

Safety in Schools: Teachers' perceptions of safety in schools were indicated by three items ("This school is located in a safe neighborhood", "I feel safe at this school", and "This school's security policies and practices are sufficient").

Order in Schools: Teachers had to evaluate the degree of order and respect in their schools ("The students behave in an orderly manner", "The students are respectful of the teachers").

Instructional Quality

To measure aspects of instructional quality, we chose four items from the 'students engaged in mathematics lessons' scale. One item was removed (namely, "I think of things not related to the lesson"), as it was negatively worded and therefore referred to students' boredom and inattention rather than their perceptions of whether or not teachers engage them; the measurement model based on five items indicated a poor model fit across countries. The remaining four items referred to the clarity of teaching ("I know what my teacher expects me to do", "My teacher is easy to understand"), and the degree to which the teacher engages students to learn mathematics ("I am interested in what my teacher says", "My teacher gives me interesting things to do").² Students had to indicate their agreement with these statements on a four-point scale (0 = I disagree a lot, to 3 = I agree a lot). These items provided valid representations of instruction aimed at engaging students in learning (Scherer and Gustafsson 2015a). These four items were used as indicators of instructional quality in all analyses presented in this chapter. Moreover, as this chapter is not concerned with teachers' perceptions of their instruction, instructional quality was measured at the student level and subsequently aggregated to the classroom level.

Motivation

Measures of student motivation were retrieved from the student questionnaire scales. These scales were developed on the basis of expectancy-value theory of achievement motivation and referred to two main components: (a) students' expectations of success in mathematics, and (b) students' subjective task values (Wigfield and Eccles 2000). The first were indicated by ability beliefs (self-concept); the second were indicated by students' intrinsic and extrinsic values. Students had to indicate the degree to which they agreed to a number of statements

²This aspect of instructional quality is closely related to teachers' motivational support aimed at engaging students to learn in their mathematics lessons.

(0 = I disagree a lot, 1 = I disagree a little, 2 = I agree a little, 3 = I agree a lot). Since methodological research has clearly indicated that a mixture between positively and negatively worded items creates construct-irrelevant multidimensionality in assessments of motivational constructs such as self-concept (see Morin et al. 2015), we decided to omit negatively worded items, as these, by and large, measure a substantively different construct than positively worded items (Marsh and Gouvenet 1989). This has been confirmed for a number of measures that relied on self-ratings (see for example, Davison and Srichantra 1988; Greenberger et al. 2003; Marsh and Gouvenet 1989; Podsakoff et al. 2003; Preckel 2014; Scherer and Gustafsson 2015b). Van Sonderen et al. (2013) further pointed out that simply recoding reversely coded items does not solve the issue of the method bias created by negatively worded items; the hope to correct for potential response bias by introducing such items has not been fulfilled. There is evidence of this method bias in the measurement of motivational constructs in TIMSS (see Bofah and Hannula 2015; Marsh et al. 2013).

Self-concept: Students' self-concept in mathematics was originally assessed by seven items corresponding to the TIMSS 2011 'students confident in mathematics' scale, four of which were negatively worded; as decided, these latter items were deleted. The resultant scale comprised three items: "I usually do well in mathematics", "I learn things quickly in mathematics", and "I am good at working out difficult problems in mathematics". Although this decision limited the overall number of indicators of self-concept, existing research has shown that a three-item scale still provides a reliable and valid measure of students' self-concept (Gogol et al. 2014).

Intrinsic Value: In order to represent the intrinsic task value, we used the TIMSS 2011 'students like learning mathematics' scale, which comprised five items.³ This scale refers to students' enjoyment and interest in learning mathematics (for example, "I enjoy learning mathematics").

Extrinsic Value: This value component of achievement motivation was represented by the 'students value mathematics' scale. The scale comprises differing aspects of the utility and attainment value of learning mathematics and its personal importance: "I think learning mathematics will help me in my daily life", "I need mathematics to learn other school subjects", "I need to do well in mathematics to get into the < university > of my choice", "I need to do well in mathematics to get the job I want", and "I would like a job that involves using mathematics".

³The original scale comprised six items, one of which was negatively formulated. As argued for the measurement of self-concept, we deleted this item to avoid method bias and construct-irrelevant multidimensionality.



Control Variables: Socioeconomic Status

To represent students' socioeconomic status, a variable derived from several items in the student questionnaire (students' ratings of the number of books at home, their parents' highest education and home study supports such as students having their own room and internet connection) was available in the TIMSS 2011 data set (the Home Educational Resources scale). We used the person estimate derived from a partial credit model in the TIMSS 2011 scaling procedure (Martin and Mullis 2012).

Control Variables: Gender

Gender served as another student-level covariate, because some research has suggested that gender differences may exist in student ratings for both achievement motivation and instructional quality (Lazarides and Ittel 2012; Meece et al. 2006a, b; Wigfield et al. 2002).

3.5.3 Statistical Analysis

We conducted a number of modeling steps comprising measurement invariance testing and multilevel structural equation modeling. In all analyses, robust maximum likelihood estimation (MLR) was used, with standard errors and tests of fit that were robust against non-normality of observations and the use of categorical variables in the presence of at least four response categories (Beauducel and Herzberg 2006; Rhemtulla et al. 2012).

Step 1 Measurement invariance testing

We applied multi-group confirmatory factor analysis (MGCFA) to test the measurement models of each construct included in the proposed research model for invariance across the 50 participating TIMSS 2011 countries. This step was necessary to ensure that the measures were, to a sufficient degree, comparable and to exclude measurement bias as a potential source of cross-country differences (Rutkowski and Svetina 2014). As instructional quality and the motivational constructs were measured at both the student and the classroom level, we tested for measurement invariance at these two levels by conducting (a) single-level MGCFA, and (b) multilevel MGCFA. For the latter, the student (individual) level was saturated, assuming only correlations among all items of a scale (Ryu 2014). For the school climate constructs, however, only (b) applied, because they were measured by teacher ratings.

Testing for measurement invariance, we specified a configural model, in which the number of factors and the pattern specified in the loading matrices were equal

across countries. Building upon this model, metric invariance additionally constrained the factor loadings to equality. Finally, scalar invariance assumed that the item intercepts were equal across countries in addition to the factor loadings. To ensure that the relations among the latent variables proposed in our research model were comparable across countries, at least metric invariance must hold (Millsap 2011). We evaluated these three invariance models with respect to their overall goodness-of-fit, and the changes in the goodness-of-fit statistics after introducing constraints on factor loadings and item intercepts. The configural model formed the basis for evaluating these changes.

To evaluate the changes in model fit, we followed the recommendations given by Rutkowski and Svetina (2014) and considered the changes of the incremental fit indices as practically insignificant if changes in the comparative fit index (CFI) were less than 0.020, and the root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR) changed by less than 0.020,⁴ compared to the configural invariance model. These statistics are particularly sensitive to deviations from the invariance of factor loadings and intercepts (Chen 2007). Although these guidelines have been studied and applied in various contexts, we consider them to be only rough guidelines in the current investigation, as the number of groups is extraordinarily high.

Step 2 Multilevel structural equation modeling (MSEM)

On the basis of the measurement invariance testing results, we applied multilevel structural equation modeling to the data of each of the 50 countries and used the factor loadings obtained from the first step of invariance testing at the classroom level as fixed parameters in the measurement models of the constructs under investigation. This procedure may circumvent convergence problems and provides results in reasonable estimation time. Although we are aware that this fixed parameters country-by-country approach may result in less precise parameter estimates than a multi-group multilevel modeling approach that estimates the factor loadings across countries, there were significant advantages in reducing the number of model parameters and therefore simplifying the model estimation. In fact, according to our research model, in a multi-group MSEM describing the relations among, for instance, an orderly school climate, instructional quality, and academic self-concept in mathematics, more than 1500 parameters had to be estimated; the fixed parameters approach resulted in 36 estimated parameters per country. These figures illustrate the substantial reduction in estimation effort and model complexity gained by this approach, with only limited loss in precision.

In all country-by-country MSEM analyses, we tested the indirect effect of school climate on achievement motivation via instructional quality against zero to check

⁴Please note that for large numbers of groups (20 or more) more liberal criteria for the Δ RMSEA and Δ SRMR may be applied. In this sense, an increase in the RMSEA of less or equal than 0.030 could still be considered acceptable.

for potential (partial) mediation. The corresponding standard errors were obtained from *Mplus* using an asymptotic estimation procedure (Muthén and Muthén 1998–2014; Preacher et al. 2010). We used an implicit latent group-mean centering and level-specific standardization (Stancel-Piątak and Desa 2014).⁵

In all analyses, missing data were handled using the full-information maximum likelihood procedure under the assumption that missing data occurred at random (Enders 2010). We furthermore included the mathematics teachers' weights in the analyses (MATWGT) to account for the sampling design applied in TIMSS 2011. The IDB (International Database) analyzer (IEA 2012) was used to prepare and merge the data. Significance testing was performed at the 5 % level.

3.6 Results

3.6.1 Measurement Invariance Testing

As already mentioned, we tested for measurement invariance for the constructs that play a role in our proposed research model (Fig. 3.1) in order to obtain evidence on sufficient degrees of comparability of measures across the 50 participating TIMSS 2011 countries. In this respect, we tested for invariance of the measurement models at the between (classroom/teacher) level for all constructs, and for invariance at the within (student) level for constructs based on student ratings.

The resulting goodness-of-fit statistics and the corresponding model comparisons indicated that, for the different aspects of achievement motivation and student ratings of instructional quality, both student- and classroom-level metric invariance could be established (see Appendix C). For the school climate scales capturing safety and order in schools, metric invariance can be assumed. Changes in goodness-of-fit indices exceeded the suggested cut-offs in only few instances, however: (a) the suggested cut-offs have not yet been evaluated in multi-group multilevel situations with a larger number of groups (in our case, countries) and can therefore only be regarded as approximate marking points; (b) the metric invariance model fitted the data reasonably well; and (c) while the CFI was substantially lower in the metric model, the TLI improved compared to the configural model, suggesting that there was mixed evidence on changes in these fit statistics. As a consequence, we accepted the metric invariance models for all constructs and levels and interpreted the invariance testing results as evidence for a sufficient degree of comparability. We therefore proceeded with comparing the relations among school climate, instructional quality, and achievement motivation across countries.

⁵The resulting standardized regression coefficients are those reported for the classroom/teacher level (in contrast to reporting the contextual or compositional effects; Marsh et al. 2012).

3.6.2 Multilevel Structural Equation Modeling

Since specifying our research model for three aspects of school climate (namely, SEAS, safety, and order in schools) and three aspects of achievement motivation (self-concept, intrinsic, and extrinsic value) has resulted in nine models and therefore a rich amount of data, we systematized the findings using the classification presented previously (see Fig. 3.2). Specifically, for each, we allocated the data of a particular country to one of our models: MED, INQ, PSC, and PSC&INQ. We here present detailed results for one of these models (results of the other models are provided in Appendix C). The goodness-of-fit statistics for the country-by-country MSEM analysis with fixed factor loadings in the measurement models of the constructs (see Sect. 3.5) were largely acceptable; in some cases, the statistics approached the cut-off value (CFI and TLI close to 0.90, RMSEA close to 0.08).

We studied the model for SEAS as an indicator of teachers' perceived school climate and students' intrinsic value. This model provided regression coefficients for the 50 countries (Table 3.1). On the basis of the direct and indirect effects, each country was assigned to one of the potential models.

Twelve countries fitted the MED model, where it was apparent that the SEAS-intrinsic value relation was at least partially mediated by instructional quality (indirect effect $\beta_1 \times \beta_2$: $M = 0.206$, $SD = 0.048$, $Mdn = 0.186$, $Min = 0.151$, $Max = 0.283$); for the South African data set, the mediation was negative due to a negative relation between SEAS and instructional quality. Twenty-six countries satisfied model INQ; the average path coefficient of instructional quality on students' intrinsic value was 0.844 ($SD = 0.087$, $Mdn = 0.861$, $Min = 0.590$, $Max = 0.985$). The remaining 11 countries fulfilled the model PSC&INQ, where both SEAS ($M[\beta_3] = 0.177$, $SD = 0.068$, $Mdn = 0.162$, $Min = 0.101$, $Max = 0.337$) and instructional quality ($M[\beta_2] = 0.865$, $SD = 0.040$, $Mdn = 0.865$, $Min = 0.787$, $Max = 0.933$) had significant effects on intrinsic value; in Turkey there was a negative relationship between SEAS and instructional quality. None of the countries fitted model PSC.

We were thus able to identify three out of the four proposed scenarios in our research. Interestingly, for the majority of countries, instructional quality was strongly associated with students' intrinsic value; for some countries, the relation between SEAS and intrinsic value was at least partially mediated via instructional quality. The latter result points to the existence of a potential mechanism among the three constructs.

We studied the proposed research model for each of the school climate aspects and aspects of achievement motivation and assigned them the appropriate model (Table 3.2).

Given the rich results, we here only highlight selected patterns. First, the results for each country are relatively consistent; the majority of countries display similar patterns across all aspects of achievement motivation, given a specific aspect of school climate. For example, the Australian data set indicated that the school climate-instructional quality-achievement motivation relation was mediated for



Table 3.1 Standardized direct and indirect effects in the model with SEAS as the school climate aspect and intrinsic value as the motivational outcome variable (see also Fig. 3.2 for explanation of scenarios)

Country	Direct effects			Indirect effect	Model
	β_1 (SE)	β_2 (SE)	β_3 (SE)	$\beta_1 \times \beta_2$ (SE)	
Armenia	-0.131 (0.098)	0.750 (0.059)*	0.118 (0.090)	-0.098 (0.073)	INQ
Australia	0.221 (0.066)*	0.827 (0.034)*	0.014 (0.054)	0.183 (0.055)*	MED
Bahrain	-0.226 (0.128)	0.855 (0.093)*	0.337 (0.092)*	-0.193 (0.120)	PSC&INQ
Chile	-0.003 (0.100)	0.897 (0.0039)**	0.070 (0.080)	-0.003 (0.090)	INQ
Chinese Taipei	-0.039 (0.091)	0.933 (0.037)*	0.188 (0.077)*	-0.037 (0.085)	PSC&INQ
England	0.173 (0.091)	0.836 (0.037)*	0.044 (0.058)	0.144 (0.076)	INQ
Finland	0.186 (0.090)*	0.902 (0.034)*	0.073 (0.063)	0.168 (0.081)*	MED
Georgia	0.096 (0.097)	0.859 (0.043)*	0.160 (0.069)*	0.083 (0.083)	PSC&INQ
Ghana	0.048 (0.112)	0.817 (0.090)*	0.061 (0.091)	0.039 (0.092)	INQ
Hong Kong SAR	0.252 (0.125)*	0.746 (0.064)*	0.277 (0.078)*	0.188 (0.089)*	MED
Hungary	0.076 (0.090)	0.916 (0.023)*	0.011 (0.053)	0.070 (0.082)	INQ
Indonesia	-0.051 (0.110)	0.943 (0.034)*	0.027 (0.061)	-0.049 (0.104)	INQ
Iran, Islamic Rep. of	-0.101 (0.087)	0.845 (0.040)*	0.238 (0.079)*	-0.086 (0.075)	PSC&INQ
Israel	-0.062 (0.078)	0.827 (0.042)*	0.147 (0.067)*	-0.051 (0.065)	PSC&INQ
Italy	0.048 (0.102)	0.985 (0.028)*	0.059 (0.057)	0.047 (0.100)	INQ
Japan	0.288 (0.090)*	0.895 (0.044)*	0.064 (0.073)	0.257 (0.078)*	MED
Jordan	0.071 (0.089)	0.963 (0.042)*	0.016 (0.064)	0.068 (0.086)	INQ
Kazakhstan	0.049 (0.103)	0.892 (0.034)*	0.130 (0.058)*	0.043 (0.092)	PSC&INQ
Korea, Rep. of	0.137 (0.072)	0.715 (0.048)*	-0.027 (0.060)	0.098 (0.053)	INQ
Lebanon	-0.095 (0.109)	0.787 (0.051)*	0.180 (0.083)*	-0.075 (0.087)	PSC&INQ

(continued)

Table 3.1 (continued)

Country	Direct effects			Indirect effect	Model
	β_1 (SE)	β_2 (SE)	β_3 (SE)	$\beta_1 \times \beta_2$ (SE)	
Lithuania	0.069 (0.082)	0.933 (0.022)*	0.044 (0.050)	0.065 (0.076)	INQ
Macedonia	0.081 (0.100)	0.878 (0.030)*	-0.035 (0.057)	0.071 (0.088)	INQ
Malaysia	0.105 (0.081)	0.883 (0.032)*	0.127 (0.048)*	0.092 (0.070)	PSC&INQ
Morocco	0.073 (0.095)	0.906 (0.063)*	0.091 (0.061)	0.066 (0.084)	INQ
New Zealand	0.206 (0.064)*	0.858 (0.030)*	-0.064 (0.062)	0.176 (0.056)*	MED
Norway	0.323 (0.105)*	0.866 (0.051)*	0.188 (0.067)*	0.280 (0.084)*	MED
Oman	0.215 (0.085)*	0.775 (0.034)*	0.103 (0.057)	0.166 (0.064)*	MED
Palestinian Nat'l Auth.	0.211 (0.084)*	0.780 (0.047)*	0.015 (0.067)	0.165 (0.067)*	MED
Qatar	0.190 (0.099)	0.860 (0.055)*	0.002 (0.082)	0.164 (0.089)	INQ
Romania	0.061 (0.079)	0.886 (0.039)*	-0.021 (0.054)	0.054 (0.070)	INQ
Russian Federation	0.214 (0.095)*	0.917 (0.028)*	0.007 (0.053)	0.197 (0.088)*	MED
Saudi Arabia	0.052 (0.114)	0.892 (0.033)*	-0.024 (0.062)	0.046 (0.102)	INQ
Singapore	0.066 (0.070)	0.803 (0.059)*	0.056 (0.073)	0.053 (0.057)	INQ
Slovenia	0.035 (0.069)	0.893 (0.031)*	0.163 (0.058)*	0.031 (0.061)	PSC&INQ
Sweden	0.206 (0.074)*	0.734 (0.050)*	0.135 (0.076)	0.151 (0.054)*	MED
Syrian Arab Rep.	-0.008 (0.124)	0.771 (0.065)*	0.006 (0.095)	-0.006 (0.096)	INQ
Thailand	-0.057 (0.112)	0.896 (0.045)*	0.087 (0.065)	-0.0051 (0.101)	INQ
Tunisia	0.002 (0.096)	0.881 (0.036)*	0.072 (0.071)	0.002 (0.084)	INQ
Turkey	-0.063 (0.091)	0.869 (0.031)*	-0.106 (0.053)*	-0.055 (0.078)	PSC(-) &INQ
Ukraine	0.266 (0.118)*	0.952 (0.027)*	0.037 (0.059)	0.253 (0.111)*	MED
United Arab Emirates	-0.098 (0.055)	0.834 (0.024)*	0.060 (0.039)	-0.082 (0.046)	INQ

(continued)

Table 3.1 (continued)

Country	Direct effects			Indirect effect	Model
	β_1 (SE)	β_2 (SE)	β_3 (SE)	$\beta_1 \times \beta_2$ (SE)	
United States of America	0.076 (0.063)	0.758 (0.029)*	-0.030 (0.050)	0.057 (0.048)	INQ
<i>Ninth grade participants</i>					
Botswana	-0.191 (0.127)	0.763 (0.067)*	0.156 (0.113)	-0.146 (0.098)	INQ
Honduras	0.295 (0.112)*	0.958 (0.033)*	-0.005 (0.065)	0.283 (0.110)*	MED
South Africa	-0.398 (0.067)*	0.818 (0.039)*	0.012 (0.061)	-0.325 (0.058)*	MED(-)
<i>Benchmarking participants</i>					
Abu Dhabi, UAE	-0.086 (0.110)	0.861 (0.033)*	0.095 (0.064)	-0.074 (0.095)	INQ
Alberta, Canada	0.160 (0.088)	0.801 (0.038)*	-0.033 (0.072)	0.128 (0.073)	INQ
Dubai, UAE	-0.136 (0.078)	0.814 (0.043)*	0.006 (0.064)	-0.111 (0.064)	INQ
Ontario, Canada	0.111 (0.117)	0.590 (0.070)*	0.158 (0.097)	0.065 (0.067)	INQ
Quebec, Canada	0.097 (0.081)	0.871 (0.029)*	0.101 (0.051)	0.084 (0.070)	PSC&INQ

Note (-) indicates a negative regression coefficient. SE standard error. UAE United Arab Emirates. * $p < 0.05$

both SEAS and order for all motivational aspects and model INQ was consistently found for the safety component. Nevertheless, there are countries where the models differ (consider Singapore and Kazakhstan).

Second, for SEAS and safety in schools, model INQ dominated; the MED model was less common, but could be identified for more than 20 % of the countries. In addition, model PSC&INQ was supported in 13 countries for students’ self-concept and in 10 countries for intrinsic value. For order in schools, the MED model was the most common, but the INQ model was also apparent for a significant number of countries (Table 3.3).

Third, differences in the assignment of a specific model across the different aspects of achievement motivation could be identified. For instance, while the PSC&INQ model can be found in more than 20 % of the countries for SEAS and for the two motivational constructs of self-concept and intrinsic value, this model is substantially less common for students’ extrinsic value. For the remaining combinations of scenarios, the overall frequencies across the motivational aspects are relatively consistent.

Fourth, we found support for model PSC in only two cases, which implies that only teachers’ perceptions of the school climate were significantly related to motivation (Table 3.3).

Table 3.2 Overview of existing models describing the relations among school climate, instructional quality, and achievement motivation (see Fig. 3.2 for explanation of scenarios)

Countries	SEAS			Safety			Order		
	Self-concept	Intrinsic value	Extrinsic value	Self-concept	Intrinsic value	Extrinsic value	Self-concept	Intrinsic value	Extrinsic value
Armenia	N/A	INQ	MED(-)	N/A	MED(-)	MED(-)	N/A	MED	INQ
Australia	MED	MED	MED	INQ	INQ	INQ	MED	MED	MED
Bahrain	INQ	PSC&INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ
Chile	PSC&INQ	INQ	INQ	PSC&INQ	INQ	INQ	INQ	INQ	INQ
Chinese Taipei	PSC&INQ	PSC&INQ	INQ	PSC&INQ	PSC&INQ	INQ	PSC&INQ	INQ	INQ
England	INQ	INQ	PSC&INQ	INQ	INQ	INQ	MED	MED	MED
Finland	MED	MED	MED	INQ	INQ	INQ	MED	MED	MED
Georgia	PSC&INQ	PSC&INQ	INQ	INQ	INQ	INQ	MED	MED	MED
Ghana	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ
Hong Kong SAR	MED	MED	MED	INQ	INQ	PSC&INQ	MED	MED	MED
Hungary	PSC&INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ
Indonesia	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ
Iran, Islamic Rep. Of	PSC&INQ	PSC&INQ	INQ	MED(-)	PSC&INQ	MED(-)	INQ	PSC&INQ	INQ
Israel	PSC&INQ	PSC&INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ
Italy	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ
Japan	MED	MED	MED	INQ	INQ	INQ	MED	MED	MED
Jordan	INQ	INQ	INQ	INQ	MED(-)	MED(-)	INQ	INQ	INQ
Kazakhstan	PSC&INQ	PSC&INQ	INQ	INQ	PSC(-) &INQ	PSC(-) &INQ	MED	MED	INQ

(continued)



Table 3.2 (continued)

Countries	SEAS				Safety				Order		
	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	MED	MED
Korea, Rep. Of	INQ				INQ				INQ	MED	MED
Lebanon	INQ	PSC&INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	PSC&INQ	PSC&INQ
Lithuania	INQ	INQ	INQ	INQ	MED	MED	MED	MED	INQ	INQ	INQ
Macedonia	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	MED	MED
Malaysia	INQ	PSC&INQ	PSC&INQ	PSC&INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ
Morocco	PSC&INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ
New Zealand	MED	MED	MED	MED	INQ	INQ	INQ	INQ	INQ	MED	MED
Norway	MED	MED	MED	MED	INQ	INQ	INQ	INQ	MED	MED	MED
Oman	MED	MED	MED	MED	INQ	INQ	INQ	INQ	INQ	MED	MED
Palestinian Nat'l Auth.	MED	MED	MED	MED	INQ	INQ	INQ	INQ	INQ	MED	MED
Qatar	INQ	INQ	MED	MED	INQ	INQ	INQ	INQ	INQ	MED	MED
Romania	PSC&INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	MED	MED
Russian Federation	MED	MED	MED	MED	MED	MED	MED	MED	MED	MED	MED
Saudi Arabia	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	PSC&INQ	PSC&INQ	INQ
Singapore	PSC&INQ	INQ	PSC(-)&INQ	PSC(-)&INQ	PSC&INQ	INQ	INQ	INQ	INQ	PSC&INQ	INQ
Slovenia	PSC&INQ	PSC&INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	MED	MED
Sweden	MED	MED	MED	MED	INQ	INQ	INQ	INQ	INQ	MED	MED
Syrian Arab Republic	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ

(continued)

Table 3.2 (continued)

Countries	SEAS			Safety			Order				
	INQ	INQ	INQ	INQ	PSC&INQ	INQ	PSC&INQ	INQ	MED	MED	MED
Thailand	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	MED	MED	MED
Tunisia	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	MED	MED	MED
Turkey	INQ	PSC(-) &INQ	INQ	INQ	INQ	INQ	PSC(-) &INQ	INQ	MED	MED	MED
Ukraine	MED	MED	MED	INQ	INQ	PSC(-) &INQ	INQ	INQ	INQ	INQ	INQ
United Arab Emirates	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ
United States of America	PSC&INQ	INQ	INQ	INQ	INQ	PSC(-) &INQ	INQ	INQ	MED	MED	MED
<i>Ninth grade participants</i>											
Botswana	INQ	INQ	PSC	INQ	INQ	INQ	N/A	INQ	INQ	INQ	N/A
Honduras	MED	MED	MED	MED	MED	MED	MED	MED	MED	MED	MED
South Africa	MED(-)	MED(-)	MED(-)	MED(-)	MED(-)	MED(-)	MED(-)	MED(-)	PSC&INQ	PSC&INQ	INQ
<i>Benchmarking participants</i>											
Abu Dhabi, UAE	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	PSC&INQ	PSC&INQ
Alberta, Canada	INQ	INQ	INQ	INQ	INQ	PSC(-) &INQ	INQ	INQ	INQ	INQ	INQ
Dubai, UAE	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ	INQ
Ontario, Canada	PSC	INQ	PSC&INQ	INQ	INQ	INQ	INQ	INQ	N/A	MED	MED
Quebec, Canada	PSC&INQ	PSC&INQ	PSC&INQ	INQ	INQ	INQ	INQ	INQ	MED	MED	MED

Note (-) indicates a negative regression coefficient. SE standard error. UAE United Arab Emirates. * $p < 0.05$



Table 3.3 Absolute frequencies of the scenarios (see Fig. 3.2) identified in the 50 participating TIMSS 2011 countries

Model	SEAS			Safety			Order		
	Self-concept	Intrinsic value	Extrinsic value	Self-concept	Intrinsic value	Extrinsic value	Self-concept	Intrinsic value	Extrinsic value
INQ	22	26	29	40	38	36	19	20	24
MED	12	12	13	3	3	4	24	26	24
MED(-)	1	1	2	2	3	4	0	0	0
PSC	1	0	1	0	0	0	0	0	0
PSC&INQ	13	10	4	4	2	3	5	4	1
PSC(-)&INQ	0	1	1	0	4	2	0	0	0
N/A	1	0	0	1	0	1	2	0	1

Note (-) indicates a negative regression coefficient. *UAE* United Arab Emirates. *N/A* indicates that either the relations among the three constructs were insignificant ($p > 0.05$) or there was only a significant relation between school climate and instructional quality without any connection to the motivational outcome variable

Fifth, looking at the role of instructional quality as a mediator between aspects of school climate and achievement motivation, the following cultural patterns could be identified:

- *Scandinavian countries*: Mediation was apparent for SEAS and order across all motivational aspects.
- *English-speaking countries*: Mediation was apparent for SEAS and order across all motivational aspects in Australia and New Zealand. In England, the USA, and Quebec, only the relation between order and motivation was mediated by instructional quality.
- *Asian countries*: In Japan and Hong Kong, the relations between SEAS and achievement motivation, and order and achievement motivation were mediated by instructional quality; in Korea and Thailand, mediation was apparent only for order.
- *Eastern and Central European countries*: instructional quality mediated the relation between order and student motivation in Georgia, Romania, Macedonia, Slovenia, and Kazakhstan. In the Russian Federation and the Ukraine, SEAS was mediated; in the Russian Federation and Lithuania, safety in schools was mediated.
- *Arabic countries*: The relation between safety and student motivation was mediated in Iran and Jordan; in addition, SEAS and order were mediated in Oman, Palestine, and Qatar.
- *North Africa*: The relation between order and student motivation was mediated by instructional quality in Tunisia and Turkey; mediation was also found for SEAS and safety in South Africa.
- *South America*: The models with SEAS, safety, and order as school climate aspects showed mediation in Honduras.

Overall, the findings indicate that differing scenarios of relations exist among school climate, instructional quality, and achievement motivation. Although there were different patterns of relations across the aspects of school climate and achievement motivation, models INQ and MED dominate.

3.7 Discussion

This study was concerned with the relations among school climate, instructional quality, and achievement motivation across the 50 participating TIMSS 2011 countries in grade eight in mathematics. We proposed a research model that allowed us to identify four potential scenarios that indicated different substantive interpretations. With the help of MSEM, we found that models INQ and MED dominated across all aspects of school climate and achievement motivation.

As a major aim in practice, policy, and teacher education is to increase the level of instructional quality, as well as boost students' motivation for mathematics, our

findings emphasize the importance of school climate for instructional quality and motivation (Creemers and Kyriakides 2008; Sammons 2009). Collective beliefs, capabilities, and trust among the included members of the school institution, as manifested by high levels of SEAS, are important for instructional quality and achievement motivation. Creating a school climate that is oriented toward academic success can therefore be associated with higher instructional quality, which in turn leads to positive student outcomes. Although we do not claim causality in this mechanism, we believe that a positive school climate is indeed beneficial for instruction (Mitchell et al. 2010; Morin et al. 2014). In this respect, we found support for the mediating role of instructional quality in the relation between SEAS and achievement motivation in a number of countries. Nevertheless, in some cases, higher levels of SEAS were associated with lower instructional quality; this finding may point to the potential negative consequences of emphasizing academic success in such a way that competition and an orientation toward performance rather than motivation emerge.

Another explanation for the mediation may refer to the conceptualization and measurement of SEAS as an aspect of teachers' perceived school climate. SEAS was measured as a latent variable, where indicators refer to parents', students', and teachers' ambitions and priorities for learning and academic success. The covariance of these indicators reflects the collective and shared beliefs among these members of the school institution (Hoy et al. 2006). The link between these members that may arise when everyone aims for the same goal seems to influence teachers' instructional quality and student motivation. Indeed, previous research has shown that a strong student-teacher relationship positively influences student achievement (Roorda et al. 2011).

Moreover, we note that the mediation model was particularly apparent for the order component; this again indicates that order in schools may serve as a prerequisite for creating learning environments of high quality. But since the strength of the mediation differed across countries, further investigation is needed to assess what determines this mechanism in the context of the specific countries. Our secondary data analysis showed cross-country differences in the occurrence of mediation.

In summary, as previous research has found that instructional quality and school climate are related to both achievement student motivation (Fauth et al. 2014; Hoy et al. 2006; Klieme et al. 2009; Klusmann et al. 2008; Nilsen and Gustafsson 2014), it is thus unsurprising that we found that school climate influences motivation.

Although the importance of school climate for students' achievement motivation has been confirmed in our study, the effects of instructional quality dominated across almost all scenarios. Indeed, previous research has identified significant interactions between instructional quality and learning outcomes (Baumert et al. 2010; Blömeke et al. 2013; Fauth et al. 2014), and our results support these findings. Moreover, while the well-recognized research in this field is often restricted to German-speaking countries (see Baumert et al. 2010; Klieme et al. 2009) and English-speaking countries (Brophy 2006; Good et al. 2009), our findings support previous research in general, but also contribute to fostering understanding of

educational policy and practice in other countries as well, including developing countries.

The strong relation between instructional quality and achievement motivation may have two potential sources. First, instruction that focuses on engaging students to learn mathematics also creates opportunities for students to become motivated by the subject. Second, the measurement of instructional quality focused mainly on the engagement part of the construct, thus aligning with the measurement of achievement motivation; this alignment of the measures may have created similarities in how students understand and rate the items presented in the TIMSS 2011 questionnaire. Regardless, this strong association was consistently found in almost all countries and therefore needs further attention.

While there have been a number of studies on the relations between school climate and achievement (see Hoy et al. 2006; Martin et al. 2013; McGuigan and Hoy 2006) and on relations between instructional quality and achievement (e.g. Baumert et al. 2010), there have been relatively few studies investigating the relations between school climate and instructional quality (Creemers and Kyriakides 2010). The findings support our expectation of an association between instructional quality and school climate, and point out the importance of SEAS, Safety, and Order as important aspects of school climate. Moreover, including all countries and using international large-scale studies may also inform policy and practice about the importance of SEAS for instructional quality.

It is noteworthy that SEAS and instructional quality play an important role not only for students' motivation in learning mathematics but also for their self-beliefs and future-oriented motivation to pursue a career in mathematics and value the subject; this points to the significance of both the school and the classroom environment for career aspirations and for students' evaluation of their own capabilities in mathematics, which in turn determine their performance.

3.8 Limitations

One limitation of the present study is the measurement of the core construct, instructional quality. Although the student ratings of teachers' clarity and support in learning provide valid indicators of instructional quality with respect to instructional practices that are aimed at engaging students in learning mathematics (Scherer and Gustafsson 2015a), it is desirable to capture further aspects, such as cognitive activation and classroom management (Fauth et al. 2014; Klieme et al. 2009). We believe that gaining conceptual breadth in the measurement of instructional quality provides (a) a better representation of this multidimensional construct, and (b) more information about whether or not different aspects of instructional quality relate differently to student outcomes.

3.9 Conclusion

Besides supporting the importance of classroom instruction for motivational outcomes, our study advocates the relevance of school climate for both instructional quality and achievement motivation in many countries, feeding into the search for ways to improve instructional quality. We encourage further research into the field of educational effectiveness, to study the effects of instructional quality on educational outcomes by accounting for the school climate context.

References

- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.
- 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. doi:10.3102/0002831209345157
- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186–203. doi:10.1207/s15328007sem1302_2
- Blömeke, S., Gustafsson, J. E., & Shavelson, R. (2013). Assessment of competencies in higher education. *Zeitschrift für Psychologie*, 221(3), 202.
- Bofah, E. A.-t., & Hannula, M. S. (2015). TIMSS data in an African comparative perspective: Investigating the factors influencing achievement in mathematics and their psychometric properties. *Large-scale Assessments in Education*, 3(1), doi:10.1186/s40536-015-0014-y
- Brophy, J. (2006). Observational research on generic aspects of classroom teaching. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 755–780). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504. doi:10.1080/10705510701301834
- Chen, P., & Vazsonyi, A. T. (2013). Future orientation, school contexts, and problem behaviors: A multilevel study. *Journal of Youth and Adolescence*, 42(1), 67–81. doi:10.1007/s10964-012-9785-4
- Creemers, B., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: a contribution to policy, practice and theory in contemporary schools*. Abingdon, Oxon: Routledge.

- Creemers, B., & Kyriakides, L. (2010). Explaining stability and changes in school effectiveness by looking at changes in the functioning of school factors. *School Effectiveness and School Improvement, 21*(4), 409–427.
- Davison, M. L., & Srichantra, N. (1988). Acquiescence in components analysis and multidimensional scaling of self-rating items. *Applied Psychological Measurement, 12*(4), 339–351. doi:10.1177/014662168801200402
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Dordrecht: Springer Science & Business Media.
- Deemer, S. (2004). Classroom goal orientation in high school classrooms: revealing links between teacher beliefs and classroom environments. *Educational Research, 46*(1), 73–90. doi:10.1080/0013188042000178836
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109–132.
- Enders, C. K. (2010). *Applied Missing Data Analysis*. New York, NY: Guilford Press.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9.
- Goddard, R. D. (Ed.) (2002). *Collective efficacy and school organization: A multilevel analysis of teacher influence in schools* (Vol. 1). Greenwich, CT: Information Age Publishing.
- Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., & Preckel, F. (2014). “My questionnaire is too long!” The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology, 39*(3), 188–205. doi:10.1016/j.cedpsych.2014.04.002
- Goldstein, S. E., Young, A., & Boyd, C. (2008). Relational aggression at school: Associations with school safety and social climate. *Journal of Youth and Adolescence, 37*(6), 641–654.
- Good, T. L., Wiley, C. R., & Florez, I. R. (2009). Effective teaching: An emerging synthesis. In L. J. Saha & A. G. Dworkin (Eds.), *International handbook of research on teachers and teaching* (pp. 803–816). Dordrecht: Springer.
- Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: do they matter? *Personality and Individual Differences, 35*(6), 1241–1254. doi:10.1016/S0191-8869(02)00331-8
- Gregory, A., Cornell, D., & Fan, X. (2012). Teacher safety and authoritative school climate in high schools. *American Journal of Education, 118*(4), 401–425.
- Harter, S. (1981). A new self-report scale of intrinsic versus extrinsic orientation in the classroom: motivational and informational components. *Developmental Psychology, 17*(3), 300.
- Hattie, J. (2009). *Visible learning: A synthesis of 800 + meta-analyses on achievement*. Abingdon, Oxon: Routledge.
- Hoy, W. K., & Tschannen-Moran, M. (1999). Five faces of trust: An empirical confirmation in urban elementary schools. *Journal of School Leadership, 9*, 184–208.
- Hoy, W. K., Tarter, C. J., & Hoy, A. W. (2006). Academic optimism of schools: A force for student achievement. *American Educational Research Journal, 43*(3), 425–446.
- IEA. (2012). *International database analyzer (version 3.1)*. (Software) Hamburg, Germany: International Association for the Evaluation of Educational Achievement (IEA). Retrieved from <http://www.iea.nl/data.html>.
- Klieme, E. (2013). The role of large-scale assessments in research on educational effectiveness and school development. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 115–147). Dordrecht: Springer, Netherlands.
- Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). New York, NY: Waxmann Publishing Co.
- Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Teachers’ occupational well-being and quality of instruction: The important role of self-regulatory

- patterns. *Journal of Educational Psychology*, 100(3), 702–715. doi:10.1037/0022-0663.100.3.702
- Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, 105(3), 805–820. doi:10.1037/a0032583
- Kythreotis, A., Pashiardis, P., & Kyriakides, L. (2010). The influence of school leadership styles and culture on students' achievement in Cyprus primary schools. *Journal of Educational Administration*, 48(2), 218–240.
- Lazarides, R., & Ittel, A. (2012). Instructional quality and attitudes toward mathematics: Do self-concept and interest differ across students' patterns of perceived instructional quality in mathematics classrooms? *Child Development Research*, 2012, 1–11. doi:10.1155/2012/813920
- Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, 34(2), 120–131. doi:10.1016/j.cedpsych.2008.12.001
- Marsh, H. W., & Gouvenet, P. J. (1989). Multidimensional self-concepts and perceptions of control: Construct validation of responses by children. *Journal of Educational Psychology*, 81(1), 57–69.
- Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J. S., Abdelfattah, F., Leung, K. C., & Parker, P. (2013). Factorial, convergent, and discriminant validity of TIMSS math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology*, 105(1), 108–128. doi:10.1037/a0029907
- Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, 47(2), 106–124. doi:10.1080/00461520.2012.670488
- Martin, M. O., Foy, P., Mullis, I. V. S., & O'Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—Implications for early learning* (pp. 109–178). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- McGuigan, L., & Hoy, W. K. (2006). Principal leadership: Creating a culture of academic optimism to improve achievement for all students. *Leadership and Policy in Schools*, 5(3), 203–229. doi:10.1080/15700760600805816
- Meece, J. L., Anderman, E. M., & Anderman, L. H. (2006a). Classroom goal structure, student motivation, and academic achievement. *Annual Review of Psychology*, 57, 487–503. doi:10.1146/annurev.psych.56.091103.070258
- Meece, J. L., Glienke, B. B., & Burg, S. (2006b). Gender and motivation. *Journal of School Psychology*, 44(5), 351–373. doi:10.1016/j.jsp.2006.04.004
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.
- Mitchell, M. M., & Bradshaw, C. P. (2013). Examining classroom influences on student perceptions of school climate: The role of classroom management and exclusionary discipline strategies. *Journal of School Psychology*, 51(5), 599–610. doi:10.1016/j.jsp.2013.05.005
- Mitchell, M. M., Bradshaw, C. P., & Leaf, P. J. (2010). Student and teacher perceptions of school climate: A multilevel exploration of patterns of discrepancy. *Journal of School Health*, 80(6), 271–279. doi:10.1111/j.1746-1561.2010.00501.x
- Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2015). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–24. doi:10.1080/10705511.2014.961800

- Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education*, 82(2), 143–167. doi:10.1080/00220973.2013.769412
- Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Muthén, B., & Muthén, L. (1998–2014). *Mplus Version 7.3*. Los Angeles, CA: Muthén & Muthén.
- Nilsen, T., & Gustafsson, J.-E. (2014). School emphasis on academic success: exploring changes in science performance in Norway between 2007 and 2011 employing two-level SEM. *Educational Research and Evaluation*, 20(4), 308–327. doi:10.1080/13803611.2014.941371
- NSF. (2012). *Science and engineering indicators 2012*. Retrieved 31 December 2015 from <http://www.nsf.gov/statistics/seind12/>.
- OECD. (2007). *PISA 2006. Science competencies for tomorrow's world* (Vol. 1). Paris: OECD Publishing.
- OECD. (2014a). *Education at a glance 2014: OECD indicators*. Paris: OECD Publishing. doi:10.1787/eag-2014-en
- OECD. (2014b). *TALIS 2013 results*. Paris: OECD Publishing.
- Pintrich, P., & Schunk, D. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Columbus, OH: Merrill Prentice Hall.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903. doi:10.1037/0021-9010.88.5.879
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15(3), 209–233. doi:10.1037/a0020141
- Preckel, F. (2014). Assessing need for cognition in early adolescence. *European Journal of Psychological Assessment*, 30(1), 65–72. doi:10.1027/1015-5759/a000170
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. doi:10.1037/a0029315
- Roorda, D. L., Koomen, H. M., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: a meta-analytic approach. *Review of Educational Research*, 81(4), 493–529.
- Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, 74(1), 31–57. doi:10.1177/0013164413498257
- Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology*, 5. doi:10.3389/fpsyg.2014.00081
- Sammons, P. (2009). The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools. *School Improvement and School Effectiveness*, 20(1), 123–129.
- Scherer, R., & Gustafsson, J.-E. (2015a). 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(1550). doi:10.3389/fpsyg.2015.01550
- Scherer, R., & Gustafsson, J.-E. (2015b). The relations among openness, perseverance, and performance in creative problem solving: A substantive-methodological approach. *Thinking Skills and Creativity*, 18, 4–17. doi:10.1016/j.tsc.2015.04.004

- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. doi:10.3102/0034654307310317
- Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology, 42*(1), 70–83. doi:10.1037/0012-1649.42.1.70
- Stancel-Piątak, A., & Desa, D. (2014). Methodological implementation of multi group multilevel SEM with PIRLS 2011: Improving reading achievement. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 75–91). Muenster, New York: Waxmann.
- Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research, 83*(3), 357–385.
- van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: let's learn from cows in the rain. *PLoS ONE, 8*(7), 1–7. doi:10.1371/journal.pone.0068967
- Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2015). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, .* doi:10.1037/edu0000075
- Wang, M.-T., & Degol, J. L. (2015). School climate: a review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review, 1*–38. doi:10.1007/s10648-015-9319-1
- Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. doi:10.1006/ceps.1999.1015
- Wigfield, A., Battle, A., Keller, L. B., & Eccles, J. (2002). Sex differences in motivation, self-concept, career aspirations and career choice: Implications for cognitive development. In A. V. McGillicuddy-De Lisi & R. De-Lisi (Eds.), *Biology, sociology, and behavior: The development of sex differences in cognition* (pp. 93–124). Greenwich, CT: Ablex.
- Wilson, D. (2004). The interface of school climate and school connectedness and relationships with aggression and victimization. *Journal of School Health, 74*(7), 293–299.



The Impact of School Climate and Teacher Quality on Mathematics Achievement: A Difference-in-Differences Approach

Jan Eric Gustafsson and Trude Nilsen

Abstract The aim of the study was to investigate causal effects of aspects of teacher quality and school climate on mathematics achievement through use of country-level longitudinal data. By investigating within-country change over time, biasing influence from omitted variables in the form of fixed country characteristics is avoided, thereby increasing the likelihood of making correct causal inferences. Data from 38 countries participating in both TIMSS 2007 and TIMSS 2011 were analyzed with structural equation modeling techniques, using both latent and manifest variables. The analyses focused aspects of teacher quality (educational level, teaching experience and major academic discipline studied, professional development, and self-efficacy) and an aspect of school climate referred to as school emphasis on academic success (SEAS). Results showed that the teachers' attained level of education had effects on mathematics achievement. Quite substantial effects of professional development on student achievement were also identified. Teacher self-efficacy, as assessed by self-reports of preparedness for teaching in different domains, showed a weakly positive, but insignificant relation to student achievement. The teacher characteristics years of teaching experience and major academic discipline studied had no effect on student achievement. SEAS did not satisfy ideals of unidimensionality, and only items reflecting parental support for student achievement and students' desire to perform well were significantly related to student achievement. OECD and non-OECD countries showed similar results and could not be differentiated.

J.E. Gustafsson (✉)

Department of Education and Special Education, University of Gothenburg, Gothenburg, Sweden

e-mail: jan-eric.gustafsson@ped.gu.se

J.E. Gustafsson

Faculty of Educational Sciences, Centre for Educational Measurement at the University of Oslo (CEMO), Oslo, Norway

T. Nilsen

Department of Teacher Education and School Research, University of Oslo, Oslo, Norway

e-mail: trude.nilsen@ils.uio.no

Keywords School climate · Teacher quality · Student achievement · Structural equation modeling · Difference in differences · Longitudinal approach

4.1 Introduction

The trend design of international large-scale assessment (ILSA) employed in studies such as TIMSS and PISA is rarely exploited in research. However, the equated achievement scales and the fact that a large number of countries participate in adjacent cycles provide opportunities to relate change in outcomes to change in explanatory factors. Such analyses can provide a stronger basis for making causal inferences than many other analytical approaches (Gustafsson 2013).

As elaborated in Chap. 1, studies within the field of educational effectiveness research have provided valuable information about explanatory factors that are likely to influence educational outcomes. A number of studies within this field have demonstrated that students' educational outcomes are influenced by school climate (Creemers and Kyriakides 2010; Hoy et al. 2006; Thapa et al. 2013; Wang and Degol 2015). An important aspect of school climate is school emphasis on academic success (SEAS) (Hoy et al. 2006; Martin et al. 2013; Nilsen and Gustafsson 2014). Conceptually SEAS reflects a clear priority of and ambition for academic success (Martin et al. 2013; Nilsen and Gustafsson 2014). Previous research has shown that SEAS is one of the strongest predictors of achievement at the school-level across a large number of countries (Martin et al. 2013). In addition to school climate, several aspects of teacher quality have been found to influence students' educational outcomes (Goe 2007).

However, many of these studies on effects of SEAS and teacher quality are cross-sectional, with varying degree of control of factors that may bias causal inference. Thus, in many cases, the studies have only established relations between SEAS and educational achievement, and between teacher quality and educational achievement (Nordenbo et al. 2010); there is a great need to place more emphasis on credible causal inference. We aim to address this research gap by investigating whether the relations found between teacher quality and educational achievement and the relations between SEAS and educational achievement, are causal. This empirical study focuses on mathematics achievement across all countries participating in TIMSS 2007 and TIMSS 2011 by using a difference-in-differences analytical approach.

4.2 Theoretical Framework

Observational cross-sectional data allow statements about correlations. However, there are several reasons why an association between two variables (X and Y) may not be given the interpretation that X causes Y. One reason may be that Y, at least

to some extent, causes X, resulting in reverse causality. For example, if poorly achieving students are allocated more resources to compensate for their poor achievement, a negative association between resources and achievement will typically be observed, even when there is a positive causal effect of resources on achievement. Another reason may be that there are omitted variables which affect both X and Y. For example, parents with higher levels of education may successfully lobby for more resources for their child(ren)'s school. An observed relation between resources and achievement may therefore be observed, simply because the "third variable", parental education, is related both to resources and to achievement. Errors of measurement in the X and Y variables form another threat to interpretation in terms of causal relations. Such errors tend to systematically cause the relation between X and Y to be underestimated, so this source of threat tends to prevent causal relations from being detected.

Several different approaches have been developed to guard against threats to valid causal inference in analyses of observational data (see for example Winship and Morgan 1999). One powerful approach is to make multiple observations of a set of units and investigate change over time in a characteristic of interest. The units also have other characteristics, some of which are more or less constant, and which in cross-sectional analyses may correlate with the characteristic of interest. However, if the units are allowed to be their own controls, information about these fixed characteristics can be omitted without causing any bias. This can, for example, be done with regression analysis, with change scores for independent and dependent variables, or with 'fixed unit effects', in which each observed unit is identified by a dummy variable (Gustafsson 2013; Winship and Morgan 1999).

Gustafsson (2007) observed that the repeated cross-sectional design used in international studies of educational achievement (such as PIRLS, TIMSS and PISA) to measure trends in the development of achievement have a longitudinal design at the country level, even though they are not longitudinal at the student or school level. Thus, with data aggregated to the country level, it is possible to take advantage of the strength of longitudinal designs.

Aggregated data also offer other advantages to combat threats to causal inference. Thus, mechanisms that at individual level cause reverse causality need not be present at other levels of observation. For example, compensatory resource allocation to low-achieving students causes bias in analyses of student-level data, but not in country-level data.

Aggregated data also have the advantage of not being as severely influenced by errors of measurement as individual data. Thus, while student responses to single questionnaire items typically have very low reliability, estimates of class means are more reliable, and estimates of country means are very reliable indeed. The downward biasing effect of errors of measurement is therefore reduced with aggregated data.



4.2.1 School Emphasis on Academic Success

School climate is a broad concept that is understood differently across studies and fields (Wang and Degol 2015). However, some key aspects have been found to be important to student learning. One such key aspect is academic climate. Hoy and colleagues (see for example Hoy et al. 2006) have published a number of studies on this dimension of school climate. Based on reviews of previous research, they merged three dimensions of academic climate, namely collective efficacy, faculty trust in parents and students, and academic emphasis, into one latent variable they called academic optimism. In their investigation of US high schools, academic optimism was found to be positively related to student achievement. Other studies using similar measures of academic climate have found positive relations with student learning outcomes.

In educational research, there are serious challenges related to shared understanding of concepts and equal operationalization of these concepts (Muijs 2012). Some of the constructs used in international large scale surveys are built on theory and remain unaltered from one survey to the next. As described in Chap. 1, one well-established school climate construct is school emphasis on academic success (SEAS), which has remained unaltered for more than a decade. This construct has been shown to have high reliability (Martin et al. 2013; Nilsen and Gustafsson 2014) and strong predictive power across almost all countries participating in TIMSS 2011 (Martin et al. 2013). Conceptually, SEAS reflects the collective beliefs, capability and trust among the members of the school institution (namely, students, parents, teachers, and school leaders) (Hoy et al. 2006; Martin et al. 2013; Nilsen and Gustafsson 2014). Schools with high levels of SEAS promote a clear priority of and ambition for academic success (Hoy et al. 2006; Martin et al. 2013). SEAS comprises teachers' beliefs in their own capabilities, schools trust in parents and students, and teachers' expectations for students' success.

4.2.2 Teacher Quality

As described in Chap. 1, a number of aspects of teacher quality have been found to be positively related to instruction and student outcomes (Goe 2007). In the current chapter, we focus on experience, certification, and professional development as aspects of teacher qualifications, and self-efficacy as a teacher characteristic. For an overarching framework of teacher quality, we refer readers to Chap. 1, and for more detailed reviews of theories and previous research on the concepts investigated in the current chapter, we refer the reader to Chap. 2.

4.2.3 *Research Questions*

The studies we reviewed indicate the importance of teacher quality and SEAS for students' learning gain. However, most of the studies have investigated associations among variables in cross-sectional data, and many are single-country studies. There may thus be limitations both in the credibility of causal interpretations of the relations and in the generalizability of findings. We address these issues by applying methods of analysis designed to provide stricter tests of causal relations. We investigate relations between within-country change in SEAS and teacher quality and change in mathematics achievement for the 38 educational systems participating in the TIMSS grade eight assessments in 2007 and 2011.

The analytical technique applied here assumes that the effect estimates are the same across all countries. However, previous research indicates that such an assumption may not be reasonable, and it has, for example, been found that resource factors have differing impact in developed and developing countries (see Falck et al. 2015). One way to investigate such interaction effects is to conduct the analysis in different groups of countries and to compare estimates across groups. We here approximate the distinction between different levels of development with a classification into OECD and non-OECD countries.

The research questions are:

1. *To what extent can effects of SEAS and teacher quality on mathematics achievement be identified in country-level longitudinal analyses?*
2. *Are the effects the same for OECD and non-OECD countries?*

4.3 **Method**

4.3.1 *Sample*

We included all countries ($n = 38$) who participated in TIMSS 2007 ($n = 170,803$ students in grade eight) and 2011 ($n = 217,427$ students in grade eight).

4.3.2 *Constructs*

School Emphasis on Academic Success (SEAS)

Teachers' ratings formed the basis for measuring SEAS (Mullis et al. 2012). In the teacher questionnaire, teachers were asked to characterize the following five aspects within their school: teachers' understanding of and success in implementing the school's curriculum, teachers' expectations for student achievement, parental support for student achievement, and students' desire to do well in school. TIMSS used

a five-point Likert scale for these questions, ranging from very low to very high. Both the scale and the questions were identical in TIMSS 2007 and TIMSS 2011.

Mathematics Education

Teachers were asked what their major or main area of study was by selecting one or more areas from a list, including for instance mathematics, physics and biology. We included the variable reflecting whether teachers' main area of study was mathematics or not (Major).

Educational Level

The teachers were asked to rate their highest level of formal education, and the responses were coded in the ISCED system, ranging from "Did not complete ISCED level 3" to "Finished ISCED 5A, second degree or higher".

Professional Development

The teachers were asked: "In the past two years, have you participated in professional development in any of the following? (a) Mathematics content, (b) Mathematics pedagogy/instruction, (c) Mathematics curriculum, (d) Improving students' critical thinking or problem solving skills, and (e) Mathematics assessment." Responses were either yes or no.

Teacher Self-efficacy

The teachers were asked: How well prepared do you feel you are to teach the following topics? They rated a number of topics within the domains Number, Algebra, Geometry and Data and Chance on a three-point Likert scale, ranging from "Not well prepared" to "Very well prepared".

Teacher Experience

The teachers were asked: By the end of this school year, how many years will you have been teaching altogether? This question was an open ended item on a continuous scale.

4.3.3 Method of Analysis

We analyzed data at country level ($n = 38$). The IDB (International Database) analyzer (IEA 2012) was used to merge micro-data for TIMSS 2007 and TIMSS 2011, and then all variables were aggregated to country-level by computing means. Differences were also computed between corresponding variables for 2011 and 2007. The aggregation of data to the country level took individual sampling weights (MATWGT) into account and was conducted using SPSS 22.

Numerous analytical techniques have been devised to aid causal inference from longitudinal data, and they go under different labels, such as difference-in-differences analysis (Murnane and Willett 2010) or fixed effects regression analysis. The basic idea underlying all the different techniques is to remove the effect of all country characteristics that remain constant over time. Such characteristics are often omitted variables, and, unless their effect is removed, they will cause bias in the estimates of relations between determinants and outcomes. This can, for example, be done by taking differences between measures at different points in time. With measurement of determinants (X) and outcomes (Y) at two points in time, a very simple technique is to first compute the difference between the two outcome measures ($\Delta Y = Y_2 - Y_1$) and also the difference between the two measures of determinants ($\Delta X = X_2 - X_1$), and then to regress ΔY on ΔX . This regression coefficient will not be influenced by country characteristics that are constant over the two time points, and it typically is very different from what is obtained from regression analyses of data from the two cross-sections. We use such an approach, but implement it in a more general and flexible form using structural equation modeling (SEM) (Bollen and Brand 2008).

In our analytical approach, we assume measurements at two time points, X and Y ; Y_1 is regressed on X_1 and Y_2 is regressed on X_2 , and the two regression coefficients are constrained to be equal (Fig. 4.1). The model also includes a latent variable (Com) that influences Y_1 and Y_2 by the same fixed amount (1.0). The Com variable captures the effect of the fixed characteristics at the two time points, and the regressions of Y on X estimate the effect of the determinant on the outcome controlling for the fixed country characteristics. Com is assumed to be uncorrelated with X_1 and X_2 (Fig. 4.1); this model is referred to as the random effects model for longitudinal data. This assumption need not be correct, however, and if there are reasons to believe that Com is correlated with X_1 and X_2 , these correlations can be

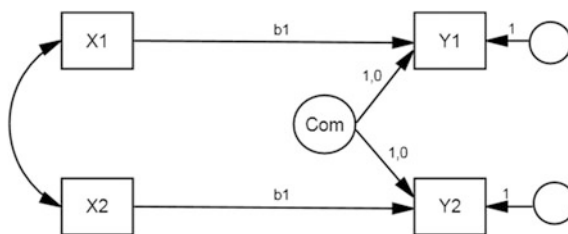


Fig. 4.1 Random effects model for two time points X and Y . Outcome Y_1 is regressed on predictor X_1 at time point 1. Similarly, outcome Y_2 is regressed on predictor X_2 at time point 2. This produces two regression coefficients that are constrained to be equal; b_1 . Com is a latent variable that captures the effect of the fixed characteristics at the two time points

added to the model. If the correlations are assumed to be equally strong, the resulting model is referred to as the fixed effects model for longitudinal data. If the correlation between Com and X1 is allowed to be different from the correlation between Com and X2, the resulting model is identical with the simple ΔX , ΔY difference model described above.

The terminology is regrettably a bit confusing. The distinction between random effects and fixed effects concern different model assumptions within fixed effects regression analysis, so the different models belong to the same difference-in-differences family.

These alternative models can easily be specified and estimated with SEM software, such as Mplus (Muthén and Muthén 1998–2014). One major advantage with this technique is that it provides information about the degree to which the model fits the data. Should it be found that the restrictive random effects model does not fit data, this suggests that one of the less restrictive models needs to be used instead. However, given that a less restrictive model is less powerful than a more restrictive model, the latter is to be preferred if it fits data.

The SEM approach also provides several other advantages. It makes it possible to also impose constraints of equality on other model parameters, such as variances, covariances and residual variances. It also allows for extensions such as use of latent variables, which can be used to investigate both the construct behind a set of items, and the individual items. SEM also allows multiple group modeling, which makes it possible to investigate whether relations between determinants and outcomes differ for different subsets of countries; we use this to investigate our second research question.

However, as the number of observations by necessity is quite limited in country-level analyses, this imposes restrictions on model complexity. It is thus not possible to estimate models with more free parameters than the number of observations, and for reasons of power, models need to be kept simple. However, a small sample size need not necessarily imply that power is low, because in SEM the amount of correlation among the variables is another important determinant of power, and in country-level longitudinal models correlations tend to be high.

Another problem associated with use of SEM techniques on aggregated country level data is that the rules of thumb developed for goodness-of-fit indices do not always apply (Bollen and Brand 2008). We therefore mainly rely on the chi-square statistic in evaluations of model fit.

4.4 Results

4.4.1 Teacher Quality

We modeled the effects of teacher quality on student achievement (Table 4.1) using the standardized estimate of parameter b1 (see Fig. 4.1). Teacher experience and teacher major showed good model fit, but had no significant effect on student mathematics achievement. For the random effects model, teachers' educational

Table 4.1 Goodness-of-fit statistics and effect estimates for all models

Model	Random effects				Fixed effects			
	Chi-square	df	Effect	t-value	Chi-square	df	Effect	t-value
<i>Teacher characteristics</i>								
Experience	1.32	3	0.05	0.44	0.16	2	-0.09	-0.50
Major	1.81	3	-0.01	-0.10	0.29	2	0.02	0.28
Educational level	8.29*	3	0.36*	2.81*	2.98	2	0.05	0.33
<i>Professional development</i>								
1 latent, 5 observed	118.16*	52	0.24	3.26*	117.14*	51	0.20	2.66*
Content	7.78	3	0.23	3.31*	6.23	2	0.19	2.72*
Instruction	10.82*	3	0.16	2.53*	4.69	2	0.10	1.80
Curriculum	3.55	3	0.16	2.89*	1.45	2	0.14	2.56*
Generic skills	7.40	3	0.04	0.61	3.46	2	0.11	1.37
Assessment	2.85	3	0.13	1.92	2.65	2	0.14	1.90
<i>Self efficacy</i>								
1 latent, 4 observed	65.23*	33	0.11	1.73	64.99	32	0.12	1.76
Algebra	2.87	3	0.12	1.70	2.53	2	0.13	1.75
Geometry	3.14	3	0.10	1.86	2.86	2	0.11	1.89
Number	8.44*	3	0.04	0.69	8.41*	2	0.04	0.71
Data and chance	2.35	3	0.08	1.62	1.28	2	0.09	1.80
<i>SEAS</i>								
1 latent, 5 observed	201.78*	58	0.31	2.83*	201.28*	57	0.37	2.46*
1 latent, 3 observed	32.99	22	-0.03	-0.42	31.15	21	-0.01	-0.11
T understanding	8.71*	3	-0.02	-0.28	8.15*	2	-0.01	-0.09
T implementation	11.16*	3	-0.04	-0.45	11.15*	2	-0.09	-0.39
T expectations	6.16	3	-0.05	-0.72	5.92	2	0.08	0.87
Parental support	1.09	3	0.48	4.82*	0.41	2	0.39	2.85*
Students' desire to learn	9.48*	3	0.13	2.34*	7.89*	2	0.17	2.51*

Note *Significant effects ($p < 0.05$). $n = 38$. Random effects and fixed effects models make different assumptions concerning correlations between the independent variable and fixed country characteristics. The chi-square test refers to the goodness-of-fit of the model and should be non-significant. Effect = standardized estimate of the coefficient for the regression of mathematics achievement on the independent variable (parameter labeled b1 in Fig. 4.1), *df* degrees of freedom, *t-value* estimate/(standard error of estimate) for effect

level yielded a significant effect, but the model fit was poor. The fixed effects model had good fit, but the effect estimate in this model was low and insignificant. While the random effects model assumes that there is no correlation between the latent variable representing the stable country characteristics and the independent variable, the fixed effects model showed that this assumption was untenable, due to a substantial positive correlation between the latent variable and teachers' educational level. It thus seems that a violation of this assumption caused the random effects model to produce a biased effect estimate. We return to this issue in the section on comparisons between OECD and non-OECD countries.

Five items were used to capture the teachers' participation in professional development. In a first step a latent variable model was specified, in which a single latent professional development variable was hypothesized to relate to all five forms of professional development. This latent variable thus reflects the countries' general tendency to involve their teachers in professional development. The model was specified with constraints of equality on corresponding factor loadings for the two waves of measurement and with covariances among residual variances of corresponding observed variables across the two measurement occasions. The fit of the model was not perfect, but with a chi-square/df ratio around two, it may be regarded as acceptable (Table 4.1). There was no difference in the fit to data of the random and fixed effects models. Significant effects on student achievement of the latent development variable were observed for both model types, with an effect expressed in terms of correlation at around 0.20.

Separate models also were fitted for each of the five items, and the results were somewhat different across items. The strongest effects were observed for content and curriculum, and they were significant for both types of models. The weakest effect was observed for generic skills, such as critical thinking or problem solving skills. For professional development in assessment, a positive effect was observed, but it was not quite significant. For instruction, the fit of the random effects model was poor, which was due to a positive correlation between the common latent variable and the independent variable. The relatively high and significant effect estimate in the random effects model should therefore not be taken seriously.

Four questions were asked about perceived preparedness for teaching in different domains, and a one-dimensional latent variable model was fitted to the four variables. The model was specified in the same way as was described for professional development above, and the fit of the model was acceptable. A weak positive effect of the latent self-efficacy variable was observed, but it was not significant (Table 4.1). Separate models also were estimated for each of the four variables. In no case was a significant effect found, but there was a weak positive effect for all variables, except for number.

4.4.2 School Emphasis on Academic Success

Five items were used to measure SEAS, and a one-dimensional model was fitted to the five variables. However, the model fit was poor, even though a relatively strong and significant effect on mathematics achievement was found. Given that there were signs of multidimensionality, an alternative model was specified that only included the three items referring to teachers (namely teachers' understanding of and their success in implementing the school's curriculum, and teachers' expectations for student achievement). This model fitted data well, but there was no relation between the latent variable and mathematics achievement (Table 4.1).

Next, separate models were estimated for each of the five variables (see Table 4.1). The three teacher items had no effect on achievement, but parental support had a strong effect on student achievement, and a smaller positive effect on students' desire to do well in school. It thus seems that the positive relation between SEAS and achievement can be accounted for by factors related to the home rather than to the school.

4.4.3 Comparisons Between OECD and Non-OECD Countries

All models used the entire set of 38 participants, thereby assuming that the same relation holds true for each and every educational system. This may be an unrealistic assumption, thus we opted to investigate to what extent this was valid across categories of educational systems. We focused on the distinction between OECD and non-OECD countries.

We estimated a two-group model with Mplus for each of the variables included in the study to investigate if the relations between the different determinants and mathematics achievement were invariant across the two categories of educational systems. The models were specified with the Mplus defaults for multiple group models, which, among other things, imply that the relations between the independent and dependent variables were constrained to be equal across groups. We therefore estimated another set of models in which this constraint was relaxed, and applied a chi-square difference test to determine the statistical significance of any difference between the regression coefficients. This procedure was repeated for both random effects and fixed effects models.

We did not identify any significant differences between pairs of regression coefficients. This is, of course, likely to be due to the low power for conducting such a test with the TIMSS 2007 and 2011 data. However, scrutiny of the estimated coefficients indicated no large differences.

Interestingly, although the random effects model for teachers' educational level did not fit the data (see Table 4.1) and produced a quite a large estimate for the effect of educational level on mathematics achievement, the fixed effects model, in

contrast, fitted well; in the latter the effect estimate was close to zero. However, the two-group models had good fit both in the random effects case and in the fixed effects case. What is even more surprising is that for both types of models there was a significant effect of educational level on student achievement of almost the same size as was found with the random effects model for the total sample. These results suggest that the estimates obtained with the random effects model for the total sample may be valid after all. This phenomenon was not found for any of the other variables.

4.5 Discussion

We posed two research questions. First, we wanted to establish whether effects of SEAS and teacher quality on mathematics achievement could be identified in country-level longitudinal analyses, and second whether such effects operated similarly in OECD and non-OECD countries? Our main reason for focusing on country-level panel data was that such data offer better opportunities for valid causal inference than cross-sectional data, because the longitudinal data makes it possible to partial out the effects of a wide range of observable and unobservable variables, which are the fixed characteristics of the participating countries.

The formal teacher characteristics of years of teaching experience and major academic discipline studied had no discernible effects. This may be due to the considerable heterogeneity among countries when it comes to the arrangements and quality of teacher education and of opportunities to learn from experience. The simple indicators employed here may thus be too blunt to capture those aspects of education and experience that are important to student achievement. There is also the related possibility that the effects vary across countries, preventing a common significant effect to appear. The lack of findings with the available variables must thus not be interpreted as supporting conclusions that teacher education and teacher experience are of no importance, but should rather be interpreted as indicating a need for further research.

The teachers' attained level of education, had, in contrast, strong effect on educational achievement. This may be because the ISCED scale on which level of education is expressed is well defined and therefore manages to capture within-country change. Note, however, that this relationship was not captured by the fixed effects model, which was because of a strong correlation between the common latent variable and level of education. However, when this correlation was removed by dividing the sample into OECD and non-OECD countries, the relationship reappeared within both categories of countries. This seems to be a case of Simpson's paradox (Simpson 1951), and there may be reason to look more closely into these kinds of complexities in future research.

In agreement with much previous research, we found quite substantial relations between student achievement and the amount of professional development activities that the teachers had participated in. The results also suggest that different domains

of professional development had differential impact, the strongest effects being found for development focusing on content and curriculum, while essentially no effect was found for generic skills, such as thinking skills and problem solving skills. These results also seem to agree with previous research.

We found teacher self-efficacy, as assessed by self-reports of preparedness for teaching in different domains, to have weakly positive, but insignificant relations with student achievement. There may, of course, be several reasons why these relations are so weak, but it is, again, reasonable to assume that measurement problems are important. Given that there are few common frames of reference among teachers for evaluating preparedness for teaching, it may be difficult to achieve sufficient reliability and validity to be able to investigate change over time.

SEAS was found to be a complex measure, unable to satisfy ideals of unidimensionality, and we also found its different components to be differentially related to achievement. The items referring to teacher knowledge and expectations were not related to student achievement, but the item reflecting parental support for student achievement was very strongly predictive of achievement and also had a weak relationship with students' desire to learn as assessed by teachers. While these results are not unreasonable, they conflict with much of the theory and research behind SEAS. Further research on the dimensionality and explanatory power of the SEAS construct is thus needed.

The comparisons between relations among variables in the groups of OECD and non-OECD countries showed these to be quite similar; no significant difference was identified. However, the limited number of observations in our study leads to such low statistical power that the chances of finding differences are limited, and it certainly is not possible to conclude that the lack of significant differences proves equality between OECD and non-OECD countries.

The methodology of our study is based on the fundamental premise that taking differences between multiple measures of the same units captures within-unit change over time. However, the actual technique with which this idea is implemented is neither transparent nor easily accessible. Nevertheless, the SEM techniques which have been applied do seem to solve the problems of estimation and testing, and offer a considerable amount of power, flexibility and generality; their potential certainly has not been exhausted. Further exploration of the advantages and disadvantages of using SEM to analyze country-level longitudinal data is encouraged.

4.6 Conclusions

The current study is based on 38 observations observed twice, although the fundamental data comprises almost 400,000 students, each observed once. In spite of these differences, there is agreement between the results from some analyses of the country-level data and the results from analyses of the student-level data. This is the

case, for example, for effects of professional development on student achievement. However, for other variables, the results from analyses of country-level data differ from the results from analyses reported in previous research. The most striking example of this is for SEAS, which in the country-level analyses was found to be multidimensional and where the components were related to achievement to strikingly different degrees. Further research is needed to clarify the meaning and importance of this finding.

References

- Bollen, K. A., & Brand, J. E. (2008). *Fixed and random effects in panel data using structural equations models*. California Center for Population Research. On-Line Working Paper Series. PWP-CCPR-2008-003. Retrieved from <http://papers.ccpr.ucla.edu/papers/PWP-CCPR-2008-003/PWP-CCPR-2008-003.pdf>.
- Creemers, B., & Kyriakides, L. (2010). Explaining stability and changes in school effectiveness by looking at changes in the functioning of school factors. *School Effectiveness and School Improvement*, 21(4), 409–427.
- Falck, O., Mang, C., & Woessmann, L. (2015). *Virtually no effect? Different uses of classroom computers and their effect on student achievement*. CESifo Working Paper, No. 5266. Retrieved from <https://www.cesifo-group.de/de/ifoHome/publications/working-papers/CESifoWP.html>
- Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. National Comprehensive Center for Teacher Quality, Washington, DC, USA. Retrieved from <http://www.gtlcenter.org/sites/default/files/docs/LinkBetweenTQandStudentOutcomes.pdf>
- Gustafsson, J. E. (2007). Understanding causal influences on educational achievement through analysis of differences over time within countries. In T. Loveless (Ed.), *Lessons learned: What international assessments tell us about math achievement* (pp. 37–63). Washington, DC, USA: The Brookings Institution.
- Gustafsson, J. E. (2013). Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement*, 24(3), 275–295.
- Hoy, W. K., Tarter, C. J., & Hoy, A. W. (2006). Academic optimism of schools: A force for student achievement. *American Educational Research Journal*, 43(3), 425–446.
- IEA. (2012). *International database analyzer (version 3.1)*. (Software) Hamburg, Germany: International Association for the Evaluation of Educational Achievement (IEA). Retrieved from <http://www.iea.nl/data.html>
- Martin, M. O., Foy, P., Mullis, I. V. S., & O'Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and*

- PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade. Implications for early learning* (pp. 109–178). Chestnut Hill, MA, USA: TIMSS & PIRLS International Study Center, Boston College
- Muijs, D. (2012). Methodological change in educational effectiveness research. In C. P. Chapman, P. Armstrong, A. Harris, D. R. Muijs, D. Reynolds, & P. Sammons (Eds.), *School effectiveness and improvement research, policy and practice: Challenging the orthodoxy* (pp. 58–66). Abingdon, UK: Routledge.
- Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Muthén, B., & Muthén, L. (1998–2014). *Mplus Version 7.3*. Los Angeles, CA: Muthén & Muthén.
- Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.
- Nilsen, T., & Gustafsson, J.E. (2014). School emphasis on academic success: Exploring changes in science performance in Norway between 2007 and 2011 employing two-level SEM. *Educational Research and Evaluation*, 20(4), 308–327. doi:<http://dx.doi.org/10.1080/13803611.2014.941371>
- Nordenbo, S. E., Holm, A., Elstad, E., Scheerens, J., Larsen, M. S., Uljens, M., Hauge, T. E. (2010). *Input, process, and learning in primary and lower secondary schools: A systematic review carried out for The Nordic Indicator Workgroup (DNI)* (Vol. 2010). Danish Clearinghouse for Educational Research, DPU, Aarhus University.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B*, 13, 238–241.
- Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research*, 83(3), 357–385.
- Wang, M.-T., & Degol, J. L. (2015). School climate: A review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review*, 1–38. doi:10.1007/s10648-015-9319-1
- Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 659–706.

The Importance of Instructional Quality for the Relation Between Achievement in Reading and Mathematics

Guri A. Nortvedt, Jan-Eric Gustafsson and Anne-Catherine W. Lehre

Abstract Students gain access to mathematical tasks through reading; consequently, low-performing readers generally perform low in mathematics. High quality instruction might help students develop comprehension strategies for reading mathematics that weakens the relationship between reading and mathematics skills. The main aim of this chapter is to investigate how instructional quality might moderate the relationship between reading and mathematics achievement. Analyzing data from 37 countries and benchmark participants who applied the same sample for TIMSS 2011 and PIRLS 2011, two different models were fitted to the data for each educational system: (1) a two-level confirmatory factor analysis (CFA) model for instructional quality and the correlation between instructional quality and reading and mathematics achievement at student and class levels, and (2) a two-level random slopes model in which the slope variation across classrooms was related to class-level instructional quality. In all educational systems, there was a strong positive correlation between reading comprehension and mathematics achievement. Further, a positive relation between instructional quality and mathematics and reading achievement was observed in a number of countries. The analysis of how instructional quality moderated the relationship between mathematics and reading was inconclusive. The influence of reading comprehension on mathematics achievement was significantly moderated by instructional quality in only six countries; nonetheless, the driving hypothesis should not be rejected.

G.A. Nortvedt (✉)
Department of Teacher Education and School Research,
University of Oslo, Oslo, Norway
e-mail: guri.nortvedt@ils.uio.no

J.-E. Gustafsson
Department of Education and Special Education, University of Gothenburg,
Gothenburg, Sweden
e-mail: jan-eric.gustafsson@ped.gu.se

J.-E. Gustafsson · A.-C.W. Lehre
Faculty of Educational Sciences, Centre for Educational Measurement
at the University of Oslo (CEMO), Oslo, Norway
e-mail: a.c.w.g.lehre@ils.uio.no

Keywords Instructional quality · Mathematics · Reading · Two-level structural equation modeling moderation models · Trends in Mathematics and Science Study (TIMSS 2011) and Progress in International Reading Literacy Study (PIRLS) 2011

5.1 Rationale

Mathematics achievement may be influenced by reading comprehension and instructional quality. Across the world, a major mandate for primary school teachers is to introduce all their young students to reading and mathematics, thus constructing the foundations for lifelong learning. The main aim of mathematics education is to develop students' mathematical competence (Niss and Højgaard 2011), namely to prepare students for further education and participation in society. Students should be prepared to use their mathematical competence in a variety of situations, to pose and solve problems, and to communicate and reason mathematically. For countries that participate in TIMSS, being able to solve both pure and applied mathematical problems is recognized as an important skill, which is in line with the definition of mathematical competence.

Reading plays a particular role in mathematical problem solving. Before students can apply their mathematical knowledge to solve a textbook or assessment item, they first need to access the problem to be solved, to understand it, and to plan how to proceed. This is generally accepted for problem solving and modeling (see, for instance, Lesh and Zawojewski 2007). Although the necessity for text comprehension is less recognized for given problems, students need to be able to read and comprehend symbolic mathematical language (Niss and Højgaard 2011). Consequently, all mathematical problem solving may be assumed to rest on students' reading skills as well as their mathematical proficiency, because they access the mathematical content through reading activities.

In primary school reading instruction, a shift from learning to read to reading to learn is usually seen around grade four (Murnane et al. 2012; Snow 2002). In the first primary school grades, a strong focus on first learning to read, that is, to decode, is followed by instruction directed toward developing students' reading comprehension in higher grades, to allow the students to apply their reading skills in other learning activities. Although primary school teachers' approaches to reading instruction differ, more effort is likely directed toward improving students' word-decoding skills than toward reading comprehension. Murnane et al. (2012) claimed that reading instruction is mainly based on literary texts, and less on science, civic, or social studies texts. Further, they claimed that the selected texts allow students few possibilities to grapple with deep comprehension. It may be questioned if such a focus benefits students when it comes to comprehending text presented in mathematics, where texts are typically short and need translated from symbolic language to everyday language.

Österholm (2005) found that the use of mathematical symbols was challenging to students in transition from secondary school to college. Language or text elements that demand deep reading strategies, such as the use of keywords, challenging syntax, or irrelevant information, influence students' success in solving mathematical problems, as is well-documented in prior research on primary school students (Abedi and Lord 2001; Cummins et al. 1988; Roth 2009; Säljö et al. 2009; Thevenot et al. 2007; Verschaffel et al. 2000; Vicente et al. 2007). Fuchs et al. (2015) even proposed that mathematical word problem solving is a form of text comprehension.

Another factor contributing to the difficulties students experience is the common belief that solving mathematical problems is about “doing numbers” (Fuentes 1998; Nortvedt 2010). This view might be shared by teachers and students. Previous research has demonstrated that teachers are concerned about the amount of text in mathematical assessments. When students struggle to solve word problems, teachers blame the amount of text, rather than recognizing students' lack of text comprehension strategies (Pearce et al. 2012). Teachers need to recognize the relationship between reading and mathematics and pay specific attention to it through the teaching-learning opportunities they offer their students. Giving special attention to reading might help students improve their mathematical problem solving (Glenberg et al. 2012; Thevenot et al. 2007).

We aim to investigate how instructional quality might moderate the relationship between reading and mathematics. A strong positive relationship between reading and mathematics indicates that students who are low performing in PIRLS are also most likely to perform low in TIMSS. Our driving hypothesis is that high-quality teaching can contribute to weakening the relationship between reading skills and mathematics achievement; that instruction might help grade four students overcome difficulties comprehending the items given in the TIMSS 2011 mathematics assessment. Three research questions are addressed:

- (1) *What is the class-level relationship between reading and mathematics achievement?*
- (2) *What is the class-level relationship between instructional quality and reading comprehension, and between instructional quality and mathematics achievement?*
- (3) *To what extent does instructional quality moderate the relationship between reading and mathematics achievement?*

5.2 The Influence of Reading on Mathematics Achievement

Previous research has demonstrated a moderate to strong positive relationship between reading and mathematics achievement (Adelson et al. 2015; Bernardo 2005; Moreau and Coquin-Viennot 2003; Nortvedt 2011; Vilenius-Tuohimaa et al.

2008). Typically, a correlation of 0.5–0.7 is observed at the student level. Most studies investigate the relationship between reading comprehension measures and some aspect of numeracy or mathematical word problem solving (Cummins et al. 1988; Palm 2008; Reusser and Stebler 1997; Thevenot et al. 2007; Verschaffel et al. 2000; Vilenius-Tuohimaa et al. 2008). Björn et al. (2014) found that text comprehension in grade four predicts mathematical word problem solving skills in secondary school. The more technical aspect of reading, namely decoding, is emphasized less often. However, even controlling for decoding, reading comprehension and mathematics have a positive relationship (Vilenius-Tuohimaa et al. 2008). The importance of reading comprehension was supported by Vista (2013), who, in a longitudinal study of Grade 3–8 Australian students, found that reading comprehension skills partially mediate the relation between problem solving ability and mathematical growth. It might be concluded that students who struggle to read will most likely also struggle to solve mathematical tasks.

Reading attention might also play a role in students' comprehension of mathematical tasks. Fuentes (1998) claimed that much of the issue with mathematical problem solving stems from students' beliefs about this activity. Many students believe that mathematics is about "doing numbers." Consequently, careful reading of the test items to unfold their underlying mathematical structure and identifying what to do is not the students' primary concern. Instead, they engage in what Verschaffel et al. (2000) termed "the word problem game," in which students typically identify an operation based on applying keywords as operation words instead of relationally. For instance, in word problems such as "Jane and Mark have 57 Euros altogether. Mark has 7 more Euros than Jane. How much does Jane have?" students who apply surface strategies typically treat 'altogether' and 'more' as operation words, signaling that the correct operation is addition. They will get 64 Euros as their solution to the item. Students who apply deep reading strategies will more likely treat the two keywords as relational words that explain the relationships between quantities and persons and get 25 Euros as the answer to their calculations.

Although the relationship between reading comprehension and mathematical performance is strong at the student level, Adelson et al. (2015) found that the relationship was even stronger at the school level, which they interpreted as indicating that the quality of instruction matters. They consequently proposed that reading strategies should be taught in mathematics classrooms. Indications that such interventions might be fruitful were documented by Glenberg et al. (2012). Initiating a small-scale intervention, Glenberg et al. (2012) improved students' mathematical performance by training Grade three and four students ($n = 58$) to apply a digital tool that assisted them in creating embodied mental models of problem texts, by manipulating on-screen pictures. Trained students, to a larger extent than the control group ($n = 39$), avoided using irrelevant numerical information in their solutions to the mathematical word problems, even without access to the tool. This indicates that training students to pay specific attention to content may later help them to identify the underlying mathematical structure in a problem, and to more easily discriminate between relevant and irrelevant information in the problem context.

Both successful and erroneous strategies displayed in student work in secondary or college education might stem from how students comprehend mathematical tasks during the primary school years. Nortvedt (2011), for instance, proposed that lower secondary school students with above-average mathematics and below-average reading achievement to some extent compensate for their weaker reading comprehension by recognizing stereotypical arithmetic word problems. However, at the same time, these students make mistakes more frequently than their peers with above-average reading and numeracy skills, because the students treat key words such as “altogether,” “less,” or “more” as signal words that indicate appropriate mathematical operations instead of as relational statements. Students at all ability levels make such mistakes, and it is likely that they are due to transferring simple surface strategies learned from early instruction, when word problems could be solved by applying one of the four operations, a well-known phenomenon (see, for instance, Cummins et al. 1988). That is, primary school mathematics education might support students in playing the “number game.” In addition, students of all ages tend to suspend sense-making and work more on solving problems than on comprehending them (Inoue 2005; Palm 2008; Schoenfeld 1992).

Lack of sense-making and overreliance on naïve methods might result from teaching–learning activities, and thus indicate issues with the quality of instruction. Graeber et al. (2012) analyzed video data from 69 teachers of grade four and six students, evaluating a total of 550 reading lessons and 600 mathematics lessons, to compare the quality of instruction in the two school subjects. They found that, overall, teachers who offered cognitively demanding teaching in the mathematics lessons were not always efficient reading teachers. Only 12 % of the teachers offered high-quality instruction in both subjects, offering teaching with high cognitive demand in terms of the quality of teacher questioning and content offered (i.e., demand of tasks and lesson content). A second important observation made by Glenberg et al. (2012) was that mathematics instruction seemed more consistent than reading instruction. For instance, while introducing reading material that had high cognitive demand, the teachers often failed to engage students in activities that reflected this level.

5.2.1 Relationships Among Reading and Mathematics: TIMSS and PIRLS 2011

When designing TIMSS 2011 and PIRLS 2011, many countries saw the opportunity to link these two studies to collect extensive information about the quality of instruction at the end of the early years of primary school (Mullis and Martin 2013). In total, 34 countries and three benchmarking entities took the opportunity to have the same grade four students take both assessments.

In PIRLS, reading achievement is reported on a scale comprising four benchmark levels. In total, 95 % of the grade four students achieve at or above the lowest benchmark, indicating that they can “locate and retrieve information from different parts of the text” (Mullis et al. 2012b). Analysis showed that reading competence is required for many of the tasks in grade four mathematics and science (Mullis et al. 2013). Mullis et al. (2013) divided the students into three groups according to their reading proficiency, and further analysis showed that better readers outperformed average readers, and average readers outperformed poor readers in mathematics. For most countries, better readers had a significant advantage over poor readers when doing mathematical tasks that had a high reading demand. Only the best readers performed well across all mathematical questions independently of the tasks’ reading demand. Less proficient readers performed relatively better on items that had a low reading demand than on items that had a high reading demand. In some countries (Austria, Chinese Taipei, Croatia, Hungary, Italy, Lithuania, Northern Ireland, Qatar, Romania, the Russian Federation, Saudi Arabia, Singapore, and the United Arab Emirates), the difference in mathematics achievement between good and poor readers was significant (Mullis et al. 2013).

Although a positive well-equipped school environment provides important support for teaching and learning, teacher quality is essential (Martin et al. 2013). Classroom teachers provide instruction directly to their students and thus influence the students’ learning environment. Teachers contribute positively when they are well-prepared and provide effective engaging instruction. Students engaged in their reading or mathematics lessons had higher scores on TIMSS and PIRLS 2011, compared to students who were “somewhat engaged” or “not engaged” in their lessons. Student engagement in reading, mathematics, and science was positively related to achievement in at least one subject in 17 countries and in all three subjects in nine countries.

To summarize, although there is a strong relationship between reading and mathematics outcomes at the student level, the relationship is even stronger at the class level (Adelson et al. 2015). Errors students make might stem from their prior experiences in the mathematics classroom and from too little focus on strategies for comprehending mathematical text. For instance, students reading at an intermediate level on PIRLS might make straightforward inferences from the text they read (Mullis et al. 2012b). However, mathematical problems might demand more advanced reading strategies. Much teaching does not offer a high cognitive demand (Graeber et al. 2012), which is most likely necessary for students to develop good comprehension skills in reading and mathematics. Glenberg et al. (2012) found that training students in comprehension strategies for reading mathematical word problems helped students develop strategies for reading mathematical word problems and become more successful problem solvers. Thus, prior research supports our hypothesis that high instructional quality might weaken the strong relationship between reading comprehension and mathematics achievement.



5.3 Method

Using data from the joint TIMSS and PIRLS 2011 database (available from <http://timssandpirls.bc.edu/timsspirls2011/international-database.html>), grade four students' outcomes in mathematics and reading were analyzed, applying multilevel structural equation modeling (SEM) with random slopes to investigate the moderating effects of instructional quality on the relationship between reading and mathematics outcomes. We aimed to capture variation across classrooms in the slope of the regression of mathematics achievement on reading achievement in a class-level latent variable and to investigate whether classroom instructional quality was negatively related to slope.

5.3.1 PIRLS 2011 and TIMSS 2011

PIRLS is an international, large-scale survey of students' reading literacy. First conducted in 2001, PIRLS assesses students in grade four every fifth year (Mullis et al. 2012b). Reading literacy is defined as

The ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment (Mullis et al. 2009a, p. 11).

Three aspects of reading literacy are assessed: (1) purpose of reading, (2) processes of comprehension, and (3) reading behaviors and attitudes. In the analysis in this chapter the achievement data from the reading test is taken as a measure of students' reading comprehension. In the following sections, we consequently refer to reading comprehension instead of reading literacy.

TIMSS and PIRLS use a matrix-sampling design, where each student is administered only a subset of the texts and associated items. All achievement scores are expressed on a common scale in the form of "plausible values," which are multiple imputed scores that take advantage of all available responses to test items and background variables (see, for instance, von Davier et al. 2009). There were five plausible values, and the information in all five was taken advantage of with the imputation procedure implemented in Mplus 7 (Muthén and Muthén 1998–2012).

5.3.2 Sample

In total, 34 countries and three benchmark participants applied the same sample for TIMSS and PIRLS 2011 (Martin and Mullis 2013), implying that a sampled student participated in both studies. These students constitute the sample used for the analyses reported in this chapter.

5.3.3 *Constructs*

Reading comprehension

The reading comprehension construct in PIRLS 2011 comprises four types of comprehension processes: (1) focus on and retrieve explicitly stated information, (2) make straightforward inferences, (3) interpret and integrate ideas and information, and (4) examine and evaluate content, language, and textual elements (Mullis et al. 2009a, p. 13).

Mathematics achievement

The grade four TIMSS 2011 assessment of student achievement in mathematics is used as a measure of students' mathematical competence in solving pure (given) and applied (word problems) mathematical tasks (Mullis et al. 2009b).

Instructional quality

Six parallel questions from the student questionnaires on TIMSS and PIRLS were used to measure instructional quality. All were four-category Likert items ranging from "disagree a lot" to "agree a lot," introduced by the statement "How much do you agree with these statements about your mathematics lessons/reading lessons?":

- MATEXP/RDTEXP: I know what my teacher expects me to do
- MATEASY/RDTEASY: My teacher is easy to understand
- MATISAY/RDTISAY: I am interested in what my teacher says.

In one measurement model, two instructional quality (InQua) factors were included: InQua-Math and InQua-Read. However, as the class-level correlation was close to unity in most countries, the two factors were collapsed into a global measure of instructional quality, InQuaB, which was related to all six items. Thus, based on the instructional quality items in the mathematics and reading questionnaire, a global class-level instructional quality measure (InQuaB) was estimated, while for the student level there were two correlated InQua variables, one for mathematics and one for reading. The factor loadings of the indicators of instructional quality on InQuaB were generally very high, often close to unity, even though there were some differences between educational systems.

Researchers have recently started to take advantage of students' ratings as a source of information about instructional quality. Although a single student's rating is not very dependable, the assessment becomes more reliable and valid when done by a whole class of students and when applied to more than one subject matter area. Scherer and Gustafsson (2015) used student questionnaire items concerning TIMSS and PIRLS 2011 reading, mathematics, and science teaching from three countries and demonstrated that a two-level latent variable modeling approach could separate different aspects of instructional quality. Here, we have adopted a similar, but simpler, approach that investigated only instructional quality in reading and mathematics.

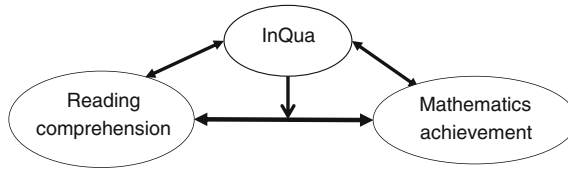


Fig. 5.1 Proposed research model describing the relationship between reading comprehension and mathematics achievement and the influence of InQua on this relationship at the class level

As student ratings are aggregated to the classroom level and both the mathematics and reading scales are included, we argue that applying data on instructional quality from the student questionnaire is valid for the analyses performed.

5.3.4 Analysis

Two-level SEM, distinguishing between student and class levels, was applied to investigate the relations among reading comprehension, mathematics achievement, and instructional quality. All five plausible values were included in the analysis of reading comprehension and mathematics achievement. Only class-level results are reported as this level is the focus of our study (Fig. 5.1).

First, a two-level confirmatory factor analysis (CFA) model for InQuaB and for the correlations between InQuaB and reading and mathematics achievement at the student and class levels was fitted. A separate model was fitted to the data for each educational system because metric invariance could not be established across all countries. These models all converged and fitted the data well, with model fit ranging from $CFI = 0.972$, $TLI = 0.948$, $RMSEA = 0.051$, $SRMR_{within} = 0.038$, $SRMR_{between} = 0.097$ (Chinese Taipei) to $CFI = 0.998$, $TLI = 0.996$, $RMSEA = 0.012$, $SRMR_{within} = 0.009$, and $SRMR_{between} = 0.066$ (Lithuania).

The second model was a two-level random slopes model, in which the slope variation for the regression of mathematics on reading comprehension across classrooms was related to class-level InQua in each educational system.

5.4 Results

We aimed to investigate the relationship between reading comprehension and mathematics achievement and how this relationship is influenced by InQuaB. We first report results from the two-level CFA model for instructional quality, investigating the correlation between InQuaB and reading and mathematics achievement at the class level. We then report results from the random slopes model focusing on the slope variation across classrooms as a function of InQuaB.

5.4.1 Relationship Among Mathematics Achievement, Reading Comprehension, and Instructional Quality

With the exception of Chinese Taipei, Singapore, Hong Kong, Malta, and Morocco, the average student score was higher in reading than in mathematics. At the classroom level, the correlation between mathematics and reading achievement ranged from 0.824 to 0.996 and was highly significant for all countries (Table 5.1). With the exception of three countries and one benchmark participant, the correlation between reading and mathematics was larger than 0.90. This is in good agreement with previous research (for instance, Adelson et al. 2015).

The relationship between InQuaB and mathematics achievement and reading comprehension at the class level was positive and significant for a number of educational systems (Australia, Botswana, Chinese Taipei, Dubai, Georgia, Hong Kong, Malta, Morocco, Oman, Portugal, Qatar, Romania, Singapore, and the United Arab Emirates), indicating that high-achieving classes value their teachers more than low-achieving classes in these countries (Table 5.1). In Azerbaijan and Saudi Arabia, the relationship was positive and significant for either mathematics or reading.

In Poland and Honduras, the correlation was negative and significant for mathematics and reading. Thus, in these two countries, low-achieving classes value their teacher and the instructional quality more than high-achieving classes do. The fact that low-performing classes tended to be more positive in their assessment of their teacher is apparent in other countries as well. In a number of educational systems (Austria, Croatia, the Czech Republic, Finland, Germany, Ireland, Quebec, the Russian Federation, Slovenia, the Slovak Republic, Spain, and Sweden) the overall correlation with InQuaB was negative, although not significant for reading achievement, mathematics achievement, or both. In addition, a number of educational systems had a non-significant positive relationship (Abu Dhabi, Hungary, Iran, Italy, Lithuania, Northern Ireland, and Norway). These non-significant relationships, both positive and negative, between InQua and achievement make some of the observed patterns difficult to interpret.

5.4.2 The Influence of Instructional Quality on the Relationship Between Reading Comprehension and Mathematics Achievement

In the final step, random slopes models were estimated by regressing the latent slope variable on InQuaB. All models converged nicely, but the regression coefficient was significant in only six cases (Table 5.1). This is probably because the statistical power was low in these models due to a limited number of students in each classroom. According to the driving hypothesis, the estimate of the regression coefficient slope should be negative if instructional quality weakens the influence of

Table 5.1 International averages of mathematics and reading achievement for 37 educational systems in TIMSS 2011 and PIRLS 2011 grade four, along with relationships between mathematics achievement, reading achievement, and instructional quality

Educational system	International averages		Measurement models			Random slope models
	Math	Read	Math with read	Math with InQuaB	Read with InQuaB	Slope on InQuaB
Azerbaijan	463	462	0.878**	0.292**	0.195	-0.110
Australia	516	527	0.990**	0.220**	0.222**	0.017
Austria	508	529	0.974**	-0.169	-0.253	0.024*
Botswana	419	419	0.983**	0.692**	0.699**	-0.089
Chinese Taipei	591	533	0.966**	0.351*	0.334*	0.089
Croatia	490	553	0.957**	-0.102	-0.090	-0.116**
Czech Republic	511	545	0.979**	0.002	-0.028	0.069**
Finland	545	568	0.967**	-0.021	-0.014	-0.016**
Georgia	450	488	0.970**	0.539**	0.499**	-0.158
Germany	528	541	0.988**	-0.042	-0.046	0.163
Honduras	396	450	0.950**	-0.296*	-0.390**	0.034
Hong Kong	602	571	0.969**	0.527**	0.512**	0.002
Hungary	515	539	0.989**	0.018	0.011	0.181
Iran	431	457	0.948**	0.077	0.155	-0.098
Ireland	527	552	0.928**	-0.048	-0.082	0.182
Italy	508	541	0.931**	0.207	0.221	0.043
Lithuania	534	528	0.987**	0.130	0.124	0.195
Malta	496	477	0.938**	0.426**	0.430**	0.012
Morocco	335	310	0.938**	0.319**	0.344**	0.029
Oman	385	391	0.957**	0.356**	0.380**	-0.056
Norway	498	507	0.955**	0.195	0.244	0.167
Poland	481	526	0.987**	-0.281*	-0.290*	0.144
Portugal	532	541	0.977**	0.290**	0.265**	0.017
Qatar	413	425	0.971**	0.333**	0.409**	-0.323
Romania	482	502	0.958**	0.303**	0.352**	0.021
The Russian Federation	542	569	0.930**	-0.041	-0.120	-0.051
Saudi Arabia	410	430	0.851**	0.225	0.444**	0.343
Singapore	606	567	0.996**	0.359**	0.330**	0.007
Slovak Republic	507	535	0.960**	0.003	-0.056	0.069
Slovenia	513	530	0.984**	-0.018	0.004	-0.002
Spain	482	513	0.884**	0.048	-0.002	-0.005*
Sweden	504	542	0.968**	-0.131	-0.172	-0.021**
UAE	434	539	0.974**	0.200**	0.218**	-0.075
Northern Ireland	562	558	0.954**	0.071	0.020	0.019

(continued)

Table 5.1 (continued)

Educational system	International averages		Measurement models			Random slope models
	Math	Read	Math with read	Math with InQuaB	Read with InQuaB	Slope on InQuaB
Dubai	468	476	0.987**	0.244**	0.264*	-0.302
Abu Dhabi	417	424	0.962**	0.172	0.190	-0.019
Quebec	533	538	0.824**	-0.073	0.115	-0.083

Note Math = mean of TIMSS 2011 mathematics achievement; Read = mean of PIRLS 2011 reading achievement; Math with Read = class-level correlation between mathematics and reading achievement; Math with InQuaB = class-level correlation between mathematics achievement and instructional quality; Read with InQuaB = class-level correlation between reading achievement and instructional quality; Slope on InQuaB = β value for the regression of InQuaB on the relationship between mathematics and reading achievement

*significant at the 0.05 level

**significant at the 0.01 level

Estimates for residuals are not reported because all residuals were estimated to be 0, even the significant residuals

reading comprehension on mathematics achievement. In four of the countries (Croatia, Finland, Spain, and Sweden), this was the case. In two of the countries (Austria and the Czech Republic), the influence of InQuaB was positive, meaning that the relationship between reading and mathematics was strengthened. The correlations between InQuaB and reading comprehension and mathematics skills were nonsignificant for all six countries where an effect was observed, indicating that the low-achieving classes in these countries resemble their peers in higher-achieving classes in terms of how they judge the instructional quality.

For the countries with significant positive relations between InQuaB and reading and mathematics achievement, the estimate of the slope on InQuaB was nonsignificant for all cases. The overall picture is mainly one of nonsignificant effects of InQuaB on the latent slope variable. This is probably because there was very little estimated slope variability across classrooms in most countries. This, in turn, is likely because the number of students in each classroom is so limited that it is difficult to achieve sufficient statistical precision in the estimation of slope variability.

5.5 Discussion and Concluding Remarks

Overall, a strong correlation between reading and mathematics was observed at the classroom level. However, a significant positive correlation between instructional quality and reading and mathematics was observed in fewer than half the participating educational systems. Moreover, an effect of instructional quality on the relationship between reading comprehension and mathematics skill was observed for

only six countries, all European. With the exception of Spain, the overall correlation between mathematics and reading achievement at the class level is in the middle range compared to the other participating educational systems (Table 5.1). In all six countries, students scored significantly above the international mean in reading (see Mullis et al. 2012b, p. 38), with reading scores ranging from 513 to 568. Some of the six countries are among the top performers on PIRLS 2011. In comparison, a range of average mathematics achievement was observed (Table 5.1), with Spain and Croatia scoring significantly below average, Sweden at the average, and Austria, the Czech Republic, and Finland significantly above the international average (see Mullis et al. 2012a, p. 90). The six countries where instructional quality had a moderating effect on the relation between reading and mathematics are all educational systems in which reading instruction is more successful compared to mathematics instruction, judging by the international results on TIMSS 2011 and PIRLS 2011.

Mullis et al. (2013), who also used the joint TIMSS and PIRLS database, investigated whether students at different reading levels could cope to the same extent with mathematics items at both high and low reading levels. They found that, in a few educational systems, students at a low reading level performed at similar levels for low- and high-demand reading-level mathematics items, indicating high quality instruction related to the reading aspects of doing mathematics. This research outcome might support our driving hypothesis regarding instructional quality. Overall, students at a low reading level scored significantly lower in mathematics than their peers at a high reading level (Mullis et al. 2013), but in Finland and Sweden, for instance, students at a low reading level not only scored at a similar level for high- and low-reading demand mathematics items, they even raised the success rate, moving from mathematics items with low reading demand to items with a high reading demand. Finland and Sweden were among the few countries where instructional quality weakened the relationship between reading and mathematics. Nonetheless, no direct conclusions should be drawn from Mullis et al. (2013) with respect to the outcomes of our study. Croatia, for instance, where instructional quality significantly weakened the relation between reading and mathematics, is among the countries where there was a remarkable drop in the success rate when low-level readers moved from mathematics items with low reading levels to items with a higher reading level.

The relationship between instructional quality and achievement (Table 5.1) is difficult to interpret. In educational systems where this relationship is negative, low-achieving classes perceive the instruction their teachers provide to be of higher quality than do their higher-achieving peers. Reading instruction should target deep reading strategies and include strategies for comprehending mathematical texts for all students, as proposed by Murnane et al. (2012). Otherwise, it is unlikely that this instruction will provide students with strategies for comprehending mathematical tasks. The student questionnaire items that together measure instructional quality do not address strategy instruction as such. Instead, students are asked about how they conceive their mathematics and reading lessons. Still, it may be argued that students and classes who feel their teacher is easy to understand and interesting, and who

think they know what their teacher expects of them during lessons report perceived quality in the instruction their teachers deliver. The varying sign and strength of the relationship between instructional quality and achievement across countries suggest that the student assessments are influenced by response styles and other factors that affect the estimated relationship. Further research on this issue is needed.

Some teachers “blame” students’ reading level and the amount of text in mathematical tasks for their students’ shortcomings with mathematical problem solving (Pearce et al. 2012). Prior research has demonstrated that when comparing the reading and mathematics instruction delivered by primary school class teachers, teachers are usually more skilled in teaching either reading or mathematics; they rarely teach both subjects equally well (Graeber et al. 2012).

Finally, some potential consequences for instruction and policy making should be discussed. First, reading is fundamental to further learning. Students who are better readers seem to be better equipped for learning in other subjects, including mathematics. In the educational systems, a high focus on reading interventions may be a good investment in students’ futures and an essential part of their lifelong learning processes. We thus advocate Murnane et al. (2012) view that reading instruction in primary school should include texts from other school subjects, including mathematical texts. However, early emphasis on reading skills is less related to achievement in TIMSS and PIRLS 2011 than teachers’ promotion of student engagement (Martin et al. 2013, p. 115). Even after controlling for home background, engaging instruction has a predictive positive effect on achievement. Students engaged in their lessons have higher achievement than students with lower or no engagement.

Thus, active teachers who make sure students know what is expected of them, strive to be easily understood, present content in engaging ways, and generally manage to maintain their students’ motivation will positively promote their students’ achievements (Martin et al. 2013, p. 115). Clearly, high-quality teaching matters to student learning. According to Scherer and Gustafsson (2015), constructing a differentiated measure for instructional quality at the class level that combines the instructional quality constructs from different subject matter domains could enable detection of different aspects of instructional quality. When the moderating effect of instructional quality on the relationship between reading and mathematics was inconclusive, this is most likely due to a combination of reliability and validity issues in the assessment of instructional quality and the lack of statistical power. Although the outcome of the analysis is inconclusive, the driving hypothesis is not rejected. Instead, more research is needed to further disentangle how instruction might weaken the influence of students’ reading comprehension to their mathematical problem solving. This research could take as its point of departure that some educational systems have students at low reading levels that perform equally well on low-reading demand and high-reading demand mathematics items as shown by Mullis et al. (2013) and the outcomes of the analysis reported in this study.



References

- Abedi, J., & Lord, C. (2001). The language factors in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234.
- Adelson, J. L., Dickinson, E. R., & Cunningham, B. C. (2015). Differences in the reading–mathematics relationship: A multi-grade, multi-year statewide examination. *Learning and Individual Differences, 43*, 118–123.
- Bernardo, A. B. I. (2005). Language and modelling word problems in mathematics among bilinguals. *The Journal of Psychology, 139*(5), 413–425.
- Björn, P. M., Aunola, K., & Nurmi, J.-E. (2014). Primary school text comprehension predicts mathematical word problem-solving skills in secondary school. *Educational Psychology, 36*, 1–16. doi:10.1080/01443410.2014.992392
- Cummins, D. D., Kintsch, W., Reusser, K., & Wiemer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*(4), 405–438.
- Fuchs, L. S., Fuchs, D., Compton, D. L., Hamlett, C. L., & Wang, A. Y. (2015). Is word-problem solving a form of text comprehension? *Society for the Scientific Study of Reading, 19*(3), 204–223.
- Fuentes, P. (1998). Reading comprehension in mathematics. *The Clearing House, 72*(2), 81–88.
- Glenberg, A., Wilford, J., Gibson, B., Goldberg, A., & Zhu, X. (2012). Improving reading to improve math. *Scientific Studies of Reading, 16*(4), 316–340.
- Graeber, A. O., Newton, K. J., & Chambliss, M. J. (2012). Crossing the borders again: Challenges in comparing quality in instruction in mathematics and reading. *Teachers College Record, 114*, 1–30.
- Inoue, N. (2005). The realistic reasons behind unrealistic solutions: The role of interpretive activity in word problem solving. *Learning and Instruction, 15*, 69–83.
- Lesh, R. A., & Zawojewski, J. (2007). Problem solving and modeling. In F. K. J. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 2, pp. 763–804). Charlotte, NC: Information Age.
- Martin, M. O., Foy, P., Mullis, I. V. S., & O’Dwyer, L. M. (2013). Effective schools in reading, mathematics and science at the fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement—implications for early learning* (pp. 109–178). Boston, MA: TIMSS & PIRLS International Study Center.
- Martin, M. O., & Mullis, I. V. S. (2013). *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—Implications for early learning*. Boston, MA: TIMSS & PIRLS International Study Center.
- Moreau, S., & Coquin-Viennot, D. (2003). Comprehension of arithmetic word problems by fifth-grade pupils: Representations and selection of information. *British Journal of Educational Psychology, 73*(1), 109–121.

- Mullis, I. V. S., & Martin, M. O. (2013). TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement-implications for early learning. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement-implications for early learning* (pp. 1–11). Boston, MA: TIMSS & PIRLS International Study Center.
- Mullis, I. V. S., Martin, A. J., Kennedy, A. M., Trong, L., & Sainsbury, M. (2009a). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009b). *TIMSS 2011 assessment frameworks*. Boston, MA: TIMSS & PIRLS International Study Center.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012a). *TIMSS 2011 international results in Mathematics*. Boston, MA: TIMSS & PIRLS International Study Center.
- Mullis, I. V. S., Martin, A. J., Foy, P., & Drucker, K. T. (2012b). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.
- Mullis, I. V. S., Martin, M. O., & Foy, P. (2013). The impact of reading ability on TIMSS mathematics and science achievement at the fourth grade: An analysis by item reading demands. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement-implications for early learning* (pp. 67–108). Boston, MA: TIMSS & PIRLS International Study Center.
- Murnane, R., Sawhill, I., & Snow, C. (2012). Literacy challenges for the twenty-first century: Introducing the issue. *The Future of Children*, 22(2), 3–15.
- Muthén, L., & Muthén, B. (1998–2012). *Mplus user's guide* (7th Edn.). Los Angeles, CA: Muthén & Muthén.
- Niss, M., & Højgaard, T. (2011). *Competencies and mathematical learning. Ideas and inspiration for the development of mathematics teaching and learning in Denmark* (English ed.). Roskilde, Denmark: IMFUMFA.
- Nortvedt, G. A. (2010). Understanding and solving multistep arithmetic word problems. *Nordic Studies in Mathematics Education*, 15(3), 23–50.
- Nortvedt, G. A. (2011). Coping strategies applied to comprehend multistep arithmetic word problems by students with above-average numeracy skills and below-average reading skills. *Journal for Mathematical Behavior*, 30(3), 255–269. doi:10.1016/j.jmathb.2011.04.003
- Österholm, M. (2005). Characterizing reading comprehension of mathematical texts. *Educational Studies in Mathematics*, 63, 325–346.
- Palm, T. (2008). Impact of authenticity on sense making in word problem solving. *Educational Studies in Mathematics*, 67(1), 37–58.
- Pearce, D. L., Bruun, F., Skinner, K., & Lopez-Mohler, C. (2012). What teachers say about student difficulties solving mathematical word problems in Grades 2–5. *International Electronic Journal of Mathematics Education*, 8(1), 1–17.
- Reusser, K., & Stebler, R. (1997). Every word problem has a solution: The social rationality of mathematical modeling in schools. *Learning and Instruction*, 7(4), 309–327.
- Roth, W.-M. (2009). On the problematic of word problems-language and the world we inhabit. In L. Verschaffel, B. Greer, W. Van Dooren, & S. Mukhopadhyay (Eds.), *Words and worlds. Modelling verbal descriptions of situations* (pp. 55–69). Rotterdam, The Netherlands: Sense.
- Säljö, R., Riesbeck, E., & Wyndhamn, J. (2009). Learning to model: Coordinating natural language and mathematical operations when solving word problems. In L. Verschaffel, B. Greer, W. Van Dooren, & S. Mukhopadhyay (Eds.), *Words and worlds. Modelling verbal descriptions of situations* (pp. 177–193). Rotterdam, The Netherlands: Sense.
- 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, 1550.
- Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334–370). New York, NY: MacMillan.

- Snow, C. E. (2002). *Reading for understanding: Toward a R&D program in reading comprehension*. Pittsburg, PA: RAND Education.
- Thevenot, C., Devidal, M., Barrouillet, P., & Fayol, M. (2007). Why does placing the question before an arithmetic word problem improve performance? A situation model account. *Quarterly Journal of Experimental Psychology*, 60(1), 43–56.
- Verschaffel, L., De Corte, E., & Greer, B. (2000). *Making sense of word problems*. Lisse, The Netherlands: Swets & Zeitlinger.
- Vicente, S., Orrantia, J., & Verschaffel, L. (2007). Influence of situational and conceptual rewording on word problem solving. *British Journal of Educational Psychology*, 77, 829–848.
- Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J.-E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, 28(4), 409–426.
- Vista, A. (2013). The role of reading comprehension in maths achievement growth: Investigating the magnitude and mechanism of the mediating effect on maths achievement in Australian classrooms. *International Journal of Educational Research*, 62(6), 21–35.
- von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series*, 2, 9–36.



The Relation Between Students' Perceptions of Instructional Quality and Bullying Victimization

Leslie Rutkowski and David Rutkowski

Abstract Instructional quality may serve as a protective factor against school bullying victimization internationally. This study investigated this using the data provided by TIMSS 2011 fourth grade students. Given the highly-skewed distribution of the bullying scale and the clustered structure of the TIMSS data, a multilevel (students nested in classes) zero-inflated Poisson regression was used and responses to the bullying items were treated as rough counts. Covariates identified as predicting bullying at the international level were controlled for. Findings from the international model indicate that better instructional quality is associated with lower rates of student self-reported bullying victimization. At the educational-system level findings are mixed. The analysis suggests that bullying begins at an early age and that, at the fourth grade level, bullying victimization is an international phenomenon. Although instructional quality is associated with lower reported bullying victimization rates internationally, cross-system differences point to the important fact that instructional quality will not, in and of itself, globally lower rates of bullying in schools.

Keywords Instructional quality • Bullying victimization • Student characteristics • Zero-inflated poisson regression • Trends in Mathematics and Science Study (TIMSS) 2011

6.1 Introduction

Over the past two decades, there has been rapid growth in understanding of bullying in schools and its many negative effects on victims (Dill et al. 2004; Jimerson et al. 2010; Olweus 1994). Bullying is a global phenomenon affecting students at all levels of achievement and socioeconomic status (Harel-Fisch et al. 2011;

L. Rutkowski (✉) • D. Rutkowski
Faculty of Educational Sciences, Centre for Educational Measurement
at the University of Oslo (CEMO), Oslo, Norway
e-mail: leslie.rutkowski@cemo.uio.no

Rutkowski et al. 2013a, b), and that immigrant students are at greater risk of deleterious effects (Rutkowski et al. 2013a, b). Based on the results of a cross-national review, Jimerson et al. (2010, p. 1) contended that “studies in all countries in which bullying has been investigated, have revealed the presence of bullying.” An increased awareness of the prevalence and impacts of bullying have accompanied a growth in prevention and related initiatives. For example, at the international nongovernmental level, nonprofit organizations, such as the International Bullying Prevention Association and No Bully, have been established to share information and work with teachers and parents from around the world to prevent and combat bullying in schools. In addition, the United Nations envoy on violence against children recently stated that bullying in schools is a “serious concern” that threatens victims’ fundamental rights to education (UN News 2015).

Related to international attention on bullying, there have also been a number of nationally-focused studies on bullying in schools and associated correlates (Bosworth et al. 1999; Haynie et al. 2001; Nansel et al. 2003; Wang et al. 2009); however, at the international level, there remain a dearth of studies examining factors associated with bullying across a large group of countries. In this study, we aim to add to international conversation and to pursue one possible correlate of bullying that does not receive a great deal of focus: teacher instructional quality.

In its most general form, bullying is understood as a behavior intended to inflict injury or discomfort upon another individual (Olweus 1994). Within the context of schools, Olweus (2010) noted that an important aspect of school bullying victimization is the exposure to “negative or aggressive acts that are carried out repeatedly and over time” (p. 11). Indeed, in much of the bullying literature, the repeated nature of the aggressive acts is a key component defining bullying (see also Cook et al. 2010). Furthermore, Olweus (2010, p. 10) recognized bullying as “a subset of aggression or aggressive behavior,” as did the US Centers for Disease Control (CDC), which included bullying as an example of violent behavior (CDC n.d.).

As prominent figures in the classroom and in children’s day-to-day lives, teachers play an important role in students’ well-being and development around the world (OECD 2005). Evidence suggests that positive student teacher relationships are associated with students’ self-esteem, academic motivation, and achievement (Frymier and Houser 2000; Skinner and Belmont 1993). Further, support from teachers has been shown to reduce student aggression (Reinke and Herman 2002) and to decrease the risk of bullying (Natvig et al. 2001). Unsurprisingly then, teachers are key players in reducing bullying prevalence within schools (Allen 2010; Crothers and Kolbert 2010). And research clearly shows that teachers are important actors in both the intervention process when bullying occurs and in preventing bullying victimization from occurring in the first place (Nicolaidis et al. 2002; Yoon and Kerber 2003).

To date, much research on the relationship between teachers and classroom bullying has largely centered on teachers’ classroom management, with a specific

emphasis on discipline practices within the classroom (Bullock 2002; Smokowski and Kopasz 2005). Advocates of a classroom management approach to ameliorating bullying argue that “teachers who are adept at managing student behavior in the classroom work to prevent student bullying through creating a classroom climate incompatible with peer victimization... work to improve children’s social skills and conflict management skills so that future bullying is less likely” (Crothers and Kolbert 2010, p. 537). Notable here is that these and other bullying studies typically place instructional quality under the umbrella of classroom management. This is in contrast to the operational definition of instructional quality used here, where instructional quality is partly comprised of classroom management (see Chap. 1). Nonetheless, these studies show the importance of teachers, with respect to school bullying victimization.

Although such studies clearly place the teacher in a central role with respect to bullying prevention and prevalence, simply viewing the teacher as a “disciplinary manager” ignores other important dimensions. In other words, teachers do more than manage disciplinary issues. This is in line with Barbetta et al. (2005, p. 17), who posited that “the first line of defence in managing student behaviors is effective instruction.” In a recent study, Kyriakides et al. (2014) found support for this argument by employing the dynamic model of educational effectiveness to design strategies and actions to counter bullying, and found that schools who support their teachers in developing an optimal and safe classroom learning environment via high quality teaching (amongst other things) may reduce bullying. To that end, they wrote: “Provision of learning opportunities for students is one of the most important aspects of school policy on teaching when dealing with bullying” (Kyriakides et al. 2014, p. 457). Hence, teachers and their instruction play an important role, not just for achievement, but also for preventing bullying. As such, our objective here was to determine the degree to which instructional quality (as defined and described in Chap. 1) is associated with less school bullying victimization internationally. The sample includes fourth grade students and their teachers who participated in TIMSS 2011. We selected fourth grade students as this is a relatively understudied age group in the international literature and because interventions are more effective at earlier ages (Smith 2010). Given the highly-skewed distribution of the bullying scale (described subsequently) and the clustered nature of the TIMSS data, we used a multilevel (students nested in classes) zero-inflated Poisson regression and treated responses to the bullying items as rough counts. We controlled for several covariates that have been found in previous research (Rutkowski et al. 2013a, b) to predict bullying at the international level. These covariates included sex of the student, student attachment to the school, student sociocultural capital, and student immigrant background.



6.2 Methods

6.2.1 Data

In this study, we consider only 48 of the 49 TIMSS participating fourth grade systems (excluding special administrative units such as the Flanders region of Belgium, the Basque Country, and Northern Ireland). We omitted Australia from our analysis because there was variance in the class-level sampling weights due to a design feature involving the indigenous population. As such, it was not possible to fit a multilevel model to these data while also properly incorporating the design features of the study.

For the models fit to the data (discussed subsequently), we treated the data as students nested within classes, recognizing that we confounded the class- and school-level variance in those countries that sample more than one class per school.

6.2.2 Measures

We used TIMSS 2011 grade four student background questionnaire responses for the entire international sample, excepting omissions.

Student measures

The TIMSS student questionnaire (Foy et al. 2013) features a six-item scale that aligns with our research question on bullying victimization. This scale is a modification of the Olweus (2007) bullying scale, with all relevant items under a single stem that asks “During this year, how often have any of the following things happened to you at school?” The individual items include (1) “I was made fun of or called names;” (2) “I was left out of games or activities by other students;” (3) “Someone spread lies about me;” (4) “I was hit or hurt by other students(s) (e.g. shoving, hitting, kicking);” (5) “I was made to do things I didn’t want to do by other students;” and (6) “Someone spread lies about me.” Students responded to one of four options: “at least once a week”; “once or twice a month”; “a few times a year”; and “never.” We found reasonable evidence that the bullying scale can be regarded as scalar invariant across all considered countries (see Appendix D, Table D.1). As such, these items were summed to create a scale from zero (when all items were ticked “never”) to 18 (when all items were ticked “at least once a week”). The reliability of this scale in the TIMSS international sample, as measured by Guttman’s λ_2 was 0.76. Although TIMSS produces a bullying scale, we opted to create our own measures because the TIMSS-produced scale is the result of an item response theory model that assumes the underlying latent variable (bullying experiences) is normally distributed. Given the frequency scale of these indicators and our interest in understanding frequency of

bullying experiences, we created our own scale that approximates count data that is best fit by a Poisson regression (Agresti 2002). The weighted average counts and standard deviations of bullying experiences in the TIMSS 2011 grade four sample (Table 6.1) show that there are meaningful differences in terms of the average levels of reported bullying victimization, from countries with relatively low average counts of bullying ($\bar{x}_{\text{Azerbaijan}} = 2.23, SD = 3.33$; $\bar{x}_{\text{Armenia}} = 2.16, SD = 3.44$) to countries with relatively high average counts ($\bar{x}_{\text{Thailand}} = 8.01, SD = 4.30$; $\bar{x}_{\text{Botswana}} = 6.56, SD = 4.79$). For reference, the pooled international average is 4.91 ($SD = 4.35$). (We further discuss the distribution of this scale in Sect. 6.2.3.)

As the primary focus of our research question, we used students' perception of instructional quality (InQua) as a predictor in our model. The variables included asked about the degree to which students agreed that, in their math lessons: (a) they know what their teacher expects them to do; (b) their teacher is easy to understand; (c) they are interested in what their teacher is saying; and (d) their teacher gives

Table 6.1 Descriptive statistics of bullying scale, by country

Country	n	Mean	SD	Country	n	Mean	SD
Azerbaijan	4882	2.23	3.33	Malta	3607	5.21	4.18
Austria	4668	4.29	4.20	Morocco	7841	5.81	4.52
Bahrain	4083	6.56	4.79	Oman	10,411	6.17	4.34
Armenia	5146	2.16	3.44	Netherlands	3229	4.73	3.85
Botswana	4198	7.95	3.68	New Zealand	5572	6.34	4.57
Chile	5585	6.03	4.87	Norway	3121	4.10	3.74
Taiwan	4284	4.31	4.09	Poland	5027	3.58	3.95
Croatia	4584	3.46	3.57	Portugal	4042	4.50	3.88
Czech Republic	4578	4.80	4.08	Qatar	4117	6.67	4.82
Denmark	3987	3.45	3.42	Romania	4673	4.76	4.19
Finland	4638	3.46	3.39	Russia	4467	4.74	3.94
Georgia	4799	3.03	3.64	Saudi Arabia	4515	5.55	4.56
Germany	3995	4.41	3.89	Serbia	4379	3.69	3.80
Honduras	3919	5.83	4.75	Singapore	6368	5.37	4.15
Hong Kong	3957	4.49	3.94	Slovakia	5616	4.81	4.21
Hungary	5204	5.40	4.27	Slovenia	4492	4.60	4.18
Iran	5760	5.14	4.21	Spain	4183	5.16	4.39
Ireland	4560	3.40	3.82	Sweden	4663	2.81	3.14
Italy	4200	4.34	3.93	Thailand	4448	8.01	4.30
Japan	4411	4.46	4.07	UAE	14,720	6.03	4.53
Kazakhstan	4382	3.34	3.97	Tunisia	4912	5.16	4.05
Korea	4334	4.13	3.84	Turkey	7479	5.96	4.65
Kuwait	4142	5.52	4.71	USA	12,569	4.55	4.43
Lithuania	4688	4.53	3.93	Yemen	8058	5.06	4.63

Note Weighted average counts (n), mean and standard deviation (SD) of bullying experiences in school. *UAE* United Arab Emirates

them interesting things to do. Items (a) and (b) tap into clarity of instruction as described in Chap. 1, while items (c) and (d) tap into the aspect of instructional quality referred to as supportive climate. A supportive climate refers among other things to teachers who support students by engaging them. This scale hence misses the aspects of instructional quality referring to cognitive activation and classroom management.

Although TIMSS also measures teachers with respect to instructional quality, we opted for the student report, as students are less prone to answering in socially desirable ways (Wagner et al. 2015).

As a proxy for sociocultural capital, we used the books in the home variable, scaled from 0 to 4, which is coded such that 0 corresponds to few books and 4 corresponds to more than 200 books. As a measure of the students' attachment to school, we used the average of two variables that asked students how much they agree that: (1) they like being in school; and (2) they feel like they belong at this school. These are Likert scaled variables, where 0 = strongly disagree and 3 = strongly agree. This short scale had an estimated international reliability of 0.77. Although the TIMSS data set has no direct measure of immigrant status, we used the frequency of speaking the language of the test at home as a rough proxy, with 0 = never; 1 = sometimes; and 2 = always or almost always. Finally, we included the student's sex such that 0 = male and 1 = female.

6.2.3 Analytic Methods

Given the inherent multilevel structure of the data (students nested in classes, classes nested in schools, schools nested in countries), we pursued a multilevel approach. Although intraclass correlations (ICCs) were relatively low, ranging from 0.01 to 0.03, the standard errors around each system's intercept variance estimate are quite small relative to the variance estimate, providing some evidence that there are meaningful between-group differences in average log-counts of bullying. Although these values are not included in the interest of space, they are available upon request from L. Rutkowski. Further, the nature of the bullying victimization scale (frequency of occurrence) resulted in a non-normal distribution. Rather, the data more closely followed a Poisson distribution, which is sensible if the scale roughly represents counts of bullying victimization experiences. Finally, the occurrence of bullying victimization in schools is, fortunately, a relatively rare occurrence, leading to more zeros than is normally expected in a Poisson distribution (Fig. 6.1). Assuming a Poisson distribution with an overall empirical mean (λ) of 4.91 (SD = 4.35) and $n = 247,338$, the number of zeros is normally expected to be $ne^{-\lambda} = 247,338 e^{-4.91} = 1824$ (Lambert 1992). Instead, we found 45,653 zeros. This confirmed our suspicion that there were too many zeros and that a typical multilevel Poisson model will not suffice. We thus chose a multilevel

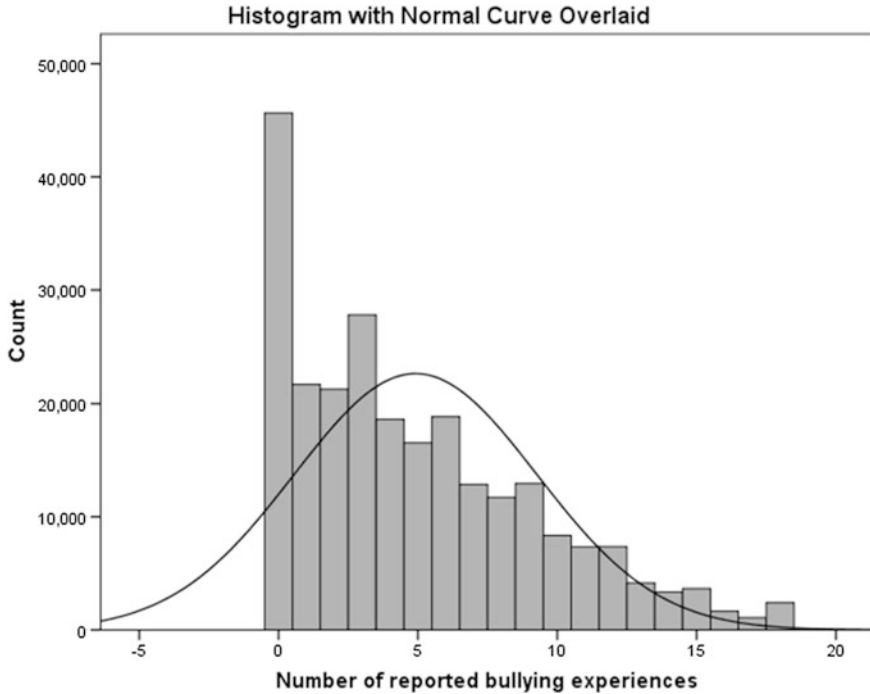


Fig. 6.1 Normal density curve overlying histogram of bullying scale distribution

zero-inflated Poisson (M-ZIP) model for our analysis, given the distribution of the outcomes and the structure of the data. In a single-level ZIP model, it is assumed that there are two separate processes at work: a latent binomial regression that predicts whether someone is in the zero category (no occurrence) and a standard count or Poisson regression (frequency of occurrence; Lambert 1992). A convenient feature of a ZIP model is that the variables that explain the zero part and count part of the model do not have to be the same. Given that our interest was in the count part of the model, we did not build a model for the zero part; however, we did estimate the coefficient associated with the odds of having no bullying victimization experiences. The model is a mixture of a Poisson distribution with parameter λ and a degenerate distribution with point mass at 0 and probability p . When excess zeros are present, a ZIP model is a better fit to the data and better predicts both zeros and counts (Hall 2000; Lambert 1992).

Although two-level M-ZIP models are theoretically well established and relatively easy to implement in commercially available software, there is little practical capacity for higher-level M-ZIP models; we thus chose to fit two sets of two-level M-ZIP models. In the first case, we fitted one pooled international model where students were nested within classrooms, and educational system was used as a clustering variable. Secondly, we produced individual models for each country

where students were nested within classes. In all cases, we assumed that InQua was metric invariant across countries and the loadings were fixed according to the results of the invariance analysis (see Appendix D, Box D.1). Further, we fitted all two-level M-ZIP models in *Mplus* 7 (Muthén and Muthén 1998–2012) and we followed recommendations (Rutkowski et al. 2010) to apply sampling weights at the student and class level. The models fitted to the TIMSS 2011 data were specified as follows, with all level-one predictor variables centered about their grand-means:

$$\text{The logistic part: } \text{logit}(p_i) = \alpha_0$$

$$\text{The Poisson part: } \log(\lambda_{ij}) = \beta_{0j} + \sum_{h=1}^5 \beta_{hj} x_{hij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} z_{1j} + U_{0j}$$

$$\beta_{hj} = \gamma_{h0} \text{ for all } h.$$

Here, α_0 is the log odds of no bullying experiences. Then $(\exp(\alpha_0)/(1 + \exp(\alpha_0)))$ expresses the probability of being in the zero category. In the Poisson part of the model, β_{0j} expresses the log counts of bullying victimization for classroom j when all h student-level covariates (x_{hij}) for student i in class j are zero. And the effect of each of the student-level predictors is expressed by β_{hj} , where $h = 1, \dots, 5$. We modelled the classroom average log bullying count (β_{0j}) as a function of an overall system-level expected log bullying count (γ_{00}) when the class-average of InQua (z_{1j}) is zero, an effect for class-average InQua (γ_{01}), plus a random classroom effect (U_{0j}). Between-classroom variance in β_{0j} is expressed as $\text{var}(U_{0j}) = \tau_0^2$. Because we treat all level-one effects as fixed across classrooms, the slopes (β_{hj}) are regarded as fixed and so $\beta_{hj} = \gamma_{h0}$ for all h . It is important to note that, although it is not represented in the above model specification, InQua is a latent variable measured by its relevant indicators at the within- and between-levels. In the pooled model, the international average estimate of log counts of bullying victimization was given as γ_{00} , whereas in the country-specific models, this parameter corresponded to the country-average bullying estimate. After controlling for other covariates, we examined whether instructional quality, as reported by students, was associated with bullying victimization experiences at either the student or classroom level.

Coefficients are interpreted similar to those in standard multilevel regression: statistically significant positive coefficients imply a positive association with counts of violence and statistically significant negative coefficients imply a negative association with counts of violence. To put the Poisson regression coefficients in a more intuitive metric, we can exponentiate them (i.e., $e^{\gamma_{h0} x_{hij}}$) and directly interpret the multiplicative effect of a one unit change in the predictor on the outcome in terms of count ratios.

6.3 Results

We were primarily interested in instructional quality; however, structural parameters for all predictors are also presented (Table 6.2). We do not report the measurement model parameters (factor loadings, intercepts, residual variance, and latent variable variance); however, they are available on request. In terms of the logistic or zero-part of the model, our analysis indicates that, internationally, 21 % of fourth grade students reported no bullying victimization experiences ($\frac{\exp(-1.56)}{1 + \exp(-1.56)} = 0.21$ 95 %CI [0.183, 0.241]).

With respect to the count or Poisson part of the model, we note the following findings associated with instructional quality at the international level. There is a statistically and practically significant negative effect for instructional quality at the student level. That is, better perceived instructional quality is associated with lower rates of student self-reported bullying victimization. In particular, for a one unit increase in instructional quality, bullying rates are expected to be 30 % lower ($e^{-0.36} = 0.698$ (95 %CI [0.541, 0.900])). In other words, for a student who reports instructional quality that is one unit higher, we expect that their reported bullying victimization will be just 70 % that of a student who reported instructional quality one unit lower. In contrast, at the classroom level, we found no relationship between instructional quality and bullying.

Our findings for the pooled international model also indicate a negative sex effect ($\exp(-0.11) = 0.895$; 95 %CI [0.861, 0.932]), with girls reporting bullying rates that are about 10 % lower than their boy peers. Students who reported higher levels of attachment to school reported bullying rates about 6 % lower than their less attached peers ($\exp(-0.06) = 0.942$; 95 %CI [0.906, 0.979]). There was a small negative association between frequency of speaking the language of the test and bullying rates ($\exp(-0.05) = 0.951$; 95 %CI [0.933, 0.970]). Given that students from an immigrant background tend to exhibit stronger associations between bullying and achievement compared to their native born peers (see for example, Rutkowski et al. 2013a, b), this finding is especially germane. Finally, at the international level, we observed a small positive association between our SES proxy and student reports of bullying victimization ($\exp(0.02) = 1.020$; 95 %CI [1.000, 1.040]), where students reporting one unit more books also reported slightly higher (2 %) bullying victimization.

Before discussing the individual educational system results, it must be noted that the language item was not administered in Slovenia, and there is thus no parameter estimate for this variable. Within the individual country analyses (Table 6.2), we generally see a highly heterogeneous pattern for the relationship between instructional quality and self-reported bullying victimization rates after controlling for other covariates in the model. For example, in several educational systems, there was a statistically significant negative association. The strongest negative associations at the within-class level were observed in Turkey ($\exp(-25.98) = 0.000$; 95 %CI [0.000, 0.000]), Tunisia ($\exp(-6.53) = 0.000$; 95 %CI [0.000, 0.046]), Honduras ($\exp(-6.45) = 0.002$; 95 %CI [0.000, 0.088]), Chile ($\exp(-5.20) = 0.006$; 95 %

Table 6.2 M-ZIP results (statistically significant effects for instructional quality are indicated in bold)

Country	Within					Between										
	InQua_w (SE)	t	Girl (SE)	t	Attach (SE)	t	Lang (SE)	t	Books (SE)	t	InQua_b (SE)	t	Int (SE)	t	Zero (SE)	t
International	-0.36 (0.13)	-2.87	-0.11 (0.02)	-7.23	-0.06 (0.02)	-3.36	-0.05 (0.01)	-5.86	0.02 (0.01)	3.13	0.03 (0.08)	0.44	1.71 (0.02)	109.32	-1.56 (0.07)	-23.36
Armenia	0.06 (0.14)	0.41	-0.12 (0.05)	-2.42	-0.02 (0.04)	-0.55	-0.06 (0.05)	-1.20	-0.02 (0.02)	-1.00	-0.45 (0.27)	-1.66	1.41 (0.05)	30.09	0.04 (0.05)	0.78
Austria	0.14 (0.16)	0.86	-0.09 (0.03)	-3.10	-0.25 (0.04)	-7.00	-0.07 (0.03)	-2.88	0.02 (0.01)	1.71	-0.14 (0.21)	-0.68	1.58 (0.03)	63.16	-1.43 (0.06)	-26.07
Azerbaijan	-1.13 (0.40)	-2.83	-0.25 (0.05)	-4.96	-0.08 (0.08)	-1.00	0.03 (0.07)	0.47	-0.02 (0.03)	-0.58	-1.32 (0.94)	-1.40	1.10 (0.08)	13.75	-0.06 (0.09)	-0.64
Bahrain	-3.34 (0.81)	-4.13	-0.22 (0.04)	-5.29	0.12 (0.03)	4.00	0.01 (0.03)	0.33	0.01 (0.01)	0.93	0.09 (0.23)	0.40	1.88 (0.03)	72.35	-2.20 (0.09)	-25.86
Botswana	-0.06 (0.05)	-1.33	-0.03 (0.02)	-1.94	0.00 (0.01)	0.00	0.04 (0.01)	3.00	0.01 (0.01)	0.71	-0.22 (0.17)	-1.35	2.10 (0.01)	149.93	-3.95 (0.07)	-23.66
Chile	-5.20 (1.27)	-4.11	-0.07 (0.03)	-2.41	0.05 (0.05)	1.00	-0.09 (0.03)	-3.40	0.01 (0.01)	0.83	0.28 (0.22)	1.27	1.73 (0.03)	69.04	-2.09 (0.07)	-30.33
Croatia	-0.39 (0.21)	-1.80	-0.14 (0.03)	-4.89	-0.02 (0.03)	-0.70	-0.12 (0.04)	-3.19	-0.02 (0.02)	-1.07	-0.03 (0.18)	-0.16	1.43 (0.07)	20.74	-1.25 (0.06)	-22.36
Czech Republic	-0.04 (0.13)	-0.32	-0.12 (0.04)	-3.54	-0.15 (0.03)	-5.58	-0.12 (0.03)	-3.56	-0.03 (0.01)	-2.00	-0.45 (0.21)	-2.09	1.71 (0.02)	85.55	-1.64 (0.07)	-25.28
Denmark	-0.68 (0.27)	-2.55	-0.07 (0.04)	-1.78	-0.16 (0.04)	-3.95	-0.13 (0.05)	-2.63	0.03 (0.02)	1.87	0.22 (0.19)	1.15	1.34 (0.04)	37.31	-1.49 (0.07)	-21.54
Finland	0.14 (0.09)	1.59	-0.11 (0.03)	-3.38	-0.27 (0.03)	-10.07	-0.17 (0.05)	-3.54	0.01 (0.02)	0.47	0.03 (0.17)	0.14	1.40 (0.02)	63.41	-1.51 (0.06)	-24.70
Georgia	-5.69 (1.30)	-4.38	-0.12 (0.04)	-2.64	0.03 (0.04)	0.97	-0.23 (0.05)	-4.78	-0.03 (0.02)	-1.47	0.84 (1.14)	0.74	1.33 (0.03)	39.21	-0.61 (0.06)	-9.45
Germany	-0.05 (0.11)	-0.47	-0.07 (0.03)	-2.34	-0.19 (0.03)	-7.42	-0.13 (0.03)	-3.74	0.01 (0.01)	0.69	0.02 (0.31)	0.05	1.63 (0.02)	77.52	-1.70 (0.06)	-28.32
Honduras	-6.45 (2.05)	-3.14	-0.15 (0.03)	-4.53	0.05 (0.06)	0.93	0.00 (0.04)	-0.09	0.02 (0.02)	1.35	0.12 (0.44)	0.26	1.76 (0.03)	58.63	-1.82 (0.09)	-19.74

(continued)

Table 6.2 (continued)

Country	Within					Between										
	InQua_w (SE)	t	Girl (SE)	t	Attach (SE)	t	Lang (SE)	t	Books (SE)	t	InQua_b (SE)	t	Int (SE)	t	Zero (SE)	t
Hong Kong	-0.14 (0.05)	-2.58	-0.14 (0.03)	-5.07	-0.06 (0.02)	-3.33	-0.05 (0.03)	-1.80	0.03 (0.01)	2.00	-0.36 (0.23)	-1.53	1.63 (0.02)	85.74	-1.75 (0.06)	-30.22
Hungary	-0.16 (0.07)	-2.35	-0.08 (0.02)	-3.65	-0.10 (0.02)	-4.85	0.01 (0.05)	0.25	-0.03 (0.01)	-2.90	-0.18 (0.23)	-0.79	1.79 (0.02)	99.56	-1.96 (0.06)	-31.66
Iran	-0.28 (0.09)	-2.94	-0.06 (0.08)	-0.79	-0.07 (0.02)	-2.87	0.01 (0.02)	0.33	-0.01 (0.01)	-1.10	-2.58 (0.77)	-3.35	1.66 (0.05)	36.93	-1.72 (0.09)	-18.26
Ireland	-0.30 (0.15)	-2.00	-0.10 (0.04)	-2.18	-0.09 (0.03)	-2.82	-0.16 (0.04)	-3.86	0.02 (0.02)	1.31	0.27 (0.29)	0.91	1.43 (0.04)	32.39	-1.02 (0.06)	-16.98
Italy	-0.12 (0.10)	-1.17	-0.06 (0.03)	-2.06	-0.07 (0.03)	-2.56	-0.12 (0.02)	-5.00	0.03 (0.01)	2.50	-1.14 (0.31)	-3.69	1.61 (0.02)	73.14	-1.56 (0.06)	-25.98
Japan	0.07 (0.12)	0.53	-0.25 (0.03)	-9.26	-0.16 (0.03)	-5.52	-0.10 (0.03)	-3.41	0.02 (0.01)	1.42	-0.14 (0.14)	-1.05	1.66 (0.02)	79.00	-1.49 (0.06)	-24.10
Kazakhstan	2.44 (0.39)	6.32	-0.07 (0.04)	-1.76	-0.36 (0.05)	-7.48	-0.04 (0.04)	-0.97	-0.05 (0.02)	-2.42	-3.39 (0.79)	-4.31	1.44 (0.06)	23.61	-0.65 (0.10)	-6.26
Korea	-0.30 (0.11)	-2.74	-0.22 (0.03)	-8.35	-0.09 (0.03)	-3.37	-0.17 (0.03)	-5.16	0.01 (0.01)	0.50	-0.44 (0.15)	-2.84	1.60 (0.02)	88.94	-1.47 (0.05)	-30.00
Kuwait	-1.53 (0.54)	-2.82	-0.23 (0.06)	-3.80	0.08 (0.03)	2.83	0.02 (0.03)	0.74	0.02 (0.01)	1.50	-0.70 (0.41)	-1.69	1.87 (0.08)	24.58	-1.52 (0.09)	-16.19
Lithuania	0.29 (0.37)	0.79	-0.06 (0.03)	-1.91	-0.28 (0.07)	-4.15	-0.15 (0.04)	-3.40	0.01 (0.02)	0.53	-0.19 (0.53)	-0.36	1.62 (0.02)	73.82	-1.75 (0.06)	-27.78
Malta	-0.11 (0.09)	-1.16	-0.12 (0.03)	-4.17	-0.07 (0.02)	-3.61	0.06 (0.02)	2.64	0.01 (0.01)	0.38	-0.38 (0.22)	-1.76	1.75 (0.02)	91.89	-1.92 (0.07)	-27.87
Morocco	-3.55 (0.91)	-3.92	-0.21 (0.03)	-7.24	0.04 (0.04)	1.03	0.05 (0.02)	2.09	0.06 (0.01)	5.00	0.07 (0.54)	0.14	1.72 (0.04)	41.88	-2.18 (0.14)	-15.83
Netherlands	-0.02 (0.37)	-0.04	-0.07 (0.04)	-2.03	-0.17 (0.06)	-2.84	-0.11 (0.04)	-3.09	0.03 (0.01)	2.43	-0.23 (0.44)	-0.52	1.61 (0.02)	80.70	-2.18 (0.08)	-28.64
New Zealand	-0.05 (0.15)	-0.31	-0.04 (0.02)	-2.00	-0.10 (0.03)	-3.38	-0.07 (0.02)	-3.00	0.01 (0.01)	1.30	0.39 (0.28)	1.37	1.90 (0.02)	112.00	-2.36 (0.07)	-34.71

(continued)

Table 6.2 (continued)

Country	Within					Between										
	InQua_w (SE)	t	Girl (SE)	t	Attach (SE)	t	Lang (SE)	t	Books (SE)	t	InQua_b (SE)	t	Int (SE)	t	Zero (SE)	t
Norway	0.04 (0.21)	0.17	-0.11 (0.04)	-3.00	-0.22 (0.04)	-5.24	-0.12 (0.04)	-2.75	0.01 (0.02)	0.59	-0.22 (0.26)	-0.85	1.53 (0.03)	58.81	-1.77 (0.09)	-20.84
Oman	-0.21 (0.08)	-2.81	-0.10 (0.02)	-5.30	-0.02 (0.02)	-1.10	-0.01 (0.01)	-0.73	0.02 (0.01)	3.52	-1.03 (0.25)	-4.08	1.90 (0.02)	109.59	-2.11 (0.08)	-25.89
Poland	-0.33 (0.16)	-2.13	-0.24 (0.03)	-7.03	-0.01 (0.03)	-0.26	-0.20 (0.05)	-4.28	0.01 (0.02)	0.63	0.00 (0.49)	0.00	1.53 (0.02)	63.71	-0.99 (0.09)	-23.57
Portugal	-0.38 (0.20)	-1.88	-0.17 (0.04)	-4.25	-0.08 (0.05)	-1.50	-0.05 (0.05)	-1.07	-0.01 (0.02)	-0.63	-0.76 (0.54)	-1.42	1.59 (0.03)	54.86	-1.81 (0.08)	-22.90
Qatar	-0.07 (0.10)	-0.70	-0.20 (0.04)	-4.70	-0.03 (0.02)	-1.39	-0.01 (0.02)	-0.67	0.02 (0.01)	1.25	-0.94 (0.43)	-2.17	1.98 (0.02)	82.50	-2.10 (0.09)	-23.57
Romania	-0.28 (0.13)	-2.10	-0.23 (0.03)	-7.70	0.01 (0.03)	0.15	-0.05 (0.05)	-1.00	-0.02 (0.01)	-1.21	-1.28 (0.69)	-1.86	1.59 (0.05)	33.08	-1.73 (0.09)	-18.96
Russia	-0.32 (0.09)	-3.64	-0.06 (0.02)	-2.38	-0.07 (0.03)	-2.68	-0.07 (0.03)	-2.46	0.00 (0.01)	0.08	-0.59 (0.33)	-1.79	1.70 (0.02)	73.87	-1.61 (0.07)	-22.72
Saudi Arabia	-3.80 (1.07)	-3.56	-0.38 (0.09)	-4.42	0.06 (0.03)	2.13	0.02 (0.04)	0.51	0.00 (0.01)	0.15	-0.38 (0.47)	-0.83	1.70 (0.04)	47.11	-2.14 (0.13)	-16.24
Serbia	-0.32 (0.24)	-1.34	-0.23 (0.04)	-6.60	-0.01 (0.05)	-0.20	-0.17 (0.04)	-4.10	0.02 (0.02)	1.11	-0.21 (0.30)	-0.70	1.54 (0.03)	46.55	-1.05 (0.07)	-15.49
Singapore	-0.11 (0.04)	-2.59	-0.21 (0.02)	-10.19	-0.08 (0.02)	-5.33	-0.03 (0.02)	-1.39	0.03 (0.01)	3.00	-0.16 (0.14)	-1.15	1.75 (0.01)	125.14	-2.16 (0.06)	-38.59
Slovakia	-0.02 (0.16)	-0.14	-0.11 (0.03)	-4.24	-0.13 (0.03)	-4.64	-0.10 (0.03)	-3.57	-0.01 (0.01)	-0.92	0.16 (0.29)	0.56	1.66 (0.03)	55.33	-1.69 (0.06)	-27.26
Slovenia	-0.38 (0.69)	-0.55	-0.19 (0.07)	-2.86	-0.05 (0.10)	-0.51	.	.	-0.01 (0.02)	-0.69	-0.55 (0.45)	-1.22	1.62 (0.03)	52.86	-1.76 (0.08)	-21.96
Spain	-0.13 (0.13)	-0.76	-0.07 (0.03)	-2.16	-0.06 (0.03)	-1.97	-0.07 (0.02)	-3.36	0.02 (0.01)	1.33	-0.05 (0.15)	-0.34	1.80 (0.02)	85.67	-1.56 (0.06)	-25.49
Sweden	-0.33 (0.13)	-2.46	-0.04 (0.04)	-1.06	-0.21 (0.03)	-6.90	-0.19 (0.04)	-5.40	0.04 (0.02)	2.50	0.19 (0.26)	0.71	1.26 (0.03)	46.63	-1.05 (0.06)	-17.75

(continued)

Table 6.2 (continued)

Country	Within						Between									
	InQua_w (SE)	t	Girl (SE)	t	Attach (SE)	t	Lang (SE)	t	Books (SE)	t	InQua_b (SE)	t	Int (SE)	t	Zero (SE)	t
Taiwan	-0.10 (0.06)	-1.70	-0.11 (0.03)	-3.93	-0.07 (0.02)	-3.04	-0.07 (0.03)	-2.46	0.00 (0.01)	-0.17	-0.26 (0.19)	-1.36	1.66 (0.02)	83.20	-1.29 (0.06)	-22.24
Thailand	-0.01 (0.10)	-0.12	-0.06 (0.02)	-3.37	-0.04 (0.02)	-2.21	-0.01 (0.02)	-0.44	0.02 (0.01)	2.20	-0.13 (0.25)	-0.50	2.09 (0.02)	87.25	-3.17 (0.14)	-23.29
Tunisia	-6.53 (1.71)	-3.82	-0.21 (0.04)	-5.58	-0.10 (0.04)	-2.49	0.03 (0.02)	1.41	0.01 (0.02)	0.71	-0.53 (0.36)	-1.46	1.62 (0.03)	62.35	-2.22 (0.11)	-20.14
Turkey	25.98 (11.46)	2.27	-0.15 (0.02)	-7.00	-0.09 (0.02)	-4.89	-0.07 (0.02)	-3.22	0.00 (0.01)	0.00	-1.18 (0.25)	-4.80	1.79 (0.02)	99.56	-1.84 (0.06)	-28.80
UAE	-0.06 (0.06)	-1.02	-0.17 (0.02)	-7.13	-0.03 (0.01)	-2.29	-0.01 (0.01)	-0.65	0.01 (0.01)	1.09	-0.51 (0.18)	-2.85	1.89 (0.02)	128.31	-1.99 (0.05)	-38.39
USA	-0.38 (0.35)	-1.10	-0.06 (0.02)	-2.77	-0.08 (0.06)	-1.34	-0.03 (0.02)	-1.42	0.03 (0.01)	2.70	-0.08 (0.24)	-0.31	1.71 (0.02)	100.35	-1.35 (0.04)	-37.42
Yemen	-0.15 (0.39)	-0.38	-0.17 (0.06)	-2.81	-0.02 (0.06)	-0.36	-0.07 (0.03)	-2.15	0.04 (0.02)	2.53	-0.97 (0.42)	-2.30	1.73 (0.07)	24.34	-1.36 (0.14)	-10.03

Note: SE standard error; t t-value (statistically significant when $t > 1.96$ or $t < -1.96$); *InQua_w* within-class instructional quality; *Attach* student attachment to the school; *Lang* frequency of speaking the language of the test at home; *Books* number of books in the home; *InQua_b* between-class instructional quality; *Int* intercept; *Zero* intercept for no bullying. UAE United Arab Emirates

CI [0.000, 0.066]), and Georgia ($\exp(-5.69) = 0.003$; 95 %CI [0.000, 0.043]). In contrast, Kazakhstan exhibited a positive association between perceived instructional quality and bullying victimization at the student level ($\exp(2.44) = 11.47$; 95 % CI [5.34, 24.64]). These findings indicate that, as students report better instructional quality practices, they also tend to report much higher incidences of bullying. The remainder of the educational systems, 25 of the 48, showed no association at the student level between instructional quality and bullying rates.

At the class level, findings were also highly varied across educational systems. Whereas Yemen showed no association at the student-level, there was a strong negative association at the class-level; we found that class-average bullying reports markedly reduced with improved instructional quality ($\exp(-0.97) = 0.379$; 95 % CI [0.17, 0.86]). Findings were similar in several countries, including the Czech Republic, Qatar, and the United Arab Emirates, who exhibited no relationship at the student level, but a negative association at the classroom-level. A small group of educational systems demonstrated both a negative student- and class-level association between instructional quality and bullying victimization. At the between-classroom level, these systems include Iran ($\exp(-2.58) = 0.076$; 95 % CI [0.017, 0.342]), Korea ($\exp(-0.44) = 0.644$; 95 %CI [0.480, 0.864]), and Oman ($\exp(-1.03) = 0.357$; 95 %CI [0.219, 0.582]). The remaining 39 educational systems demonstrated no association at the classroom-level between instructional quality and bullying victimization. In summary, at the country level, we observed significant negative relations at both the student and class level; however more educational systems (23) exhibited significant negative relations at the student level than at the classroom level (10).

6.4 Discussion

In this study, we aimed to determine the degree to which instructional quality is associated with bullying victimization at both the international level and within TIMSS-participating educational systems. Although bullying levels vary substantially in each participating system, unfortunately, in all systems, there is a prevalence of self-reported bullying victimization at the fourth grade. This is in line with previous similar research at the eighth grade (Rutkowski et al. 2013a, b). Fortunately, in many countries, the rates of self-reported bullying victimization are relatively low, leading to many students reporting no bullying victimization. We accounted for this skewed distribution in our analysis via a zero-inflated Poisson modeling approach. Further, to help isolate the relationship between instructional quality and bullying, we controlled for several covariates that have been demonstrated to predict bullying at the international level including: sex of the student, student attachment to the school, socio-cultural capital of the student, and immigrant background of the student, as measured by language spoken at home. Finally, given the wide variety of cultural differences among TIMSS countries, our analysis allowed for a unique cross-national examination. We believe that these findings have important implications for future research

and point to a need for a better understanding of the teacher's role in bullying prevention, both within and between countries.

Internationally, there is a negative association between student reported instructional quality and bullying rates. This association, however, is limited to the student level. In other words, better perceived instructional quality is related to fewer reports of bullying victimization within classes. These results suggest that, after accounting for other covariates, the sampled students' perceptions of teachers and their teaching quality play a role in explaining differences at the student level in bullying victimization internationally. These results are particularly interesting when we consider the economic, geographic, cultural, and linguistic diversity of TIMSS participant systems. Stated simply, we have some evidence that, at the international level, students' perceptions of instructional quality correlate with outcomes beyond achievement. In contrast, at the classroom level, we found that better instructional quality (as reported by the students, aggregated to the classroom level) was not associated with bullying rates internationally.

At the country level, we found a highly heterogeneous pattern of instructional quality results, with a mix of primarily negative or null associations at both analyzed levels. This is in contrast to other covariates in the model. For example, student sex explained differences in bullying experiences in most analyzed educational systems. Specifically, bullying rates were generally lower for girls in all but six countries (Lithuania, Kazakhstan, Denmark, Botswana, Sweden, and Iran). In addition, student attachment to the school was predictive of lower bullying rates in 28 out of 48 educational systems. These findings are consistent with previous research that analyzed similar outcomes at the eighth grade (Rutkowski et al. 2013a, b) and found that girls and more attached students tended to report fewer incidences of bullying victimization. When we consider previous research that has pointed to clear differences in the kind of bullying victimization experienced by boys and girls (Hong and Espelage 2012; Smith et al. 1999), interventions that are sex-specific might be reasonable in many of the countries considered. Specific to the findings around attachment, the causal direction of this relationship cannot be established through our results, unfortunately. Some studies suggest that students who are frequently victimized feel a lower sense of belonging (Eisenberg et al. 2009), while other studies reported less bullying at schools where students report higher school attachment (Hong and Espelage 2012; Richard et al. 2012), suggesting that decreases in victimization are associated with better school attachment. Nonetheless, the consistency of this finding indicates that school attachment is important and that further research in this area is justified. Policies for fostering school attachment should approach the issue holistically, considering both the academic and social aspects of a student's sense of belonging (Akiba 2010).

Teachers should be integral to any policy, as students spend most of their days in the presence of teachers, placing teachers in the most obvious position to directly foster a safe environment for students, with the support of school leadership. Our findings suggest that student perceptions of instructional quality have the potential to reduce bullying victimization; however, given the mixed findings at the educational

system level and at the classroom level, it is unclear whether an emphasis on improving instructional quality would translate into better outcomes. Rather, it is likely that other interventions and policies would be important in this regard.

6.5 Limitations

We note several limitations to our study. First, our bullying measure includes both physical and verbal victimization and social exclusion, and does not identify the perpetrator. Our data do not indicate whether bullying victimization took place within or outside of the classroom. The stem of the question only situates the occurrence at school. Furthermore, the data available for measuring instructional quality did not include items measuring classroom management or cognitive activation. Our construct is hence not as broad as that commonly used in the literature (see for example Klieme et al. 2009). Had the TIMSS data included such items, we may have obtained different significant findings. TIMSS data is cross-sectional and observational, making any conclusions correlational only. To that end, further research in each of the identified areas is important for establishing the causal direction of the relationships. In particular, it might be that bullied students hold generally negative feelings toward schooling and their teachers, thus explaining the pattern of negative associations at the student level between perceptions of teaching quality and bullying. Finally, we note that given the state-of-the-art in commercially available software, we made a trade-off between fully capturing the nested structure of the data (students in classes in schools in countries) and most appropriately modeling the distribution of bullying victimization. Despite these limitations, we found meaningful associations between instructional quality and bullying victimization internationally, pointing to a need for country- and school-level policies and interventions to foster safe and productive learning environments for all students.

6.6 Conclusions

Our analysis shows that bullying begins at an early age and that, at the fourth grade level, bullying victimization is an international phenomenon. Although we found that positive student perceptions of instructional quality were associated with lower reported bullying victimization rates internationally, cross-system differences indicate that instructional quality is unlikely to be a universal solution. The lack of a homogeneous solution that can be applied internationally speaks to the complexity of bullying victimization and how, in spite of a global prevalence and near-universal consequences (Akiba 2010; Engel et al. 2009; Rutkowski et al. 2013a, b), the problem of bullying in schools necessitates local solutions. Hence, educational system-level policy makers must address the issue by carefully examining their own context and by using tools that are proven to work best in a given setting. To that

end, it is important to recognize that this study is one piece of evidence in the international bullying literature, and further research, especially at the system-level, is clearly needed in terms of identifying interventions and policies that foster a safe secure learning environment for the youngest students. Nevertheless, the power of such an analysis, with many countries and representative samples, demonstrates that bullying victimization is happening across a wide range of heterogeneous countries, regardless of geography, dominant race/ethnicity, language, culture, and economic development.

References

- Agresti, A. (2002). *Categorical data analysis*. Chichester: John Wiley & Sons.
- Akiba, M. (2010). What predicts fear of school violence among U.S. adolescents? *Teachers College Record*, *112*(1), 68–102.
- Allen, K. P. (2010). Classroom management, bullying, and teacher practices. *Professional Educator* *34*(1). Retrieved from <http://eric.ed.gov/?id=EJ988197>.
- Barbetta, P. M., Norona, K. L., & Bicard, D. F. (2005). Classroom behavior management: A dozen common mistakes and what to do instead. *Preventing School Failure: Alternative Education for Children and Youth*, *49*(3), 11–19. doi:10.3200/PSFL.49.3.11-19.
- Bosworth, K., Espelage, D. L., & Simon, T. R. (1999). Factors associated with bullying behavior in middle school students. *The Journal of Early Adolescence*, *19*(3), 341–362.
- Bullock, J. R. (2002). Bullying among children. *Childhood Education*, *78*(3), 130–133.
- CDC. (n.d.). *About school violence*. Atlanta, GA: Centers for Disease Control and Prevention. Retrieved from <http://www.cdc.gov/violenceprevention/youthviolence/schoolviolence/>.
- Cook, C. R., Williams, K. R., Guerra, N. G., & Kim, T. E. (2010). Variability in the prevalence of bullying and victimization: A cross-national and methodological analysis. In S. R. Jimerson, S. M. Swearer, & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective*. (pp. 347–362). New York, NY: Routledge. Retrieved from <http://psycnet.apa.org/psycinfo/2010-06797-000>.
- Crothers, L. M., & Kolbert, J. B. (2010). *Teachers' management of student bullying in the classroom* (pp. 535–546). *Handbook of Bullying in Schools: An International Perspective*.
- Dill, E. J., Vernberg, E. M., Fonagy, P., Twemlow, S. W., & Gamm, B. K. (2004). Negative affect in victimized children: The roles of social withdrawal, peer rejection, and attitudes toward bullying. *Journal of Abnormal Child Psychology*, *32*(2), 159–173.
- Eisenberg, M. E., Neumark-Sztainer, D., & Perry, C. L. (2009). Peer harassment, school connectedness, and academic achievement. *Journal of School Health*, *73*(8), 311–316.

- Engel, L. C., Rutkowski, D., & Rutkowski, L. (2009). The harsher side of globalisation: violent conflict and academic achievement. *Globalisation, Societies and Education*, 7(4), 433–456. doi:10.1080/14767720903412242.
- Foy, P., Arora, A., & Stanco, G. A. (Eds.) (2013). TIMSS 2011 user guide for the international database. Supplement 1: International version of the TIMSS 2011 background and curriculum questionnaires. Chestnut Hill, MA: TIMSS & PIRLS International Study Center and International Association for the Evaluation of Educational Achievement.
- Frymier, A. B., & Houser, M. L. (2000). The teacher-student relationship as an interpersonal relationship. *Communication Education*, 49(3), 207–219. doi:10.1080/03634520009379209.
- Hall, D. B. (2000). Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4), 1030–1039. doi:10.1111/j.0006-341X.2000.01030.x.
- Harel-Fisch, Y., Walsh, S. D., Fogel-Grinvald, H., Amitai, G., Pickett, W., Molcho, M., et al. (2011). Negative school perceptions and involvement in school bullying: A universal relationship across 40 countries. *Journal of Adolescence*, 34(4), 639–652.
- Haynie, D. L., Nansel, T., Eitel, P., Crump, A. D., Saylor, K., Yu, K., & Simons-Morton, B. (2001). Bullies, victims, and bully/victims: Distinct groups of at-risk youth. *The Journal of Early Adolescence*, 21(1), 29–49.
- Hong, J. S., & Espelage, D. L. (2012). A review of research on bullying and peer victimization in school: An ecological systems analysis. *Aggression and Violent Behavior*, 17(4), 311–322.
- Jimerson, S. R., Swearer, S. M., & Espelage, D. L. (2010). *Handbook of bullying in schools: An international perspective*. New York, NY: Routledge. Retrieved from <http://psycnet.apa.org/psycinfo/2010-06797-000>.
- Kyriakides, L., Creemers, B. P. M., Papastilianou, D., & Papadatou-Pastou, M. (2014). Improving the school learning environment to reduce bullying: An experimental study. *Scandinavian Journal of Educational Research*, 58(4), 453–478. doi:10.1080/00313831.2013.773556.
- Klieme, E., Pauli, C., & Reusser, K. (2009). The pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). New York, NY: Waxmann Publishing Co.
- Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1), 1–14. doi:10.2307/1269547.
- Muthén, L., & Muthén, B. O. (1998–2012). *Mplus user's guide*. (Seventh edition). Los Angeles, CA: Muthén & Muthén.
- Nansel, T., Craig, W., Overpeck, M., Saluja, G., & Ruan, W. (2003). Cross-national consistency in the relationship between bullying behaviors and psychosocial adjustment. *Chives of Pediatric and Adolescent Medicine*, 158, 730–736.
- Natvig, G. K., Albrektsen, G., & Qvarnström, U. (2001). School-related stress experience as a risk factor for bullying behavior. *Journal of Youth and Adolescence*, 30(5), 561–575. doi:10.1023/A:1010448604838.
- Nicolaidis, S., Toda, Y., & Smith, P. K. (2002). Knowledge and attitudes about school bullying in trainee teachers. *British Journal of Educational Psychology*, 72, 105.
- OECD. (2005). *Teachers matter*. Paris: Author. Retrieved from <http://www.oecd-ilibrary.org/content/book/9789264018044-en>.
- Olweus, D. (1994). Bullying at school: Basic facts and effects of a school based intervention program. *Journal of Child Psychology and Psychiatry*, 35(7), 1171–1190. doi:10.1111/j.1469-7610.1994.tb01229.x.
- Olweus, D. (2007). *Olweus sample school report (Sample)*. Center City, MN: Hazelden.
- Olweus, D. (2010). Understanding and researching bullying: Some critical issues. In S. R. Jimerson, S. M. Swearer, & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective*. (pp. 9–34). New York, NY: Routledge. Retrieved from <http://psycnet.apa.org/psycinfo/2010-06797-000>.
- Reinke, W. M., & Herman, K. C. (2002). Creating school environments that deter antisocial behaviors in youth. *Psychology in the Schools*, 39(5), 549–559. doi:10.1002/pits.10048.

- Richard, J. F., Schneider, B. H., & Mallet, P. (2012). Revisiting the whole-school approach to bullying: Really looking at the whole school. *School Psychology International, 33*(3), 263–284.
- Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151.
- Rutkowski, D. J., Rutkowski, L., & Wild, J. (2013a). Predictors of school violence internationally: The importance of immigrant status and other factors. Paper presented at the 5th IEA International Research Conference, Singapore. Retrieved from <http://www.iea.nl/irc-2013.html>.
- Rutkowski, L., Rutkowski, D., & Engel, L. (2013b). Sharp contrasts at the boundaries: School violence and educational outcomes internationally. *Comparative Education Review 57*(2), 232–259. doi:10.1086/669120.
- Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571–581. doi:10.1037/0022-0663.85.4.571.
- Smith, P. K. (2010). Bullying in primary and secondary schools: Psychological and organizational comparisons. In S. Jimerson, S. Swearer, & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 137–150). New York, NY: Routledge.
- Smith, P. K., Morita, Y. E., Junger-Tas, J. E., Olweus, D. E., Catalano, R. F., & Slee, P. E. (1999). *The nature of school bullying: A cross-national perspective*. New York, NY: Routledge.
- Smokowski, P. R., & Kopasz, K. H. (2005). Bullying in school: An overview of types, effects, family characteristics, and intervention strategies. *Children & Schools, 27*(2), 101–110.
- UN News. (2015). UN envoy calls for efforts to eliminate bullying. UN News 10 June 2015. Retrieved from <http://www.un.org/sustainabledevelopment/blog/2015/10/un-envoy-calls-for-concerted-efforts-to-eliminate-bullying-in-all-regions/>.
- Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2015). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*. Retrieved from <http://doi.org/10.1037/edu0000075>.
- Wang, J., Iannotti, R. J., & Nansel, T. R. (2009). School bullying among adolescents in the United States: Physical, verbal, relational, and cyber. *Journal of Adolescent Health, 45*(4), 368–375. doi:10.1016/j.jadohealth.2009.03.021.
- Yoon, J., & Kerber, K. (2003). Bullying: Elementary teachers' attitudes and intervention strategies. *Research in Education, 69*(1), 27–35. doi:10.7227/RIE.69.3.

Final Remarks

Jan-Eric Gustafsson and Trude Nilsen

Abstract This book contributes to educational policy, the field of educational effectiveness and practice. In this chapter, the findings from the five studies are summarized and discussed. After a comprehensive examination of the methodological issues related to measurement, causality, analysis, and design, implications for educational practice are proposed.

Keywords Discussion • Instructional quality • Teacher quality • Methodological issues • Contribution

7.1 Overview of the Five Studies

We first begin with a brief synopsis of the findings by chapter (Table 7.1). In summarizing the contributions of individual chapters, we employ the conceptual framework described in Chap. 1 (Fig. 1.1), which is based on the dynamic model of educational effectiveness (Creemers and Kyriakides 2008). This framework outlines the relations between the educational levels, ranging from the national level, via school and class levels to the student level.

As TIMSS 2011 does not explicitly provide detailed information on educational systems at a national level, this level is only implicitly included in the analyses. However, it is evident that the national level has influenced the findings, in that

J.-E. Gustafsson (✉)

Department of Education and Special Education, University of Gothenburg,
Gothenburg, Sweden
e-mail: jan-eric.gustafsson@ped.gu.se

J.-E. Gustafsson

Faculty of Educational Sciences, Centre of Educational Measurement
at the University of Oslo (CEMO), Oslo, Norway

T. Nilsen

Department of Teacher Education and School Research, University of Oslo, Oslo, Norway
e-mail: trude.nilsen@ils.uio.no

Table 7.1 A summary of the objective and the findings of each chapter

Chapter	Objective	Results
2	Investigate the relations between instructional quality, teacher quality and achievement in mathematics	Findings from the international model indicated that professional development and preparedness had, on average, the strongest relations with instructional quality and achievement. Teachers' experience influenced instructional quality and students' mathematics achievement. The teachers' attained level and major in math or math education did not matter for instructional quality, but were significantly related to mathematics achievement. Achievement was not influenced by instructional quality. At the educational-system level, findings were mixed, although professional development and preparedness had significant relations to instructional quality and student achievement in a large number of countries
3	Investigate the relations between school climate, instructional quality, and student motivation in mathematics	There was a significant positive relation between instructional quality and achievement motivation in all countries. In a number of countries, instructional quality partially mediated the relation between school climate and achievement motivation. Mediation was most apparent for an orderly school climate, and then for school emphasis on academic success. A safe climate was a mediator in only seven educational systems
4	Investigate the effects of school climate and teacher quality on mathematics achievement using country-level longitudinal analyses	Teachers' attained level of education was found to be quite strongly related to educational achievement. There were also quite substantial relations between student achievement and professional development. Teacher self-efficacy, as assessed by self-reports of preparedness for teaching in different domains, was weakly positively, but insignificantly related to student achievement. The teacher characteristics, years of teaching experience and the major academic discipline studied, had no effect. School emphasis on academic success did not satisfy ideals of unidimensionality, and only items reflecting parental support for student achievement and students' desire to perform well were significantly related to student achievement

(continued)



Table 7.1 (continued)

Chapter	Objective	Results
5	Investigate how instructional quality influences the relation between reading and mathematics achievement	All educational systems revealed a strong positive correlation between reading comprehension and mathematics achievement. Further, a number of countries demonstrated a positive relation between instructional quality and mathematics achievement and between instructional quality and reading achievement. The analysis of the moderation of the relationship between mathematics and reading by instructional quality was inconclusive; moderation was present in only six countries
6	Determine the degree to which instructional quality serves as a protective factor against school bullying victimization	Findings from the international model indicated that higher instructional quality was associated with lower rates of student self-reported bullying victimization. At the educational-system level, findings were mixed. In all systems there was a prevalence of self-reported bullying victimization at the fourth grade. However, girls and students who were more attached to their school tended to report fewer incidences of bullying victimization

most analyses yielded heterogeneous results across educational systems (from here on referred to as countries). Moreover, there was some evidence that countries with similar cultures and educational policies, such as the Nordic or Arabic countries, had similar patterns of results.

At the school level, two chapters examined different aspects of school climate. In line with the conceptual model, the results showed that school climate influenced both teachers’ instructional quality and students’ educational outcomes. At the classroom level, the results indicated that aspects of teacher quality were associated with instructional quality and student achievement. Instructional quality, as rated by students, had a positive relation with motivation and, in many cases, with achievement. These results are in general agreement with the conceptual model. However, the data revealed huge variations in the strengths of the different relations, something that probably is related to the specific constructs chosen, the indicators selected, and the countries examined. This result emphasizes that great care needs to be applied before generalizations can be made.

The student level included variables describing student background; although not the focus of the current research, these were used as control variables. As expected, SES, gender, and migration background were related to student outcomes

in many countries, however, the strength and the direction of relations varied. For example, it is not a given that girls are worse at mathematics than boys.

The conceptual model included both cognitive and affective outcomes. The cognitive outcomes (mostly achievement in mathematics, but also in reading) were related to variables at the teacher level (such as instructional quality as rated by students, and aspects of teacher quality) and at the school level (namely aspects of school climate). The affective outcomes (including bullying and motivation) were also related to instructional quality; motivation was also related to school climate.

7.2 Discussion of Substantive Issues

In the following discussion, we follow the structure of our conceptual framework (Chap. 1, Fig. 1.1). We start by discussing the school level, and then proceed with the teacher level and then the student level. Discussions related to the national level are included where required.

Two chapters examined reported school climate within the TIMSS 2011 grade eight data: Scherer and Nilsen (Chap. 3) investigated safe and orderly school climate and school emphasis on academic success (SEAS), while Gustafsson and Nilsen (Chap. 4) examined the role of SEAS in their country-level longitudinal design.

Scherer and Nilsen (Chap. 3) found that the three aspects of school climate were positively related to both perceived instructional quality and motivation in a number of countries. Instructional quality was positively related to motivation, and mediated the influences of an orderly climate in about half of the countries, and the influence of SEAS in about 30 % of the countries.

These results are interesting, given that few studies have established the relations between school climate, instructional quality, and student motivation (Thapa et al. 2013; Wang and Degol 2015). The findings were heterogeneous across countries, but indicated patterns for countries with similar educational systems, cultures and educational policies. For instance, the influence of an orderly climate on motivation was mediated by instructional quality in all English-speaking countries.

Gustafsson and Nilsen (Chap. 4) included the 38 countries that participated in both the 2007 and 2011 cycles of TIMSS, in a longitudinal approach. They found that only the part of SEAS that reflects parental support and students' desire to do well had a positive influence on achievement. SEAS reflects teachers, parents' and students' priorities and ambitions for academic success. The results from this chapter emphasize the importance of the parents' and students' contributions to the academic school climate, and hence extends existing research, which has, for the most part, focused on school leaders and teachers (see for example, Wang and Degol 2015).

Although both these studies accessed different outcome and school climate variables (Scherer and Nilsen focused on students' motivation and a broad range of school climate, whereas Gustafsson and Nilsen focused on student achievement and

SEAS only), their findings both confirm a positive relation between good school climate and educational outcomes in mathematics. This confirms the expectations of our conceptual model and previous reviews (see Thapa et al. 2013; Wang and Degol 2015).

At the class level, teacher quality and instructional quality were the two main constructs. Instructional quality was included in every study (except for Chap. 4, where data was not available), and the studies reported in Chap. 3 (Scherer and Nilsen), Chap. 5 (Nortvedt et al.), and Chap. 6 (Rutkowski and Rutkowski) investigated instructional quality as rated by students, and aggregated these ratings to the classroom level.

Scherer and Nilsen (Chap. 3) found that, for grade 8 students, instructional quality was positively related to all three motivational constructs (intrinsic and extrinsic motivation and self-concept) in 48 of the 50 countries. This finding is a major extension of previous research. Our literature review indicated that most previous studies conducted on relations between instructional quality and motivation were single country studies conducted primarily in Germany or the USA (see for example, Covington 2000; Fauth et al. 2014; Kunter et al. 2013; Stroet et al. 2013; Wang and Eccles 2013). Given that the sample includes countries from all continents, with diverse cultures and educational policies, the findings in Chap. 3 emphasize the need for teachers to support their students by engaging them, and providing clear and comprehensive instruction, in order to promote students intrinsic and extrinsic motivation and self-concept in mathematics.

Nortvedt et al. (Chap. 5) found perceived instructional quality to be positively related to mathematics and reading achievement in 40 % of the countries. Because their construct measured the aspects of instructional quality related to a supportive climate and clarity, this finding is important for two reasons. First, the aspects found to be strongly related to achievement in the existing research were, first and foremost, cognitive activation and classroom management (see Kunter et al. 2013); the findings of Nortvedt et al. emphasize the additional importance of these two other aspects of instructional quality. Second, the bulk of previous research on instructional quality has been conducted in Germany or the USA, and very seldom in developing countries; these findings extend this research to a more international level. Taking these two considerations together, Nortvedt et al.'s study indicates that the cultural context and the educational system play key roles in the aspects of instructional quality that are important for student achievement. For instance, aspects that are important for student achievement in Germany may differ from those important in Oman. In general though, the findings of Nordtvedt et al. are in agreement with existing research and support the idea that quality of instruction matters for student achievement (Baumert et al. 2010; Klieme et al. 2009; Good et al. 2009; Pianta and Hamre 2009; Scherer and Gustafsson 2015; Wayne and Youngs 2003).

Rutkowski and Rutkowski (Chap. 6) identified instructional quality to be negatively related to bullying internationally, in other words, higher instructional quality was associated with reduced levels of bullying. They observed significant negative relations at both the student and class level; however more educational

systems (23) exhibited significant negative relations at the student level than at the classroom level (10). At the student level, the measure refers to students' individual perceptions of instructional quality. Hence, more significant findings at the student level could reflect students' overall perceptions and attitudes to schooling. Students who are not bullied may tend to be more positive in their ratings.

Very few studies have investigated relations between bullying and instructional quality, although Kyriakides et al. (2014) examined this in five countries (Belgium, Cyprus, England, Greece, and The Netherlands) and found that the risk of bullying is reduced when students are provided with opportunities to learn and high quality teaching. Rutkowski and Rutkowski extend the limited research there is in the field by including 48 countries with diverse cultures, and by emphasizing the need for supportive teachers with clear instruction to reduce bullying.

Blömeke et al. (Chap. 2) found that instructional quality as rated by teachers was significantly related to teacher quality, and teacher quality was also related to student achievement. While previous studies have addressed sub-questions of this general relation (for example, Baumert et al. 2010; Blömeke and Delaney 2012; Goe 2007; Wayne and Youngs 2003), this is the first time a comprehensive model has been applied to almost 50 countries using a broad set of specific indicators. Specifically, in a large number of countries, Blömeke et al. identified positive relations between teachers' experience, attained level of education, and major in math or math education, and students' mathematics achievement, with countries from the same region revealing similar relational patterns. In contrast to their hypothesis, Blömeke et al. did not find significant relations between instructional quality, as rated by teachers, and achievement.

Gustafsson and Nilsen (Chap. 4) found that teachers' attained level of education and professional development had positive effects on mathematics achievement in grade eight. No study has investigated teacher quality with a longitudinal approach and with so many countries (38) before, thus these findings are important, and also support previous research (see Timperley et al. 2007). In contrast to Chap. 2, the study in Chap. 4 found that fewer aspects of teacher quality had a significant influence on achievement. Whether this was because the study in Chap. 2 investigated grade four students and Chap. 4 investigated grade eight students, or because Chap. 4 had a longitudinal approach while Chap. 2 pursued a cross-sectional approach, is difficult to disentangle, and calls for further research.

Student characteristics were related to both cognitive and affective outcomes. Blömeke et al. (Chap. 2) demonstrated a strong relation between SES and achievement, as did the other chapters and previous research (Hansen and Munk 2012). Gender differences in mathematics achievement were found in 28 countries, particularly in Europe and Latin America, unanimously in favor of boys. In Western Asian/Arabian and African countries, gender inequality was less prevalent, and, when present, the differences favored girls. This pattern may be distinguished in the international TIMSS reports (see for example, Mullis et al. 2012), although these reports present descriptive statistics; the gender patterns identified by Blömeke et al. thus extend previous research.

Rutkowski and Rutkowski (Chap. 6) found that bullying rates were generally lower for girls and for students who spoke the language of the test. Moreover, student attachment to the school was predictive of lower bullying rates in 28 out of 48 educational systems. These findings contribute in a major way to policy and to the field of bullying; the large number of countries included in Rutkowski and Rutkowski's analysis considerably extend previous research in this area (Rutkowski et al. 2013).

7.3 Methodological Issues

The complex structure of the large-scale data sets used in this report gave rise to some methodological issues, which point to the need for further research and development.

7.3.1 *Measurement*

The constructs investigated here are complex and challenging to measure. However, the existence of relations between different aspects of teacher quality on the one hand and student achievement on the other hand suggests that measurement of these variables has been reasonably successful. In contrast, the relative paucity and inconsistency of relations between measures of instructional quality and other variables indicate problems with the ways in which this construct has been operationalized in TIMSS.

There is long-standing controversy over whether teacher or student ratings should be used to assess instructional quality (Desimone et al. 2009; Marsh et al. 2012; Scherer et al. 2016; Schlesinger and Jentsch 2016; Wagner et al. 2015). Our results do not resolve the controversy given that they indicate that there are problems with both approaches. However, as was noted by Blömeke et al. (Chap. 2), teacher and student ratings may capture different aspects of instructional quality, which implies that they will not be highly correlated and that both may be needed for the construct of instructional quality to be adequately captured.

Instructional quality as reported by teachers or students may be affected by response-style bias caused by, for instance, cultural differences (Wagner et al. 2015). In Chap. 2, Blömeke et al. discussed further potential problems with the measurement of instructional quality by teacher reports.

Instructional quality as reported by students included the important aspects of teachers' clarity and support in learning. However, it is also desirable to capture the other two central aspects of instructional quality, namely cognitive activation and classroom management (Fauth et al. 2014; Klieme et al. 2009). TIMSS includes items in the teacher questionnaire that measure an orderly atmosphere in the school, but this refers to the school level and not to the classroom level. Including more

items and also capturing these two additional aspects of instructional quality would contribute to conceptual breadth and provide more information about whether or not different aspects of instructional quality relate differently to student outcomes.

We would like to reiterate the recommendations made by Blömeke et al. (Chap. 2) concerning development of improved measures of instructional quality. These should: (1) reflect both students' and teachers' experiences; (2) have a broader scope, including the four core components, clarity of instruction, cognitive activation, classroom management and supportive climate; (3) cover each of these aspects in depth by including separate, but related, constructs; (4) be subject-specific rather than generic; and (5) include scales aimed at capturing the qualities rather than the frequency of various activities.

In the studies that employed student ratings of instructional quality, the class-level relations with achievement had differing magnitude and signs. Such differences may be due to the influence of response styles, and it is important that further research investigates these issues more closely. The TIMSS and PIRLS 2011 database would be excellently suited as a basis for initial attempts to deal with these issues, because it includes both student and teacher ratings of instructional quality in three different subjects.

While the need to broaden the measures is most clearly felt for instructional quality, teacher quality also lacks sufficient variables to measure the full breadth of this construct. Teacher quality should, for instance, also include teachers' beliefs (see Goe 2007), but TIMSS does not include such measures.

7.3.2 *Causal Inference*

It is a well-known limitation of TIMSS that each cycle only collects cross-sectional data. With such data, it is essential not to interpret correlations as expressing causal relations. Gustafsson and Nilsen (Chap. 4) included data from two TIMSS cycles (2007 and 2011) and used analytical methods that provide better support for causal inference than data from one time-point only, because the analysis removed the effect of the omitted variables that are fixed characteristics of the educational systems. While it compensates for many omitted variables, this longitudinal approach does not protect against effects of omitted time-varying characteristics. It also assumes that the estimated causal effect has the same magnitude in all countries. The limitations of both the cross-sectional and the longitudinal techniques invite questions surrounding the agreement of results across these two approaches.

The cross-sectional study of grade four students that Blömeke et al. (Chap. 2) undertook comes closest to Gustafsson and Nilsen's longitudinal study of grade eight students. Several teacher quality indicators were available in both studies, and

with these we can make some comparisons between the pooled models in both the cross-sectional and longitudinal analyses.

For the highest level of formal education completed, there was a significant relation with achievement in the cross-sectional analysis, and this was also the case in the longitudinal analysis when separate analyses were made for OECD and non-OECD countries. A major qualification in mathematics had a significant relation with achievement in the cross-sectional analysis, but not in the longitudinal analysis, and the same pattern was found for number of years of teaching experience. Professional development showed no relation with achievement in the cross-sectional model, but was linked in the longitudinal one. Conversely, preparedness (or self-efficacy) had a significant relation with achievement in the cross-sectional analysis, but not in the longitudinal analysis.

Thus, in some cases, the findings of the two studies overlapped, and, in other cases, not. This calls for further research. The longitudinal analysis estimates a common effect for all educational systems, thus if there are differences between educational systems in the strength and sign of effects this may cause the effects of positive and negative relations to cancel each other out. This could partly explain the lack of significant findings in Chap. 4.

We know from many published studies that there is a positive relation between SEAS and achievement at both student and class levels (see for example, Nilsen and Gustafsson 2014; Martin et al. 2013). In the longitudinal study, there was a significant effect of SEAS on achievement in a latent variable model with SEAS defined by five items. However, when separate analyses were performed for each item, the relation was found to be due to one item asking about parental support, and to one item asking about students' desire to learn. These results suggest that parents and students are more important for the relation between achievement and SEAS than are school factors. The results also suggest that the SEAS construct is multi-dimensional, which needs to be further explored.

The country-level longitudinal approach is a simple way of strengthening the credibility of causal statements based on ILSA data (Gustafsson 2013). As more countries participate in adjacent cycles, the approach becomes more powerful. However, the approach also requires that items are maintained unchanged over cycles; it is thus a great pity to see that few of the items included in TIMSS 1995 are still in use. It is therefore essential that, as questionnaires are changed to improve the measurement of constructs, new items are added while, at the same time, old items are kept.

7.3.3 Design

TIMSS is designed so that only one class per selected school is typically included in the sample; this causes the school and class levels to be confounded. This is unfortunate given that there are both theoretical and empirical reasons to expect that school- and class-level factors and processes are differentially related to

achievement (Yang Hansen et al. 2014). Furthermore, software is now available to allow powerful analyses of such three-level data. We therefore recommend sampling of two classes from each school, when possible.

7.4 Implications

This book can contribute to educational policy, the field of educational effectiveness and practice. Educational policy may benefit from the study findings that point to the importance of teacher quality, and especially teacher education and professional development for high instructional quality and for student achievement in mathematics. Instructional quality was also found to be related to school climate and to student motivation in mathematics. Hence, providing first and foremost an orderly school climate, but also a climate where teachers, students and parents collectively prioritize success and learning, may create the foundations for high instructional quality and boost student motivation in mathematics. This finding is extremely important in addressing the international challenges related to the decline of students' participation in STEM-related studies and careers (OECD 2014). Moreover, the results identify the potential importance of instructional quality in reducing bullying.

The studies found that there were large cultural diversities and heterogeneous findings across the educational systems with respect to the relation between teacher quality, instructional quality and student achievement. Nevertheless, patterns could be identified within groups of countries, confirming previous research that identified countries clustering (Olsen 2006; Olsen et al. 2005). Further research in this area could result in policy-relevant differentiation of knowledge for different categories of educational systems.

Our findings extend existing research on the importance of school climate by: (1) including a wide range of countries across all continents, (2) including three aspects of school climate in the same study (SEAS, safety, and order), and (3) identifying relations with student motivation and instructional quality.

Our work also contributes by applying advanced methodology in the context of international large-scale surveys. Some of the methods used in this book are new and were not previously applied in the field of educational research (such as SEM for longitudinal country-level analyses). The results highlight the integral challenges with some methods (such as using random slopes on the class level), and suggest the need for further methodological research.

7.5 Concluding Remarks

We have investigated countries from all over the world and performed both cross-sectional and trend analyses, while incorporating school and student home background contexts. The studies demonstrated the importance of teacher quality, school climate, and instructional quality for educational outcomes, and although there is not yet a coherent and international understanding of these relations, this research demonstrates progress and the value of international large-scale surveys. ILSAs view the world as a global educational laboratory, providing golden opportunities to investigate questions important to educational policy, research and practice.

References

- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133–180.
- Blömeke, S., & Delaney, S. (2012). Assessment of teacher knowledge across countries: A review of the state of research. *ZDM*, 44, 223–247.
- Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology*, 51(1), 171–200. doi:10.1146/annurev.psych.51.1.171.
- Creemers, B., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. Abingdon: Routledge.
- Desimone, L., Smith, T., & Frisvold, D. (2009). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy*, 24, 267–329.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9.
- Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from <http://www.gtlcenter.org/sites/default/files/docs/LinkBetweenTQandStudentOutcomes.pdf>.
- Good, T. L., Wiley, C. R., & Florez, I. R. (2009). Effective teaching: An emerging synthesis. In *International handbook of research on teachers and teaching* (pp. 803–816). Dordrecht: Springer.
- Gustafsson, J.-E. (2013). Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement 1. *School Effectiveness and School Improvement*, 24(3), 275–295.

- Hansen, K., & Munk, I. (2012). Exploring the measurement profiles of socioeconomic background indicators and their differences in reading achievement: A two-level latent class analysis. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 5, 49–77.
- Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). New York, NY: Waxmann Publishing Co.
- Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology*, 105(3), 805.
- Kyriakides, L., Creemers, B., Muijs, D., Rekers-Mombarg, L., Papastilianou, D., Van Petegem, P., & Pearson, D. (2014). Using the dynamic model of educational effectiveness to design strategies and actions to face bullying. *School Effectiveness and School Improvement*, 25(1), 83–104. doi:10.1080/09243453.2013.771686.
- Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, 47(2), 106–124. doi:10.1080/00461520.2012.670488.
- Martin, M. O., Foy, P., Mullis, I. V. S., & O'Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—Implications for early learning* (pp. 109–178). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Nilsen, T., & Gustafsson, J. E. (2014). School emphasis on academic success: Exploring changes in science performance in Norway between 2007 and 2011 employing two-level SEM. *Educational Research and Evaluation*, 20(4), 308–327.
- OECD. (2014). *Education at a Glance 2014: OECD Indicators*. Paris: OECD Publishing. doi:10.1787/eag-2014-en.
- Olsen, R. V. (2006). A Nordic profile of mathematics achievement: Myth or reality? *Northern lights on PISA 2003: A reflection from the Nordic countries*, 33–45.
- Olsen, R. V., Kjærnsli, M., & Lie, S. (2005). *Similarities and differences in countries' profiles of scientific literacy in PISA 2003*.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38, 109–119.
- Rutkowski, D. J., Rutkowski, L., & Wild, J. (2013). Predictors of school violence internationally: The importance of immigrant status and other factors. Paper presented at the 5th IEA International Research Conference, Singapore. Retrieved from <http://www.iea.nl/irc-2013.html>
- 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, 1550. doi:10.3389/fpsyg.2015.01550.
- Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology*, 7, 110. doi:10.3389/fpsyg.2016.00110.
- Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM*, 1–12. doi:10.1007/s11858-016-0765-0.

- Stroet, K., Opendakker, M.-C., & Minnaert, A. (2013). Effects of need supportive teaching on early adolescents' motivation and engagement: A review of the literature. *Educational Research Review*, 9, 65–87. doi:10.1016/j.edurev.2012.11.003.
- Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research*, 83(3), 357–385.
- Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration*. Wellington: Ministry of Education.
- Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2015). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*. Retrieved from <http://dx.doi.org/10.1037/edu0000075>.
- Wang, M.-T., & Degol, J. L. (2015). School climate: A review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review*, 1–38. doi:10.1007/s10648-015-9319-1.
- Wang, M.-T., & Eccles, J. S. (2013). School context, achievement motivation, and academic engagement: A longitudinal study of school engagement using a multidimensional perspective. *Learning and Instruction*, 28, 12–23. doi:10.1016/j.learninstruc.2013.04.002.
- Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73, 89–122.
- Yang Hansen, K., Gustafsson, J.-E., & Rosén, M. (2014). School performance differences and policy variations in Finland, Norway and Sweden. In *Northern lights on TIMSS and PIRLS 2011*. Denmark: Nordic Council of Ministers.

Appendix A

Country-specific descriptives, including information about their distribution in terms of skewness and kurtosis (see Tables A.1, A.2, A.3, A.4, A.5 and A.6.).

Table A.1 Country-specific descriptives of teacher's years of experience

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
Poland	0.00	3.00	2.74	0.63	-2.72	(0.02)	7.18	(0.04)
Lithuania	0.00	3.00	2.68	0.56	-1.77	(0.06)	3.26	(0.13)
Armenia	0.00	3.00	2.61	0.71	-2.03	(0.07)	3.93	(0.13)
Italy	0.00	3.00	2.55	0.77	-1.74	(0.01)	2.37	(0.03)
Serbia	0.00	3.00	2.52	0.68	-1.41	(0.04)	1.74	(0.08)
Hungary	0.00	3.00	2.48	0.91	-1.67	(0.04)	1.60	(0.07)
Georgia	0.00	3.00	2.43	0.83	-1.56	(0.05)	1.87	(0.10)
Azerbaijan	0.00	3.00	2.39	0.83	-1.18	(0.03)	0.50	(0.06)
Romania	0.00	3.00	2.29	0.82	-0.94	(0.02)	0.14	(0.04)
Croatia	0.00	3.00	2.28	0.93	-1.19	(0.05)	0.47	(0.10)
Slovenia	0.00	3.00	2.26	0.93	-1.09	(0.08)	0.18	(0.16)
Slovak Republic	0.00	3.00	2.22	0.98	-1.06	(0.05)	-0.01	(0.09)
Tunisia	0.00	3.00	2.16	1.06	-0.95	(0.03)	-0.45	(0.06)
Spain	0.00	3.00	2.15	1.11	-0.98	(0.02)	-0.52	(0.04)
Portugal	0.00	3.00	2.14	0.81	-0.73	(0.03)	0.08	(0.06)
Czech Republic	0.00	3.00	2.13	1.07	-0.92	(0.04)	-0.49	(0.07)
Iran	0.00	3.00	2.10	0.97	-0.90	(0.01)	-0.15	(0.02)
Finland	0.00	3.00	2.07	0.98	-0.83	(0.04)	-0.35	(0.09)
Germany	0.00	3.00	2.00	1.11	-0.71	(0.01)	-0.91	(0.03)
Morocco	0.00	3.00	1.99	0.99	-0.68	(0.02)	-0.60	(0.03)
Thailand	0.00	3.00	1.99	1.16	-0.68	(0.01)	-1.07	(0.03)
Sweden	0.00	3.00	1.97	0.93	-0.63	(0.04)	-0.45	(0.08)

(continued)



Table A.1 (continued)

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
Chinese Taipei	0.00	3.00	1.93	0.84	-0.60	(0.02)	-0.05	(0.05)
Northern Ireland	0.00	3.00	1.92	0.93	-0.39	(0.08)	-0.81	(0.17)
Norway	0.00	3.00	1.85	0.99	-0.44	(0.04)	-0.87	(0.09)
Hong Kong	0.00	3.00	1.81	0.98	-0.67	(0.06)	-0.50	(0.12)
Japan	0.00	3.00	1.81	1.24	-0.36	(0.01)	-1.53	(0.02)
Denmark	0.00	3.00	1.79	1.10	-0.36	(0.05)	-1.21	(0.09)
Korea	0.00	3.00	1.77	1.14	-0.31	(0.02)	-1.34	(0.04)
Chile	0.00	3.00	1.76	1.21	-0.39	(0.03)	-1.41	(0.05)
Yemen	0.00	3.00	1.76	0.82	-0.72	(0.02)	0.14	(0.04)
Netherlands	0.00	3.00	1.72	1.05	-0.18	(0.03)	-1.22	(0.06)
United States	0.00	3.00	1.68	1.01	-0.28	(0.01)	-1.00	(0.01)
Botswana	0.00	3.00	1.64	1.06	-0.23	(0.07)	-1.17	(0.13)
Bahrain	0.00	3.00	1.63	0.85	-0.24	(0.11)	-0.51	(0.22)
Malta	0.00	3.00	1.61	1.03	-0.03	(0.28)	-1.16	(0.54)
Saudi Arabia	0.00	3.00	1.59	1.00	-0.28	(0.02)	-0.99	(0.04)
Ireland	0.00	3.00	1.58	1.19	-0.07	(0.04)	-1.52	(0.08)
New Zealand	0.00	3.00	1.57	1.10	-0.11	(0.04)	-1.30	(0.08)
Honduras	0.00	3.00	1.53	1.04	-0.18	(0.03)	-1.15	(0.05)
Turkey	0.00	3.00	1.48	1.06	-0.08	(0.01)	-1.24	(0.02)
England	0.00	3.00	1.46	1.17	0.00	(0.02)	-1.49	(0.03)
Qatar	0.00	3.00	1.42	1.11	0.10	(0.11)	-1.33	(0.21)
United Arab Emirates	0.00	3.00	1.26	1.01	0.18	(0.06)	-1.12	(0.11)
Singapore	0.00	3.00	1.18	1.02	0.33	(0.07)	-1.08	(0.14)
Oman	0.00	3.00	1.14	0.79	0.53	(0.06)	0.08	(0.12)
Kuwait	0.00	3.00	1.01	0.83	0.22	(0.08)	-1.00	(0.15)

Note A normal distribution has a skewness of 0 and a kurtosis of 3. Countries are ordered according to the mean of the categories “less than 5 years” (0), “at least 5 years but less than 10 years” (1), “at least 10 years but less than 20 years” (2), or “20 years or more” (3) of experience

Table A.2 Country-specific descriptives of teacher’s degree from teacher education

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
Slovak Republic	1.00	5.00	4.98	0.25	-13.04	(0.05)	179.26	(0.09)
Poland	2.00	5.00	4.95	0.29	-7.44	(0.02)	64.39	(0.04)
Czech Republic	1.00	5.00	4.74	0.93	-3.37	(0.04)	9.66	(0.07)
Georgia	2.00	5.00	4.70	0.59	-2.28	(0.05)	5.65	(0.10)
Finland	1.00	5.00	4.67	0.78	-3.44	(0.04)	12.98	(0.08)
United States	4.00	5.00	4.61	0.49	-0.46	(0.01)	-1.79	(0.01)
England	2.00	5.00	4.37	0.53	-0.09	(0.02)	-0.30	(0.03)
Armenia	0.00	5.00	4.34	1.27	-1.57	(0.06)	0.88	(0.13)
Northern Ireland	3.00	5.00	4.28	0.52	0.23	(0.08)	-0.51	(0.17)

(continued)

Table A.2 (continued)

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
Chinese Taipei	2.00	5.00	4.22	0.53	-0.60	(0.02)	3.81	(0.05)
Qatar	2.00	5.00	4.17	0.65	-1.20	(0.10)	3.62	(0.21)
Bahrain	2.00	5.00	4.17	0.47	-0.30	(0.11)	5.08	(0.21)
Hong Kong	3.00	5.00	4.15	0.50	0.30	(0.06)	0.53	(0.12)
Ireland	3.00	5.00	4.14	0.43	0.73	(0.04)	1.36	(0.08)
Korea	3.00	5.00	4.13	0.50	0.26	(0.02)	0.67	(0.04)
Thailand	1.00	5.00	4.07	0.47	-2.41	(0.01)	19.08	(0.03)
Kuwait	2.00	5.00	4.04	0.31	-0.89	(0.08)	19.94	(0.15)
United Arab Emirates	1.00	5.00	4.03	0.61	-1.13	(0.06)	3.78	(0.11)
Hungary	3.00	5.00	4.02	0.18	2.31	(0.04)	26.52	(0.07)
Chile	3.00	5.00	4.01	0.46	0.05	(0.03)	1.69	(0.05)
Spain	4.00	5.00	4.01	0.10	9.73	(0.02)	92.75	(0.03)
New Zealand	3.00	5.00	4.01	0.59	0.00	(0.04)	-0.13	(0.08)
Lithuania	2.00	5.00	4.01	0.56	-1.16	(0.06)	4.33	(0.12)
Netherlands	1.00	5.00	3.99	0.30	-7.37	(0.03)	78.76	(0.06)
Norway	3.00	5.00	3.98	0.29	-0.59	(0.04)	8.65	(0.09)
Portugal	3.00	5.00	3.96	0.28	-1.28	(0.03)	9.39	(0.06)
Sweden	0.00	4.00	3.90	0.44	-6.10	(0.04)	43.58	(0.08)
Turkey	3.00	5.00	3.89	0.40	-0.81	(0.01)	2.30	(0.02)
Japan	2.00	5.00	3.86	0.62	-2.06	(0.01)	4.60	(0.02)
Denmark	1.00	5.00	3.83	0.50	-2.08	(0.05)	8.72	(0.09)
Germany	1.00	5.00	3.75	0.79	-2.55	(0.01)	6.31	(0.03)
Malta	0.00	5.00	3.74	1.14	-1.82	(0.28)	3.19	(0.56)
Oman	0.00	5.00	3.74	0.88	-1.69	(0.06)	2.91	(0.12)
Singapore	1.00	5.00	3.65	0.84	-0.91	(0.07)	0.50	(0.14)
Slovenia	3.00	5.00	3.62	0.50	-0.33	(0.08)	-1.44	(0.16)
Azerbaijan	0.00	5.00	3.61	0.80	-1.42	(0.03)	5.38	(0.06)
Saudi Arabia	2.00	5.00	3.42	0.93	-0.82	(0.02)	-1.17	(0.04)
Croatia	1.00	5.00	3.35	0.55	0.21	(0.05)	1.13	(0.10)
Serbia	1.00	5.00	3.12	1.09	-0.36	(0.04)	-1.48	(0.08)
Iran	1.00	5.00	3.07	0.99	-0.91	(0.01)	0.09	(0.02)
Botswana	1.00	5.00	3.07	0.61	-0.14	(0.06)	2.97	(0.13)
Romania	1.00	5.00	2.73	1.34	-0.12	(0.02)	-1.40	(0.04)
Honduras	0.00	4.00	2.57	1.35	-0.19	(0.02)	-1.72	(0.05)
Yemen	0.00	4.00	2.24	1.23	0.47	(0.02)	-1.30	(0.04)
Tunisia	0.00	5.00	2.14	1.13	0.32	(0.03)	-1.32	(0.06)
Morocco	0.00	5.00	1.95	1.54	0.59	(0.02)	-1.40	(0.03)
Italy	1.00	5.00	1.71	1.34	1.44	(0.01)	0.25	(0.03)

Note Countries are ordered according to the mean of the categories “Not completed ISCED level 3” (0), “Finished ISCED level 3” (1), “Finished ISCED level 4” (2), “Finished ISCED level 5B” (3), “Finished ISCED level 5A, first degree” (4), or “Finished ISCED level 5A, second degree or higher” (5)



Table A.3 Country-specific descriptives of teacher’s major in mathematics or mathematics education

Country	Min	Max	M	SD	Skewness (<i>SE</i>)	Kurtosis (<i>SE</i>)
Kuwait	0.00	1.00	0.97	0.18	-5.18 (0.08)	24.90 (0.15)
Bahrain	0.00	1.00	0.93	0.26	-3.38 (0.11)	9.44 (0.21)
Oman	0.00	1.00	0.84	0.37	-1.84 (0.06)	1.37 (0.12)
United Arab Emirates	0.00	1.00	0.81	0.40	-1.55 (0.05)	0.41 (0.11)
Azerbaijan	0.00	1.00	0.80	0.40	-1.50 (0.03)	0.25 (0.06)
Saudi Arabia	0.00	1.00	0.78	0.41	-1.35 (0.02)	-0.17 (0.04)
Georgia	0.00	1.00	0.77	0.42	-1.28 (0.05)	-0.37 (0.10)
Armenia	0.00	1.00	0.76	0.43	-1.19 (0.06)	-0.58 (0.13)
Qatar	0.00	1.00	0.70	0.46	-0.89 (0.10)	-1.22 (0.21)
Sweden	0.00	1.00	0.68	0.47	-0.77 (0.04)	-1.41 (0.08)
Hong Kong	0.00	1.00	0.66	0.48	-0.66 (0.06)	-1.57 (0.12)
Singapore	0.00	1.00	0.65	0.48	-0.65 (0.07)	-1.58 (0.14)
Thailand	0.00	1.00	0.63	0.48	-0.53 (0.01)	-1.72 (0.03)
Denmark	0.00	1.00	0.58	0.49	-0.32 (0.05)	-1.90 (0.09)
Germany	0.00	1.00	0.51	0.50	-0.04 (0.01)	-2.00 (0.03)
Yemen	0.00	1.00	0.48	0.50	0.07 (0.02)	-1.99 (0.04)
Botswana	0.00	1.00	0.44	0.50	0.26 (0.06)	-1.94 (0.13)
Tunisia	0.00	1.00	0.42	0.49	0.31 (0.03)	-1.90 (0.06)
Chile	0.00	1.00	0.36	0.48	0.59 (0.03)	-1.65 (0.05)
Romania	0.00	1.00	0.35	0.48	0.61 (0.02)	-1.63 (0.05)
Morocco	0.00	1.00	0.34	0.47	0.69 (0.02)	-1.53 (0.03)
Chinese Taipei	0.00	1.00	0.34	0.47	0.69 (0.02)	-1.52 (0.05)
Norway	0.00	1.00	0.32	0.47	0.77 (0.04)	-1.41 (0.09)
Serbia	0.00	1.00	0.31	0.46	0.83 (0.04)	-1.32 (0.08)
Spain	0.00	1.00	0.31	0.46	0.84 (0.02)	-1.29 (0.03)
Portugal	0.00	1.00	0.27	0.45	1.02 (0.03)	-0.95 (0.06)
Iran	0.00	1.00	0.26	0.44	1.10 (0.01)	-0.80 (0.02)
Netherlands	0.00	1.00	0.24	0.43	1.22 (0.03)	-0.51 (0.06)
Turkey	0.00	1.00	0.24	0.43	1.24 (0.01)	-0.47 (0.02)
England	0.00	1.00	0.23	0.42	1.31 (0.02)	-0.29 (0.03)
Japan	0.00	1.00	0.20	0.40	1.48 (0.01)	0.19 (0.02)
Poland	0.00	1.00	0.20	0.40	1.53 (0.02)	0.34 (0.04)
Honduras	0.00	1.00	0.19	0.39	1.57 (0.03)	0.48 (0.05)
Italy	0.00	1.00	0.18	0.38	1.70 (0.02)	0.88 (0.04)
Croatia	0.00	1.00	0.17	0.37	1.78 (0.05)	1.16 (0.10)
Malta	0.00	1.00	0.15	0.36	1.95 (0.28)	1.85 (0.56)
New Zealand	0.00	1.00	0.15	0.36	1.97 (0.04)	1.87 (0.08)
Slovak Republic	0.00	1.00	0.14	0.35	2.06 (0.05)	2.25 (0.09)

(continued)

Table A.3 (continued)

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
Finland	0.00	1.00	0.14	0.35	2.08	(0.04)	2.31	(0.08)
Ireland	0.00	1.00	0.13	0.33	2.27	(0.04)	3.15	(0.08)
United States	0.00	1.00	0.11	0.32	2.44	(0.01)	3.93	(0.01)
Northern Ireland	0.00	1.00	0.10	0.30	2.63	(0.08)	4.95	(0.16)
Lithuania	0.00	1.00	0.10	0.30	2.73	(0.06)	5.45	(0.12)
Korea	0.00	1.00	0.09	0.29	2.83	(0.02)	6.01	(0.04)
Hungary	0.00	1.00	0.08	0.27	3.20	(0.04)	8.24	(0.07)
Czech Republic	0.00	1.00	0.07	0.26	3.25	(0.04)	8.60	(0.07)
Slovenia	0.00	1.00	0.04	0.20	4.57	(0.08)	18.88	(0.16)

Note Countries are ordered according to the proportion of teachers with a major in mathematics or mathematics education

Table A.4 Country-specific descriptives of the item parcel indicating teacher's participation in professional development (PD) activities preparing for specific challenges of mathematics instruction (out of three item-parcels of the latent construct "PD")

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
Kuwait	0.00	1.00	0.78	0.33	-1.32	(0.08)	0.49	(0.15)
Singapore	0.00	1.00	0.70	0.37	-0.83	(0.07)	-0.77	(0.14)
New Zealand	0.00	1.00	0.68	0.41	-0.76	(0.04)	-1.12	(0.08)
Thailand	0.00	1.00	0.67	0.40	-0.67	(0.01)	-1.18	(0.03)
Hong Kong	0.00	1.00	0.66	0.37	-0.59	(0.06)	-1.07	(0.12)
Saudi Arabia	0.00	1.00	0.64	0.40	-0.61	(0.02)	-1.25	(0.04)
United States	0.00	1.00	0.64	0.40	-0.58	(0.01)	-1.26	(0.01)
Honduras	0.00	1.00	0.63	0.39	-0.53	(0.02)	-1.27	(0.05)
Armenia	0.00	1.00	0.62	0.41	-0.46	(0.06)	-1.44	(0.13)
Azerbaijan	0.00	1.00	0.59	0.34	-0.46	(0.03)	-0.92	(0.06)
Northern Ireland	0.00	1.00	0.58	0.44	-0.30	(0.08)	-1.69	(0.16)
Portugal	0.00	1.00	0.58	0.44	-0.32	(0.03)	-1.68	(0.06)
England	0.00	1.00	0.57	0.42	-0.27	(0.02)	-1.60	(0.03)
Sweden	0.00	1.00	0.54	0.40	-0.08	(0.04)	-1.52	(0.08)
Croatia	0.00	1.00	0.53	0.38	-0.11	(0.05)	-1.41	(0.10)
Qatar	0.00	1.00	0.53	0.44	-0.14	(0.11)	-1.72	(0.21)
Romania	0.00	1.00	0.52	0.43	-0.06	(0.02)	-1.67	(0.04)
United Arab Emirates	0.00	1.00	0.51	0.39	-0.05	(0.06)	-1.48	(0.11)
Serbia	0.00	1.00	0.49	0.40	0.01	(0.04)	-1.56	(0.08)
Bahrain	0.00	1.00	0.48	0.42	0.08	(0.11)	-1.64	(0.21)
Chinese Taipei	0.00	1.00	0.47	0.42	0.10	(0.02)	-1.64	(0.05)
Poland	0.00	1.00	0.46	0.38	0.08	(0.02)	-1.42	(0.04)

(continued)



Table A.4 (continued)

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
Germany	0.00	1.00	0.44	0.38	0.16	(0.01)	-1.41	(0.03)
Oman	0.00	1.00	0.44	0.40	0.28	(0.06)	-1.49	(0.12)
Japan	0.00	1.00	0.42	0.40	0.22	(0.01)	-1.52	(0.02)
Lithuania	0.00	1.00	0.40	0.39	0.39	(0.06)	-1.33	(0.13)
Iran	0.00	1.00	0.40	0.42	0.38	(0.01)	-1.54	(0.02)
Korea	0.00	1.00	0.40	0.39	0.36	(0.02)	-1.40	(0.04)
Tunisia	0.00	1.00	0.37	0.39	0.55	(0.03)	-1.19	(0.06)
Chile	0.00	1.00	0.35	0.37	0.53	(0.03)	-1.19	(0.06)
Slovenia	0.00	1.00	0.33	0.36	0.69	(0.08)	-0.85	(0.16)
Italy	0.00	1.00	0.32	0.38	0.70	(0.01)	-1.01	(0.03)
Ireland	0.00	1.00	0.29	0.37	0.91	(0.04)	-0.71	(0.08)
Hungary	0.00	1.00	0.28	0.33	0.89	(0.04)	-0.44	(0.07)
Slovak Republic	0.00	1.00	0.26	0.33	1.02	(0.05)	-0.08	(0.09)
Georgia	0.00	1.00	0.26	0.35	1.06	(0.05)	-0.28	(0.10)
Yemen	0.00	1.00	0.26	0.35	1.03	(0.02)	-0.39	(0.04)
Malta	0.00	1.00	0.24	0.35	1.20	(0.28)	0.02	(0.55)
Denmark	0.00	1.00	0.23	0.34	1.13	(0.05)	-0.14	(0.10)
Norway	0.00	1.00	0.22	0.33	1.16	(0.04)	-0.17	(0.09)
Spain	0.00	1.00	0.21	0.32	1.32	(0.02)	0.47	(0.03)
Netherlands	0.00	1.00	0.19	0.30	1.38	(0.03)	0.62	(0.06)
Czech Republic	0.00	1.00	0.18	0.28	1.47	(0.04)	1.27	(0.07)
Morocco	0.00	1.00	0.15	0.33	1.94	(0.02)	2.04	(0.03)
Botswana	0.00	1.00	0.13	0.28	2.12	(0.07)	3.28	(0.13)
Turkey	0.00	1.00	0.11	0.26	2.48	(0.01)	5.10	(0.02)
Finland	0.00	1.00	0.10	0.21	1.99	(0.04)	2.91	(0.08)

Note Countries are ordered according to the proportion of teachers that took part in specific PD activities

Table A.5 Country-specific descriptives of teacher’s sense of preparedness to teach geometry (item-parcel)

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
Poland	0.50	2.00	1.95	0.17	-4.31	(0.02)	24.04	(0.04)
Denmark	1.43	2.00	1.93	0.13	-2.24	(0.05)	4.38	(0.10)
Kuwait	1.00	2.00	1.93	0.18	-3.46	(0.08)	12.62	(0.15)
Romania	1.00	2.00	1.92	0.19	-2.96	(0.02)	9.04	(0.04)
Portugal	0.57	2.00	1.90	0.16	-2.51	(0.03)	10.19	(0.06)
United States	0.00	2.00	1.90	0.22	-3.44	(0.01)	17.20	(0.01)
Croatia	0.00	2.00	1.89	0.26	-4.01	(0.05)	20.87	(0.16)
Saudi Arabia	0.00	2.00	1.89	0.27	-3.94	(0.02)	19.79	(0.04)

(continued)

Table A.5 (continued)

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
England	0.86	2.00	1.89	0.23	-2.76	(0.02)	7.81	(0.03)
Czech Republic	0.75	2.00	1.88	0.26	-2.58	(0.04)	6.27	(0.07)
Northern Ireland	0.00	2.00	1.88	0.24	-3.26	(0.08)	16.94	(0.17)
Lithuania	1.00	2.00	1.87	0.22	-1.86	(0.06)	3.03	(0.13)
Slovak Republic	0.57	2.00	1.86	0.25	-1.98	(0.05)	3.49	(0.10)
Qatar	0.00	2.00	1.86	0.32	-3.22	(0.11)	12.52	(0.21)
Malta	1.00	2.00	1.86	0.26	-1.98	(0.28)	3.25	(0.56)
Chinese Taipei	0.86	2.00	1.84	0.30	-1.77	(0.02)	1.88	(0.05)
Botswana	0.14	2.00	1.84	0.30	-2.53	(0.07)	7.57	(0.13)
Oman	0.29	2.00	1.84	0.28	-1.87	(0.06)	3.29	(0.12)
Ireland	0.71	2.00	1.83	0.28	-1.70	(0.04)	2.13	(0.08)
Spain	0.67	2.00	1.83	0.31	-1.95	(0.02)	2.74	(0.03)
Chile	0.71	2.00	1.83	0.28	-1.64	(0.03)	1.98	(0.06)
Singapore	0.50	2.00	1.83	0.31	-1.92	(0.07)	3.04	(0.14)
United Arab Emirates	0.00	2.00	1.83	0.38	-2.71	(0.06)	7.40	(0.12)
Georgia	0.00	2.00	1.82	0.33	-2.66	(0.05)	8.98	(0.10)
Slovenia	0.43	2.00	1.82	0.30	-2.04	(0.08)	4.51	(0.16)
Tunisia	0.57	2.00	1.82	0.32	-2.03	(0.03)	3.51	(0.07)
Serbia	0.40	2.00	1.81	0.31	-1.74	(0.04)	2.31	(0.08)
Morocco	0.00	2.00	1.80	0.37	-2.31	(0.02)	5.56	(0.04)
Bahrain	0.75	2.00	1.80	0.33	-1.59	(0.11)	1.24	(0.22)
Hungary	0.50	2.00	1.78	0.33	-1.41	(0.04)	0.84	(0.08)
Norway	0.71	2.00	1.77	0.35	-1.43	(0.04)	0.75	(0.09)
Finland	0.00	2.00	1.77	0.32	-1.96	(0.04)	4.87	(0.08)
Armenia	0.00	2.00	1.77	0.36	-2.22	(0.06)	5.86	(0.13)
Netherlands	0.00	2.00	1.76	0.37	-1.98	(0.03)	4.69	(0.06)
New Zealand	0.14	2.00	1.73	0.34	-1.37	(0.04)	1.38	(0.08)
Turkey	0.29	2.00	1.73	0.37	-1.60	(0.01)	2.20	(0.02)
Iran	0.33	2.00	1.73	0.37	-1.45	(0.01)	1.49	(0.03)
Germany	0.57	2.00	1.73	0.33	-1.30	(0.01)	1.03	(0.03)
Korea	0.00	2.00	1.72	0.42	-1.56	(0.02)	1.77	(0.04)
Hong Kong	0.00	2.00	1.71	0.42	-1.47	(0.06)	1.75	(0.12)
Sweden	0.00	2.00	1.70	0.35	-1.39	(0.04)	2.18	(0.08)
Azerbaijan	0.29	2.00	1.67	0.34	-1.06	(0.03)	0.87	(0.07)
Italy	0.67	2.00	1.66	0.38	-0.86	(0.02)	-0.58	(0.03)
Yemen	0.00	2.00	1.61	0.52	-1.38	(0.02)	0.95	(0.05)
Honduras	0.00	2.00	1.51	0.48	-0.89	(0.02)	0.30	(0.05)
Japan	0.14	2.00	1.48	0.46	-0.23	(0.01)	-1.26	(0.03)
Thailand	0.00	2.00	1.40	0.46	-0.30	(0.01)	-0.37	(0.03)

Note Countries are ordered according to the mean of the categories “Not well prepared” (0), “Somewhat prepared” (1) and “Very well prepared” (2)

Table A.6 Country-specific descriptives of the item-parcel indicating the frequency with which teachers implemented InQua in terms of cognitive activation (out of three item-parcels of the latent construct “InQua”)

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
England	1.50	3.00	2.89	0.25	-2.55	(0.02)	7.73	(0.03)
Poland	1.00	3.00	2.87	0.35	-2.75	(0.02)	7.52	(0.04)
Qatar	0.50	3.00	2.85	0.38	-2.99	(0.10)	10.15	(0.21)
Slovak Republic	1.00	3.00	2.84	0.34	-2.21	(0.05)	4.35	(0.09)
Romania	1.50	3.00	2.84	0.36	-2.19	(0.02)	3.70	(0.04)
Georgia	1.50	3.00	2.84	0.33	-1.87	(0.05)	2.28	(0.10)
Hungary	1.00	3.00	2.83	0.37	-2.91	(0.04)	9.39	(0.07)
United States	1.00	3.00	2.83	0.34	-2.50	(0.01)	7.55	(0.01)
Malta	1.50	3.00	2.81	0.38	-1.93	(0.27)	2.85	(0.54)
Lithuania	1.00	3.00	2.80	0.38	-2.40	(0.06)	6.76	(0.12)
United Arab Emirates	0.50	3.00	2.79	0.40	-2.26	(0.06)	5.78	(0.11)
Azerbaijan	0.50	3.00	2.78	0.42	-2.51	(0.03)	7.08	(0.06)
Czech Republic	1.00	3.00	2.78	0.36	-1.69	(0.04)	2.75	(0.07)
Portugal	1.00	3.00	2.78	0.41	-1.82	(0.03)	2.56	(0.06)
Northern Ireland	0.00	3.00	2.78	0.43	-2.91	(0.08)	12.67	(0.16)
Italy	1.50	3.00	2.77	0.41	-1.61	(0.01)	1.41	(0.03)
Armenia	1.00	3.00	2.76	0.42	-2.06	(0.06)	4.22	(0.13)
Serbia	1.50	3.00	2.76	0.39	-1.42	(0.04)	0.93	(0.08)
Japan	1.50	3.00	2.76	0.37	-1.38	(0.01)	1.12	(0.02)
Croatia	1.50	3.00	2.74	0.40	-1.52	(0.05)	1.54	(0.10)
Korea	0.00	3.00	2.74	0.39	-2.38	(0.02)	10.75	(0.04)
Oman	0.50	3.00	2.70	0.48	-1.79	(0.06)	3.36	(0.12)
Slovenia	1.50	3.00	2.70	0.45	-1.41	(0.08)	0.95	(0.16)
Chile	1.50	3.00	2.70	0.43	-1.20	(0.03)	0.29	(0.05)
Iran	1.00	3.00	2.69	0.43	-1.58	(0.01)	2.58	(0.02)
Botswana	0.50	3.00	2.65	0.57	-1.63	(0.06)	1.84	(0.13)
Singapore	1.00	3.00	2.64	0.52	-1.37	(0.07)	0.98	(0.14)
Thailand	1.00	3.00	2.64	0.48	-1.24	(0.01)	0.96	(0.03)
Spain	1.00	3.00	2.62	0.49	-0.93	(0.02)	-0.34	(0.03)
Turkey	0.00	3.00	2.62	0.54	-1.56	(0.01)	2.44	(0.02)
Bahrain	1.50	3.00	2.62	0.47	-0.86	(0.11)	-0.48	(0.21)
Ireland	1.00	3.00	2.61	0.45	-0.85	(0.04)	-0.17	(0.08)
Hong Kong	1.00	3.00	2.61	0.50	-1.20	(0.06)	0.91	(0.12)
Saudi Arabia	1.00	3.00	2.60	0.52	-1.16	(0.02)	0.50	(0.04)
New Zealand	1.00	3.00	2.58	0.46	-1.01	(0.04)	0.54	(0.08)
Honduras	1.00	3.00	2.57	0.58	-1.14	(0.02)	0.23	(0.05)
Tunisia	1.00	3.00	2.53	0.56	-0.88	(0.03)	-0.20	(0.06)
Kuwait	1.00	3.00	2.51	0.56	-0.91	(0.08)	0.09	(0.15)

(continued)

Table A.6 (continued)

Country	Min	Max	M	SD	Skewness (<i>SE</i>)		Kurtosis (<i>SE</i>)	
Morocco	1.00	3.00	2.50	0.61	-0.97	(0.02)	-0.15	(0.03)
Germany	0.50	3.00	2.44	0.50	-0.77	(0.01)	0.46	(0.03)
Netherlands	1.00	3.00	2.38	0.56	-0.64	(0.03)	-0.46	(0.06)
Yemen	0.50	3.00	2.34	0.60	-0.61	(0.02)	-0.24	(0.04)
Norway	1.00	3.00	2.33	0.47	-0.37	(0.04)	-0.28	(0.09)
Chinese Taipei	1.00	3.00	2.33	0.62	-0.55	(0.02)	-0.68	(0.05)
Sweden	1.00	3.00	2.25	0.55	-0.64	(0.04)	-0.25	(0.08)
Finland	0.00	3.00	2.23	0.54	-0.18	(0.04)	-0.47	(0.08)
Denmark	1.00	3.00	2.02	0.53	-0.04	(0.05)	-0.56	(0.09)

Note Countries are ordered according to the mean of the categories “Never” (0), “Some lessons” (1), “Half the lessons” (2), or “Every lesson” (3)



Appendix B

Establishing measurement invariance across 47 countries of the latent constructs used in Chap. 3. (Note that as the configural model fits perfectly, a comparison with the metric and scalar models is not meaningful.) (Tables B.1, B.2 and B.3).

Table B.1 Professional development (three item parcels, TG10 and TM11)

Model	χ^2	df	p	RMSEA	CFI	Δ RMSEA	Δ CFI
Metric	93.76	92	0.43	0.00	1.00	–	–
Scalar	1561.09	184	<0.01	0.04	0.39	0.04	–0.61

Table B.2 Preparedness (three item parcels, TM12)

Model	χ^2	df	p	RMSEA	CFI	Δ RMSEA	Δ CFI
Metric	242.23	92	<0.01	0.02	0.96	–	–
Scalar	867.60	184	<0.01	0.03	0.84	0.01	–0.13

Table B.3 Instructional quality (three item parcels, TG15)

Model	χ^2	df	p	RMSEA	CFI	Δ RMSEA	Δ CFI
Metric	231.91	92	<0.01	0.02	0.92	–	–
Scalar	1132.83	184	<0.01	0.03	0.47	0.01	–0.45

Appendix C

Measurement invariance testing of measures across 50 countries (Table C.1).

Table C.1 Results of measurement invariance testing of measures across 50 countries

Invariance model	SB- χ^2 (df)	CFI	TLI	RMSEA	SRMR _w	SRMR _b	Δ CFI	Δ TLI	Δ RMSEA	Δ SRMR _w	Δ SRMR _b
<i>School emphasis on academic success (5 items)—L2 invariance</i>											
Configural	434.2 (101)*	0.975	0.876	0.025	0.000	0.050	-	-	-	-	-
Metric	1245.5 (396)*	0.936	0.919	0.020	0.000	0.066	-0.029	+0.043	-0.005	0.000	+0.016
Scalar	5507.5 (592)*	0.630	0.688	0.039	0.000	0.073	-0.345	-0.288	+0.014	0.000	+0.023
<i>Safety and order in schools (5 items)—L2 invariance</i>											
Configural	538.9 (200)*	0.979	0.948	0.018	0.000	0.044	-	-	-	-	-
Metric	957.0 (348)*	0.963	0.946	0.018	0.000	0.056	-0.016	-0.002	0.000	0.000	+0.012
Scalar	2279.3 (495)*	0.891	0.890	0.026	0.000	0.057	-0.088	-0.058	+0.008	0.000	+0.013
<i>Instructional quality (4 items)—L1 invariance</i>											
Configural	2421.2 (100)*	0.988	0.963	0.064	0.019	-	-	-	-	-	-
Metric	4455.5 (247)*	0.978	0.973	0.055	0.051	-	-0.010	+0.010	-0.009	+0.032	-
Scalar	27,385.8 (394)*	0.857	0.891	0.110	0.142	-	-0.131	-0.072	+0.046	+0.123	-
<i>Instructional quality (4 items)—L2 invariance</i>											
Configural	921.0 (100)*	0.995	0.973	0.038	0.002	0.079	-	-	-	-	-
Metric	1992.2 (271)*	0.990	0.979	0.034	0.002	0.077	-0.005	+0.006	-0.004	0.000	-0.002
Scalar	15,037.6 (413)*	0.919	0.882	0.079	0.003	0.260	-0.076	-0.091	+0.041	+0.001	+0.181
<i>Academic self-concept in mathematics (3 items)—L1 invariance</i>											
Configural	0.0 (0) [#]	1.000	1.000	0.000	0.000	-	-	-	-	-	-
Metric	2548.0 (98)*	0.988	0.981	0.067	0.043	-	-	-	-	-	-
Scalar	11,568.8 (196)*	0.943	0.957	0.102	0.078	-	-	-	-	-	-
<i>Academic self-concept in mathematics (3 items)—L2 invariance</i>											
Configural	197.2 (0)*	0.999	1.000	0.000	0.001	0.030	-	-	-	-	-
Metric	763.8 (148)*	0.997	0.993	0.027	0.001	0.042	-0.002	-0.007	+0.027	0.000	+0.012
Scalar	11,235.5 (295)*	0.940	0.939	0.081	0.002	0.079	-0.059	-0.061	+0.054	+0.001	+0.049

(continued)

Table C.1 (continued)

Invariance model	SB- χ^2 (df)	CFI	TLI	RMSEA	SRMR _w	SRMR _b	Δ CFI	Δ TLI	Δ RMSEA	Δ SRMR _w	Δ SRMR _b
<i>Intrinsic value (6 items)—L1 invariance</i>											
Configural	2521.8 (350)*	0.995	0.990	0.033	0.013	—	—	—	—	—	—
Metric	9371.7 (595)*	0.982	0.977	0.051	0.067	—	-0.013	-0.013	+0.018	+0.054	—
Scalar	39,177.0 (840)*	0.920	0.929	0.090	0.127	—	-0.075	-0.046	+0.057	+0.114	—
<i>Intrinsic value (6 items)—L2 invariance</i>											
Configural	1418.9 (350)*	0.998	0.991	0.023	0.001	0.077	—	—	—	—	—
Metric	3218.6 (645)*	0.995	0.988	0.027	0.002	0.116	-0.003	-0.003	+0.004	+0.001	+0.039
Scalar	28,400.2 (939)*	0.944	0.910	0.072	0.003	0.165	-0.054	-0.081	+0.049	+0.001	+0.088
<i>Extrinsic value (5 items)—L1 invariance</i>											
Configural	2535.3 (150)*	0.990	0.966	0.053	0.015	—	—	—	—	—	—
Metric	5870.3 (346)*	0.976	0.966	0.053	0.046	—	-0.014	0.000	0.000	+0.031	—
Scalar	26,179.7 (542)*	0.889	0.898	0.092	0.094	—	-0.101	-0.068	+0.039	+0.079	—
<i>Extrinsic value (5 items)—L2 invariance</i>											
Configural	629.7 (150)*	0.998	0.988	0.024	0.001	0.096	—	—	—	—	—
Metric	1517.3 (396)*	0.996	0.989	0.022	0.002	0.120	-0.002	+0.001	-0.002	+0.001	+0.024
Scalar	18,820.1 (641)*	0.931	0.892	0.071	0.002	0.186	-0.067	-0.096	+0.047	+0.001	+0.090

Note SB- χ^2 (df) = Satorra-Bentler corrected chi-square statistic with degrees of freedom (Satorra and Bentler 2010). For model fit comparisons, the configural models served as references. # As the student-level configural model is fully identified (i.e., $df = 0$), a comparison with the metric and scalar model is only trivial

* $p < 0.05$



Appendix D

Measurement invariance testing of bullying (Table D.1) and instructional quality (Box D.1).

Table D.1 Measurement invariance: bullying

Model	χ^2	df	p	RMSEA	CFI	TIL	$\Delta\chi^2$	p	RMSEA	CFI
Configural invariance	8215.61	918	<0.01	0.039	0.961	0.939	–	–	–	–
Metric invariance	15446.28	1481	<0.01	0.043	0.926	0.928	7230.67	<0.01	0.004	–0.035
Scalar invariance	21519.02	1574	<0.01	0.05	0.895	0.904	6072.74	<0.01	0.007	–0.031

Box D.1 Measurement invariance for instructional quality

The item SM2A-E reflecting instructional quality in the student questionnaire in grade 4 TIMSS 2011 was analyzed

<i>Student-level models</i>	<i>Student- and class-level models</i>
In the first step a series of one-level models was run on student data, in order to get a first impression of the dimensionality of the data. There were indications of a two-factor structure, the two items SM2D and SM2E forming one factor, presumably related to interest, and the other four items another factor. However, the two factors were highly correlated and in several countries the correlation was close to unity. An alternative one-factor model with a covariance among the residuals of items D and E therefore was tested. It also was observed that item SM2B	Here we are of course interested in investigating metric invariance at both student- and class-levels, the important thing being that metric invariance can be established at class-level given that both outcomes and other independent variables are at this level of observation In the first step a one-factor two-level model was fitted, with the covariances for items D and E included at the student level. This 52-group model was fitted under the Mplus default, which imposes constraints on intercepts and factor loadings, implying an

(continued)

(continued)

The item SM2A-E reflecting instructional quality in the student questionnaire in grade 4 TIMSS 2011 was analyzed

in most of the countries had low loadings on the latent variable, so this item was excluded, leaving five items for the ensuing analyses. The student-level one-factor model had good fit to data and when the covariance between the residuals of items D and E was added fit improved. When the metric invariance requirement was imposed fit was worsened. However, with ΔCFI at 0.02 it could be argued that metric invariance is at least marginally supported with student-level data

assumption of scalar invariance. This model had poor fit, as indicated by a CFI of 0.874. Relaxing the equality constraints on the intercepts across countries improved fit dramatically (CFI = 0.962, ΔCFI = 0.088). Relaxing the equality constraints on the student-level factor loadings caused some further improvement (CFI = 0.979, ΔCFI = 0.017), but small enough to support the claim of student-level metric invariance. Relaxing the equality constraints on the class-level factor loadings caused little improvement (CFI = 0.982, ΔCFI = 0.003), supporting class-level metric invariance

In conclusion, these results show that across all the 52 countries investigated there was metric invariance at class- and student-levels.

Reference

- Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, 75(2), 243–248.