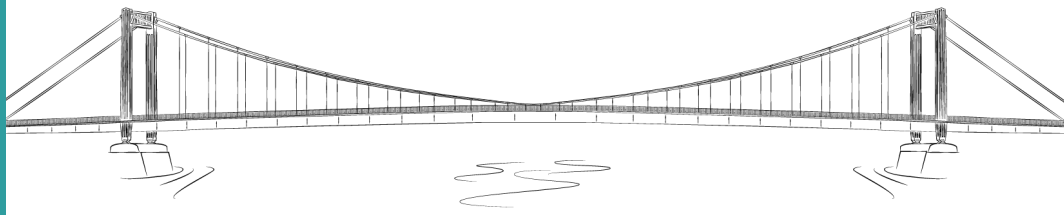
ACTA UNIVERSITATIS GOTHOBURGENSIS
GOTHENBURG STUDIES IN EDUCATIONAL SCIENCES 470



Linking recent and older IEA studies on mathematics and science

Erika Majoros

UNIVERSITY OF GOTHENBURG

Linking Recent and Older IEA Studies on Mathematics and Science

ACTA UNIVERSITATIS GOTHOBURGENSIS
GOTHENBURG STUDIES IN EDUCATIONAL SCIENCES 470

Linking recent and older IEA studies on mathematics and science

Erika Majoros

ISBN 978-91-7963-109-3 (printed)
ISBN 978-91-7963-110-9 (pdf)
ISSN 0436-1121

Editor: Elisabet Öhrn
The publication is also available in full text at:
http://hdl.handle.net/2077/71965

Subscriptions to the series and orders for individual copies sent to: Acta Universitatis Gothoburgensis, PO Box 222, SE-405 30 Göteborg, Sweden or to acta@ub.gu.se

Cover image: *Elisabeth Bridge (Erzsébet híd), Budapest, Hungary* by Zita Sramkó

Photographer: Victoria Rolfe

Print:
Stema Specialtryck AB, Borås, 2022

Abstract

Title:	Linking recent and older IEA studies on mathematics and science
Author:	Erika Majoros
Language:	English with a Swedish summary
ISBN:	978-91-7963-109-3 (printed)
ISBN:	978-91-7963-110-9 (pdf)
ISSN:	0436-1121
Keywords:	linking and scaling, international large-scale assessments, mathematics achievement, science achievement, mathematics motivation, TIMSS

The purpose of this thesis was to develop procedures that allow researchers to make reasonable comparisons of grade-eight mathematics and science achievement and motivation scales over a long time period, despite changes to the instruments, populations, and procedures between administrations.

The data were selected from international large-scale assessments administered by the International Association for the Evaluation of Educational Achievement. Student data were used from the Trends in International Mathematics and Science Study (TIMSS) and its four predecessors conducted before 1995.

The assessments have targeted slightly different populations (13-year-olds, 14-year-olds, and eighth-grade students), and the constructs changed somewhat across administrations. The thesis, therefore, aimed to: 1) evaluate the degree of comparability across these assessments; 2) link the cognitive test results onto the TIMSS reporting scale with the use of item response theory (IRT) modeling; 3) explore the feasibility of linking the motivational scales in these assessments with different approaches in the IRT and structural equation modeling frameworks.

Despite the assessments being carried out since the 1960s, a high level of stability in the inferences and cognitive constructs of the assessments was found. The measurement of the motivational constructs was found to be less stable. Overall, the results indicated that the comparability of the scales has improved over time. Different linking approaches were explored, and they yielded similar results in the country-level trend descriptions of achievement and motivation.

The linking techniques applied in this thesis may be beneficial for linking other large-scale assessments, in which changes have occurred between administrations. Moreover, with the scales established in this thesis, it is possible to examine long-term changes in educational systems. Powerful statistical approaches may be applied to these system-level longitudinal data to address causal research questions.

Contents

List of tables and figures

List of abbreviations

Abbreviation	Meaning	Page
2PL	two-parameter logistic model	29
3PL	three-parameter logistic model	29
BIB	balanced incomplete block	46
CFA	confirmatory factor analysis	33
CFI	comparative fit index	49
COMPEAT	Center for Comparative Analyses of Educational Achievement	43
CTT	classical test theory	18
DIF	differential item functioning	36
FIML	full information maximum likelihood	52
FIMS	First International Mathematics Study	22
FISS	First International Science Study	22
GPCM	generalized partial credit model	39
ICC	item characteristic curve	28
IEA	International Association for the Evaluation of Educational Achievement	17
ILSA	international large-scale assessment	17
IRT	item response theory	18
MGCFA	multiple-group confirmatory factor analysis	33
MLR	maximum likelihood estimation with robust standard errors	51
NAEP	National Assessment of Educational Progress	23
OECD	Organisation for Economic Co-operation and Development	24
PCM	partial credit model	40
PISA	Program for International Student Assessment	24
PV	plausible value	34
RCS	Reading Comprehension Study	22
RMSEA	root mean square error of approximation	49
SEM	structural equation modeling	33
SIMS	Second International Mathematics Study	20
SISS	Second International Science Study	23
SRMR	standardized root mean square residual	49
STEM	science, engineering, mathematics, and technology	68
TALIS	Teaching and Learning International Survey	33
TIMSS	Trends in International Mathematics and Science Study	17
TLI	Tucker-Lewis index	49

To Erzsébet

Acknowledgments

First and foremost, I wish to thank my supervisors, Monica Rosén and Jan-Eric Gustafsson for their support in guiding me through the Ph.D. process. I am very grateful for your availability and the space you held for me to follow my interests over the years. Our discussions, your thorough reading of my drafts, and the comments you provided were invaluable to the completion of this dissertation. Camilla, Stefan, and Victoria, your feedback at the late stages of my writing are also highly appreciated.

This thesis was completed as part of the OCCAM project, in which I had the utmost privilege of meeting excellent researchers and making great friends with fellow doctoral students. I was hosted in Gothenburg by the research group FUR. This stimulating group has been a space for me to grow as a researcher, and I thank you all. I also had the fortune to spend some time on two secondments, at ETS and IEA. I would like to extend my gratitude to my mentors at these host institutions, Eugenio J. Gonzalez and Agnes Stancel-Piątak.

I would like to thank the discussants in my planning, halfway, and final seminars, Rolf Strietholt, Björn Andersson, and Rianne Janssen for their feedback, which has supported the development of this work. I would also like to thank Rolf Vegar Olsen for serving as my defense opponent.

Probably none of the exciting experiences mentioned above would have happened if it was not for my supervisor during my master's programme in Hungary and my statistics teacher during my first year in Gothenburg, in the IMER programme. Csaba, I thank you for making me believe that I have a place at this table. Kajsa, thank you for supporting the ambitious idea of applying for a Ph.D. position.

This journey has been transformative, filled with both personal and global events that were quite challenging to endure. I am eternally grateful for your friendship, Leah. Our conversations, both personally and professionally, kept me going, inspired me to grow, or held me.

I thank you, doctoral student pals, for sharing this journey with me. I am grateful for learning from all of you, especially the Andrés, the other Andrés, Edwin, Emilie, Jonas, Lea, Tina, and everyone else in OCCAM and at GU. I am so proud of knowing every one of you.

I appreciate everyone who was part of my life during these years. I thank my friends in Gothenburg, especially Victor, for bearing with me. I am grateful for the patience of my family and friends in Hungary and elsewhere, who did not see me too

often during these four years. Csaba, Dóri, Ildi, Kati, Melitta és Médea, köszönöm, hogy biztattatok és tartottátok bennem a lelket, amikor arra volt szükségem. Andi, Ági, Blanka és Gyula, nektek is köszönöm, hogy mindig számíthattam rátok.

Finally, Erzsébet, my mother, probably would not be surprised that I ended up doing a Ph.D. training even though she never saw me finishing my BA degree. Things took unexpected turns a few times, but I tried to make the most of it and this would not have been possible without her.

Preface

Before you lies the dissertation "Linking recent and older IEA studies on mathematics and science", the basis of which are international comparative surveys that have been conducted since 1964. This thesis has been written to fulfill the graduation requirements of the Doctor in Education Program at the University of Gothenburg, Sweden. I was engaged in researching and writing this dissertation from August 2018 to August 2022.

When I was in high school, I tutored a few other students from different schools, who needed help with mathematics. This experience was profoundly influential in my interests. I saw that with enough time and patience, anyone can understand high school mathematics and that a dislike for this subject can easily grow from a lack of confidence or worse, from being shamed by teachers. I had the privilege of learning in a class specialized in natural sciences and my mathematics education was advanced and highly individualized.

When I finished high school, I first started a master's program in applied mathematics. Soon after it somehow did not feel right. I felt a need for a wider perspective of the world than "just numbers". While I still loved mathematics, I turned to special education and educational sciences. Of course, in almost every group, I turned out to be the "math nerd", who loved teaching mathematics using an abacus to young children living with visual impairment, or who genuinely enjoyed learning the programming language R.

My doctoral research started with an interest in international comparative studies of student achievement in mathematics and science. This led me to investigate how to make comparable measures of country-level educational outcomes. The most exciting school outcomes to me are the affective aspects of learning, such as motivation or academic self-confidence, especially in mathematics for the reasons mentioned above.

This background reveals my interest in the topic of the thesis. During my doctoral studies, I also had the privilege of receiving excellent methodological training and space to nerd on methods and tools beyond my courses. My hope with this work is that the long-term scales will serve interesting research on mathematics and science education and maybe spark methodological interest as a side effect.

Chapter 1 Introduction

In this thesis, recent and older international assessments on mathematics and science are explored with the aim to link these surveys and scale the system-level educational outcomes onto a common metric. The main reason for linking the assessments is to provide researchers with comparable data of grade-eight mathematics and science achievement and motivation scales over a long time period.

The scales achieved in this thesis combined with powerful analytical approaches such as country-level longitudinal modeling techniques and advanced econometric methods allow for investigating changes in educational systems. For instance, educational reforms that take effect in the long term can be evaluated on the national level. In the comparative context, longitudinal studies are useful to explore global phenomena, such as trends toward a "global curriculum" (Johansson & Strietholt, 2019; Rutkowski & Rutkowski, 2009) or changes in the "socioeconomic achievement gap" (Broer et al., 2019; Chmielewski, 2019).

Two types of educational outcomes are the focus of this thesis. Firstly, cognitive outcomes, i.e., student achievement in mathematics and science in grade eight. Secondly, affective outcomes, i.e., how motivated students are for learning mathematics. These outcomes have been measured through international large-scale assessments (ILSAs) administered by the International Association for the Evaluation of Education Achievement (IEA).

The IEA has been maintaining trend scales of mathematics and science achievement since 1995. Before that, the IEA conducted four ILSAs in these subjects, but these early assessments have not been officially linked to the Trends in International Mathematics and Science Study (TIMSS) scales. In this thesis, the ILSAs administered before 1995 are referred to as the first-phase studies, while those after as the second-phase assessments (Gustafsson, 2008).

The decision not to link the studies from the two phases was motivated by the changes that have been made to the instruments, populations, and procedures between the early survey administrations (Martin & Kelly, 1996a). Technological and methodological challenges at that time might also have constrained the feasibility of linking. During the decades since the first assessment, technical decisions have been made concerning e.g., sampling of items and test-takers or item wording. These decisions pose challenges to comparability and consequently to linking the

assessments. However, recent technical and methodological advancements allow for tackling such challenges.

Previous research has shown that it is possible to link cognitive outcomes of the early IEA ILSAs to the recent assessments, with various linking approaches. One approach has been to link scales from surveys that include common items taking advantage of item response theory (IRT) modeling. Afrassa (2005) and Strietholt and Rosén (2016) linked cognitive outcomes of mathematics and reading achievement with this approach. However, the linking study on mathematics (Afrassa, 2005) remained limited in terms of evaluating the comparability with the TIMSS reporting scale and the scope of educational systems included in the linking.

Another linking approach has been applied to scores from different regional, national, or international assessments over a long period, which relies on classical test theory (CTT). In this approach, not all surveys have overlapping items, therefore, the linking is performed under stronger assumptions related to ability distributions (see e.g., Chmielewski, 2019; Hanushek & Woessmann, 2012).

The trend measurement of affective outcomes began with the 2011 administration of TIMSS. Certain context questionnaire scales that included common items across TIMSS 2011, TIMSS 2015, and TIMSS 2019 were linked to common metrics (Martin et al., 2012, 2016; Yin & Fishbein, 2020). To the best of my knowledge, there is no previous research on extending these longitudinal affective scales.

It can be concluded that with the recent methodological advancements and the increasing role that ILSAs play in educational systems, it is worth exploring the possibilities that lay in legacy data. The contribution of this thesis is twofold. First, the linking techniques may be applied to other large-scale assessments, in which changes have occurred between administrations. Second, the achieved scales are of potential use for future longitudinal studies.

This thesis consists of an integrative essay and three empirical studies. Issues of change, comparability, linking, and scaling are investigated in the studies. The purpose of the integrating essay is to provide a comprehensive picture of the complex background of the international trend scales from a measurement perspective. The essay has been written with the intent of elaborating on the results of the empirical studies.

This thesis is guided by two overarching research questions:

1. To what extent are the student outcomes comparable across the first- and second-phase IEA assessments on mathematics and science?
2. How do different linking approaches influence the descriptions of the system-level trends?

Outline of the thesis

In the integrative essay, the second chapter reviews the evolution of IEA surveys on mathematics and science from a historical perspective. The third chapter presents the theoretical framework of the thesis. The theoretical aspects of measuring trends of educational outcomes in an international comparative context are focused on. The fourth chapter reviews the applied analytical methods for linking the assessments and the validity framework of the thesis. Then the fifth chapter summarizes the results of the empirical studies. The sixth chapter concludes the thesis, discussing limitations, highlighting its contribution and implications, and outlining future research ideas. Following the integrative essay, the empirical studies (I-III) are presented. Each study concerns linking the early IEA assessments to the TIMSS administrations from different aspects.

Study I: Measures of long-term trends in mathematics: Linking large-scale assessments over fifty years

The main purpose of this study is to evaluate the feasibility of constructing comparable trend measures of mathematics achievement as assessed by IEA from 1964 onwards. In contrast to a previous study by Afrassa (2005), the scope of the inquiry is extended by discussing the degree of similarity across the assessments, including more educational systems, and using a different linking method. Afrassa used equating procedures on data from the first three IEA studies (between 1964 and 1995) of students in Australia. In Study I, the data of the grade eight population are used from the four countries that participated in all ILSAs between 1964 and 2015: England, Israel, Japan, and the USA.

Study II: Linking the first- and second-phase IEA studies on mathematics and science

This study builds on Study I and extends the scope of the linking in three aspects. Firstly, it includes data from all participating educational systems in grade eight, i.e., 83 educational systems in the studies on mathematics and 85 on science. Secondly, an alternative linking approach is employed to place the mathematics assessments administered before 1995 onto the TIMSS trend scale. Thirdly, this study also links the early science surveys of IEA with the TIMSS administrations. After evaluating the comparability of the science achievement tests, the linking approaches are compared.

Study III: Motivation towards mathematics from 1980 to 2015: Exploring the feasibility of trend scaling

This study explores the feasibility of establishing long-term student motivational scales using eighth-grade data from the Second International Mathematics Study (SIMS) and each TIMSS administration between 1995-2015. The analyses involve five countries that participated at the seven time points: England, Hong Kong, Hungary, Israel, and the United States. The study investigates the comparability of the scales both across cultures and over time. Three approaches to scaling and linking are compared for intrinsic and extrinsic motivation. Finally, the influence of different approaches to scaling and linking on the results is discussed.

Chapter 2 The beginning of international large-scale assessments

For more than half a century, ILSAs have provided a large body of data on student achievement and background from a vast number of educational systems. What exactly are ILSAs? When looking for definitions in recent studies in educational research, a definition of large-scale international comparative assessments by Bos (2002) was found to be adopted by several researchers (e.g., Hernández-Torrano & Courtney, 2021; Johansson, 2016; Olsen, 2005): "Studies in which both achievement of certain age/grade in one or more subjects is compared across education systems and *effects* [emphasis added] of contextual factors at the system, school, classroom and student level on achievement are studied" (p. 2. Italics in original).

The issue of studying cause and effect in the context of ILSAs emerges from the above definition. The international reports of ILSAs primarily present descriptive results of student outcomes alongside background and process factors. Therefore, the effects of contextual factors on achievement are not directly studied in ILSAs but left to researchers. For this reason, in the present thesis, a modified definition of ILSAs is applied: *large-scale studies that are designed to measure system-level student achievement in one or more subjects, and contextual indicators at the system-, school-, classroom-, and student levels among educational systems with comparable target populations across systems.*

The current state of the IEA ILSAs is a result of evolution for almost six decades. In the following section, this evolution is outlined with a focus on IEA studies measuring student knowledge in the subjects of mathematics and science.

Historical overview of the early studies

The IEA was established in 1958 as an independent, international cooperation of national research institutions and governmental research agencies. As Heyneman and Lee (2013) wrote: "It began as an educational experiment. In the late 1950s, Torsten Husen [sic] from the University of Stockholm was visiting his friends Benjamin Bloom and C. Arnold Anderson at the University of Chicago" (p.38). Husén reportedly has asked, "Why don't we test for academic achievement internationally?", and made the renowned statement "The world could be our laboratory" (Heyneman & Lykins, 2008, p. 106).

The first step in the IEA international comparative research was a feasibility study known as the Pilot Twelve-Country Study administered between 1959 and 1962, which compared the educational achievements of 13-year-olds in 12 countries in four subjects: mathematics, reading comprehension, geography, and science (Härnqvist, 1975). On the rationale behind international comparisons, Härnqvist (1975) wrote:

> Almost all the authors of IEA reports are eager to point out that the project is not intended to be a sort of Olympic games in the area of school achievement. Why, then, are international comparisons considered necessary? One of the original motives for starting the project was discontent with earlier work in comparative education which provided a description of systems and curriculum but practically no information about learning outcomes. (p. 87)

The first-phase studies in mathematics and science are listed in Table 1. During the first phase, the IEA conducted separate ILSAs in mathematics and science on four occasions; data were collected on mathematics in 1964 and 1980-82 and on science in 1970-71 and 1983-84. The first ILSA including science was the Six Subject Survey and in addition to a science study (later known as the First International Science Study, FISS), it involved reading comprehension (known as the Reading Comprehension Study; RCS), literature education, English as a foreign language, French as a foreign language, and civic education.

Table 1 IEA ILSAs on mathematics and science administered in the first phase

Assessment	Time of data collection	Number of participating educational systems
First International Mathematics Study	1964	12
First International Science Study	1970-71	17
Second International Mathematics Study	1980-82	20
Second International Science Study	1983-84	23

Husén and Postlethwaite (1967) described the main goal of the First International Mathematics Study (FIMS) as follows:

> the overall aim is, with the aid of psychometric techniques, to compare outcomes in different educational systems. The fact that these comparisons are cross-national should not be taken as an indication that the primary interest was, for instance, national means and dispersions in school achievements at certain age or school levels. … The main objective of the study is to investigate the "outcomes" of various school systems by relating as many as possible of the relevant input variables (to the extent that they could be assessed) to the output assessed by international test instruments. (p. 30)

There was thus an explanatory element in the purpose of FIMS by studying the input and output of educational systems. In light of the current state of ILSAs, it is

interesting to note how it was emphasized from the start that system-level means were not the main focus of the assessment. When summarizing the first phase of IEA, Husén (1979) noted that "what in the minds of some academics was perceived as a major exercise in basic research was perceived by others as an international contest in mathematics" (p. 379). However, in the second phase of IEA, the country rankings based on ILSA results have gained considerable and increasing public attention in many countries. In line with the original intention, Klemenčič and Mirazchiyski (2018) argued that league tables should not be perceived as the ultimate product of ILSAs.

Regarding the historical context of FIMS, Howson (1999) noted that it was carried out at a time of economic and educational expansion. Husén and Postlethwaite (1967) described why the subject of mathematics was chosen to be the focus of the first international comparative assessment with two main arguments. First, the study had to be restricted in its subject matter and in the allotted time. Second, the twelve participating educational systems were concerned with improving their scientific and technical education, and mathematics was regarded to form the basis of this improvement. This effort was referred to as the *New Mathematics*, which was an international agreement on new ways of teaching mathematics. This shift in curricula and teaching practices occurred shortly after the so-called Sputnik crisis (see e.g., Truman, 1959).

The crisis was prominent in the United States and led to the development of the National Assessment of Educational Progress (NAEP). According to the description of the beginning of NAEP by Kirsch et al. (2013), Francis Keppel, the US Commissioner of Education from 1962, was responsible for reporting to Congress about the condition of education in the United States. Keppel was concerned about the lack of systematic data and pointed out that most of the information that had been previously collected focused on the inputs of education rather than on the output, such as skills and knowledge. A technical advisory group was formed in 1965 and was chaired by John Tukey, head of the Department of Statistics at Princeton University. Their work led to the NAEP, which conducted its first assessment of 17-year-old students in citizenship, science, and writing in 1969.

SIMS and the Second International Science Study (SISS), which were conducted in the early 1980s, took place in a different economic climate in which countries were experiencing the consequences of the first oil crisis (Howson, 1999). Hence, they were more interested in the value production of their educational systems. Interestingly, in their overview of SIMS, Travers and Weinzweig (1999) argued that the choice of mathematics in FIMS

> was more a matter of convenience than interest in mathematics achievement per se. The organizers believed that it would be easier to make international comparisons in mathematics than in any other area, and they felt that

> mathematics achievement would serve as a surrogate for school achievement. (p.19)

This statement refers to the choice of school subject from a different perspective than the previously mentioned two arguments by Husén and Postlethwaite (1967). The importance of mathematics as the base of scientific improvement is replaced by its function to represent general school achievement.

In the second phase, the Third International Mathematics and Science Study in 1995 was the first IEA study to test mathematics and science together. Much of the NAEP methodology was adapted and adopted for use in this study (Beaton et al., 2011). This survey is repeated every fourth year, most recently in 2019. Since 1999, the assessment is named the Trends in International Mathematics and Science Study.

After the 1995 administration of the mathematics and science study, Plomp (1998) identified two main purposes of the IEA studies: informing policymakers and educational practitioners about the quality of their education system concerning relevant reference groups and facilitating the interpretation of the observed differences. Concerning the subsequent TIMSS cycles, Gustafsson (2008) observed that there was a shift in the aim of the assessments, i.e., from explanatory to descriptive purposes.

The Organisation for Economic Co-operation and Development (OECD) launched the first Programme for International Student Assessment (PISA) in 2000 to measure 15-year-old students' reading, mathematics, and science literacy. It has been administered every third year since then with the major subject rotating among reading, mathematics, and science in each cycle. Referring to former IEA ILSAs, the first PISA assessment framework stated that

> the quality and scope of these surveys have greatly improved over the years but they provide only partial and sporadic information about student achievement in limited subject areas. The three science and mathematics surveys conducted by the IEA provide some indication of how things have changed over 30 years but the view is limited by the restricted numbers of countries participating in the early surveys and by limitations on the extent to which the tests can be compared. (OECD, 1999, p.10)

Consequently, to the first-mentioned limitation of the former IEA surveys, the most apparent difference between the IEA and OECD studies lies in the abilities they intend to measure. The IEA studies are focused on curriculum-based knowledge and skills. The OECD studies focus on competences that are considered to be important in adult life and for lifelong learning. However, the IEA and OECD studies are similar in many ways (Johansson, 2016; Olsen, 2005). As Olsen (2005) summarized, they are all sample-based with clearly defined populations, use comparable types of

instruments and processes, and apply similar psychometric methods. Furthermore, the assessments have cyclical designs with a focus on measuring trends. Both organizations provide country rankings, which attract substantial public attention. Finally, as Johansson (2016) pointed out, the results of these studies have been argued to be comparable (see e.g., M. Wu, 2010).

To summarize the differences between the goals in the two phases of IEA ILSAs, Gustafsson (2008) stated that in the first phase, "the goal was to generate knowledge about determinants and mechanisms behind educational achievement, while in phase two the goal is to describe the outcomes of different educational systems, leaving it to the different participating countries to find the explanations" (p. 2). The differences in the technical aspects of the two phases are partly due to the inevitable shortcomings of a pioneering project (Husén, 1979) and partly to the methodological improvements in the field of testing.

The methodological improvements in the second phase were largely due to developments in modern test theory, which provided the opportunity to administer much more test items and link the assessments onto a common scale (Gustafsson, 2018). In the next chapter, the theoretical frameworks for measuring educational outcomes across cultures and connecting the measurement points are outlined.

Chapter 3 Measurement of system-level educational outcomes in an international and longitudinal context

In this chapter, the theoretical underpinnings of ILSAs are discussed from a measurement perspective. This overview unfolds the layers of measurement in international comparative assessments. First, the frameworks applied for measuring educational outcomes are outlined. Second, the score scaling framework is described. Third, a conceptual framework for cross-cultural measurement is presented. Finally, the methodological framework of linking scales concludes this chapter.

In the present thesis, *measurement* is referred to as the process defined by Stevens (1968): "the assignment of numbers to aspects of objects or events according to one or another rule or convention" (p. 850). Furthermore, measurement is assumed to be based on a theory about the phenomenon to be measured because "what is relevant to measure can be determined only within an implicit or explicit theory about the phenomenon one wishes to study" (Pedhazur & Pedhazur Schmelkin, 1991, p.16). *Educational measurement* is referred to as measurement that takes place in an educational setting.

Educational systems are complex and multilevel structures. Csíkos et al. (2020) defined the function of educational measurement as providing feedback for the planning and management of educational processes at different levels of the structure. At the system level, feedback is provided by international and/or national assessments. ILSAs may be beneficial for medium- or long-term planning due to the relatively long feedback time. Participation in ILSAs may be particularly beneficial for countries that do not maintain national assessments.

In the ILSAs in the focus of the present investigation, measurements have been conducted via achievement tests and background questionnaires. Educational testing takes advantage of theories and methodologies developed in *psychometrics*, i.e., "the science concerned with evaluating the attributes of psychological tests" (Furr & Bacharach, 2014, p.9). Cronbach (1960) defined a *psychological test* as "a systematic procedure for comparing the behavior of two or more people" (p.21).

The principles underlying psychological test development have changed dramatically in recent decades (Embretson & Reise, 2000). IRT, which is also referred to as modern test theory, has become the mainstream theoretical basis for

psychological assessments, replacing the principles that are referred to as CTT. As Embretson and Reise (2000) argued, "standardized tests are developed from IRT due to the more theoretically justifiable measurement principles and the greater potential to solve practical measurement problems" (p.12).

In the CTT framework, methods and measurement procedures have several shortcomings (Hambleton et al., 1991). The most important limitation is that the test-taker characteristics such as ability and test properties such as difficulty can be interpreted only in the context of the other. The notion of ability is expressed by the *true score*, which is defined as "the expected value of observed performance on the test of interest" (p. 2), i.e., only in terms of a particular test. The difficulty of a test item is defined as "the proportion of examines in a group of interest who answer the item correctly" (p. 2-3). Therefore, under CTT, test and item characteristics change as the examinee context changes, and vice versa, in other words, the scores are test-dependent, which has negative consequences on the comparability of test scores. This and other limitations of CTT led psychometricians to develop alternative theories and models of psychological measurement. In the following, the two main frameworks for measuring latent, i.e., not directly observable constructs, such as ability are presented.

Item response theory

IRT is a model-based measurement framework (Embretson & Reise, 2000). The two fundamental ideas in IRT are that the performance of a test-taker on an item can be predicted by latent traits or abilities and that the relationship between item performance and the latent trait can be expressed by the item characteristic function or item characteristic curve (ICC; see e.g., Embretson & Reise, 2000; Hambleton et al., 1991).

IRT models include a set of assumptions about the data. Two basic assumptions underlie all IRT models. First, the ICC reflects the true relationship between the unobservable variable (ability) and the observable variables (item responses). Second, the items are locally independent. This requires that the item parameters and person parameters fully account for interrelationships between items and persons, hence, no further relationships exist in the data. One assumption common to the most widely used IRT models is unidimensionality, i.e., that only one ability is measured by the items that constitute the test.

The major distinction among the IRT models lies in the number and type of item parameters assumed to affect performance (Embretson & Reise, 2000). The discrimination or slope parameter expresses the item's relation to the latent trait. The difficulty or location parameter indicates the difficulty level at which the person has

an equal probability to get an item correct or incorrect when applying a two-parameter logistic model (2PL). Including a guessing or lower-asymptote parameter in a three-parameter logistic model (3PL) may be necessary when an item can be solved by guessing, as in multiple-choice cognitive items.

The ICC of a hypothesized item, Item 1, is shown in Figure 1. The item is binary, i.e., scored as correct (1) or incorrect (0). Item 1 has a discrimination parameter (denoted by a) set to 1.00, a difficulty parameter (denoted by b) of 1.00, and the guessing parameter (denoted by c) is 0.00, as in 2PL models. Therefore, the ability level, which is required to solve the item with a 50% probability is 1.00.

The ICC of Item 2 in Figure 2 represents the logit curve with a discrimination parameter set to 1.00, a difficulty parameter of 0.50, and the guessing parameter of 0.00. Item 2 is equally discriminating among ability levels as Item 1 but easier. The ability level, which is sufficient for solving the item with a 50% probability is 0.50.

Item 3 has a discrimination parameter set to 0.50, a difficulty parameter set to 1.00, and a guessing parameter set to 0.00. If we take a person with the ability level 1, their probability of solving Item 3 is 50%, as in the case of Item 1. However, Item 3 discriminates less well among trait levels as does Item 1, which manifests in a gentler slope of the ICC, as shown in Figure 3.

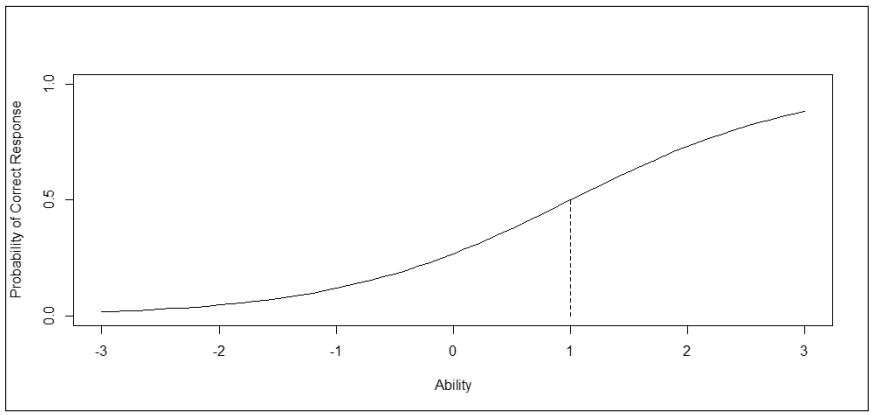


Figure 1 Item characteristic curve of a binary item – Item 1
a=1.00; b=1.00; c=0.00

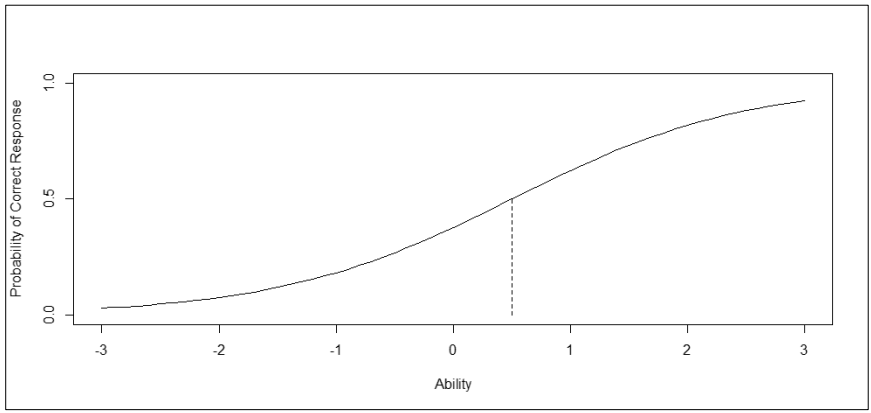


Figure 2 Item characteristic curve of a binary item – Item 2
a=1.00; b=0.50; c=0.00

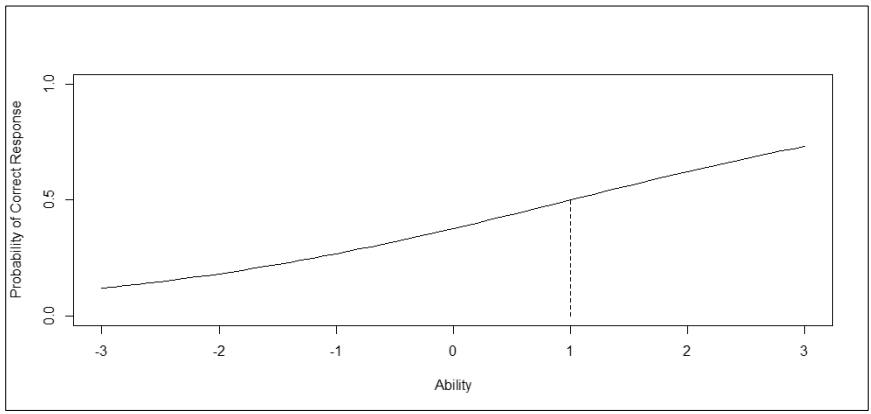


Figure 3 Item characteristic curve of a binary item – Item 3
a=0.50; b=1.00; c=0.00

Finally, in Figure 4, the discrimination and difficulty parameters of Item 4 are both set to 1 and its guessing parameter is 0.25. Hence, the chance of successfully solving this item is at least 25 %, even for test-takers with the lowest trait levels.

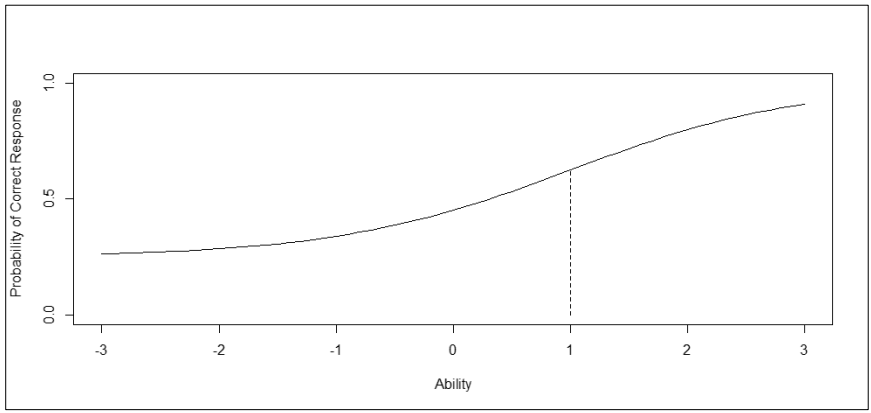


Figure 4 Item characteristic curve of a binary item – Item 4
a=1.00; b=1.00; c=0.25

The example items above were dichotomously scored, like most of the tasks in the TIMSS achievement test. However, some test items are worth more than one score point, and the background questionnaires include items with multiple ordered-response categories. For such items, polytomous IRT models are employed to represent the relationship between the latent trait and the probability of choosing a particular response category of an item.

Hambleton and Jones (1993) pointed out four advantages of applying the IRT framework for testing over CTT approaches. First, item statistics are independent of the sample from which they were estimated. Second, proficiency scores are independent of test difficulty. Third, IRT models provide a basis for matching test items to ability levels. Finally, IRT models do not require strictly parallel tests for assessing reliability. The main differences between CTT and IRT models are presented in Table 2. In this table, CTT item statistics refer to item difficulty (denoted by *p*) and item discriminating power (denoted by *r*), whereas IRT item statistics refer to the previously mentioned difficulty, discrimination, and guessing parameters.

Another advantage of IRT is that it allows for the matrix sampling of items and test-takers, which sampling method is frequently used in large-scale assessments. The practice of IRT-based large-scale assessment trend scales started with NAEP (Kirsch et al., 2013). The original NAEP reporting was done at the item level, and as this survey progressed, one of the criticisms that arose was that interpretations were fixed to the individual items used in the assessments. Messick et al. (1983) introduced the idea of using the analytic approach of IRT that supports the creation of comparable scales across multiple forms of a test.

Table 2 Main differences between classical and item response theories and models

Area	Classical test theory	Item response theory
Model	Linear	Nonlinear
Level	Test	Item
Assumptions	Weak (i.e., easy to meet with test data)	Strong (i.e., more difficult to meet with test data)
Item-ability relationship	Not specified	Item characteristic functions
Ability	Test scores or estimated true scores are reported on the test-score scale (or a transformed test-score scale)	Ability scores are reported on the scale $-\infty$ to $+\infty$ (or a transformed scale)
Invariance of item and person statistics	No – item and person parameters are sample dependent	Yes – item and person parameters are sample independent, if model fits the test data
Item statistics	p, r	b, a, and c (for the three-parameter model) plus corresponding item information functions
Sample size (for item parameter estimation)	200 to 500 (in general)	Depends on the IRT model but larger samples, i.e., over 500, in general, are needed

Note. From "Comparison of classical test theory and item response theory and their applications to test development" by Hambleton and Jones (1993, p. 43)

Except for FIMS and FISS, in all achievement tests considered in this thesis, a matrix sampling approach was applied. It means that the assessments contain more items in total than what is presented to each individual test-taker. The reasons for this are twofold. Firstly, the breadth of the skills and knowledge measured requires a correspondingly large item pool. Secondly, testing students' knowledge and application of science and mathematics content requires time-consuming item types. Essentially, under matrix sampling, samples of items are administered to samples of students (Mazzeo et al., 2006).

Matrix sampling reduces the testing burden for each student. By employing equating or linking methods, comparable results can be obtained using different test forms. Furthermore, by utilizing matrix sampling to present a subset of questions to each student, the design of ILSAs counters a potential validity threat from student attrition (Newton & Shaw, 2014) within the test-taking procedure as students are more likely to complete the test within the given time. However, matrix item sampling also imposes challenges, e.g., reporting the achievement scores on a common scale for all test-takers and communicating the approach to the different stakeholders. Matrix sampling also requires large sample sizes to appropriately calibrate test items (Hambleton & Jones, 1993).

Confirmatory factor analysis

As explained earlier, the measurement of latent traits or abilities is carried out via modeling in the IRT framework. Another approach to latent variable modeling is confirmatory factor analysis (CFA; Jöreskog, 1969). CFA is a statistical technique in the framework of structural equation modeling (SEM).

CFA considers the relationship between observed variables that serve as *indicators* and latent variables, i.e., *factors*. A factor is an underlying construct that is hypothesized to influence the indicators and accounts for the correlations among them. The pattern of relationships is expressed in *factor loadings*. *Factor scores* may be calculated using factor loadings. Factor scores represent a manifest score of the test-takers on the latent dimension (Brown, 2015).

CFA is a frequently used tool during scale development. This method is commonly employed for instance to evaluate the psychometric properties of measurement instruments, validate constructs, test measurement invariance across groups, or investigate method effects (Brown, 2015).

Questionnaire data surveying latent constructs can be scaled via CFA methods. An example of employing CFA modeling for scaling affective items is the Teaching and Learning International Survey (TALIS) administered by the OECD since 2008. Furthermore, measurement invariance testing with multiple-group confirmatory factor analysis (MGCFA) was employed to evaluate the comparability across participating educational systems (OECD, 2019).

Scaling

Stevens (1968) classified scales into four types: nominal (e.g., categories but no ordering like the numbering of football players), ordinal (e.g., a limited number of categories with an order like the Likert scale, such as the hardness scale of minerals), interval (e.g., continuous and equidistant numbers, but no true zero like degrees Celsius and Fahrenheit, calendar time), and ratio scales (e.g., continuous, equidistant, and true zero like age, length). Kolen (2006) defined *scaling* in the educational measurement framework as "the process of associating numbers or other ordered indicators with the performance of examinees on an educational test" (p.156). Test items are typically scored as correct, partially correct, or incorrect and the *raw score* is based on the number of correct answers. Different types of transformations exist to produce *scale scores* from raw scores.

In the context of ILSA methodology for producing the achievement scales, the IRT maximum likelihood scoring is the function, which is employed to achieve such transformation. The IRT proficiency scaling conventions produce scale scores with a

mean of 0 and a standard deviation of 1, which need linear transformation to have meaningful interpretations (Kolen, 2006). For instance, the TIMSS international achievement scales, at the first administration, were set with a mean of 500 and a standard deviation of 100 (Martin & Kelly, 1996b).

The plausible value (PV) methodology (see e.g., Beaton & Johnson, 1992; Mislevy, 1991; Mislevy et al., 1992; L. Rutkowski et al., 2010) is employed for generating population-level proficiency estimates from a test design with matrix sampling. Under the PV methodology, student achievement is treated as a missing value. Each student responds to a subset of test items and these responses are used to generate a student ability distribution for the population. For estimating the ability distribution of the population, a measurement-model-based extension of Rubin's (1987) multiple imputation approach is applied. From the estimated ability distribution, random draws, each referred to as a PV, are selected for each student and reported as their scale scores.

The PV methodology applied in modern ILSAs is different from other approaches that focus on individual proficiency scores (Mazzeo et al., 2006; L. Rutkowski et al., 2010). The PV approach facilitates the use of large item pools for more accurate proficiency estimation but given the matrix sampling, the reported scores are intended to be used on the group level. These types of assessments are referred to as *group-score assessments* (Mazzeo et al., 2006). The role of group-score assessments in educational measurement is

> to answer questions about what groups of students know and can do, as well as to provide comparisons of student performance across jurisdictions, to provide information concerning educational performance trends, and to indicate the extent to which jurisdictions are meeting their educational goals. (Mazzeo et al., 2006, p. 681)

ILSAs are designed to compare student outcomes across educational systems and to monitor system-level progress. There are specific challenges concerning the measurement of educational outcomes in an international context. In the next section, these challenges are outlined.

Measurement invariance

Assessments of student outcomes involving multiple educational systems need to consider cultural differences concerning, e.g., the meaning of the constructs that are to be measured. *Measurement invariance* or *equivalence* refers to the property of an assessment, which regards whether scores have the same meaning under different conditions, e.g., time of assessment, administration methods, countries, or populations (Kline, 2016; Meade & Lautenschlager, 2004).

As (Kline, 2016) pointed out, there are several labels in the literature to refer to the same type of invariance. Most terminologies derive from the hierarchy of invariance levels established by Meredith (1993). These levels of invariance represent a hierarchy of constraints on parameters across groups that increasingly attributes group differences to the common factors in the SEM framework (H. Wu & Estabrook, 2016).

The hierarchy of factorial invariance is comprised of four levels: configural (pattern), weak (metric), strong (scalar), and strict (residual) invariance (Kline, 2016; Meredith, 1993; Meredith & Teresi, 2006; Putnick & Bornstein, 2016). Configural invariance is the least restrictive level, and it means that the basic organization of the constructs, i.e., the pattern of factor loadings is identical across groups, e.g., countries. Weak factorial invariance builds upon configural invariance and means that the factor loadings are equivalent across groups. Strong factorial invariance means in addition that factor means are equivalent across groups. At the top of the hierarchy is strict factorial invariance, and it means that, in addition, the variances of the residuals are equivalent across groups.

The definitions of the level of measurement invariance need to also consider the type of indicators of the construct. Continuous indicators can be described as having means, variances, and covariances with other variables. Kline (2016) reviewed the consequences of the levels of invariance concerning the comparability of interval scale scores. If factor scores are calculated assuming only configural invariance, a different weighting scheme would be needed for each group. If weak invariance is established, factor scores would be calculated using the same weighting scheme in all groups. Strong invariance is the minimal level required for meaningful interpretation of group mean comparisons. Strict invariance means that the indicators measure the same factors in each group with the same degrees of precision.

A general definition of measurement invariance for an ordinal indicator, such as a Likert-scale item, means that the probability of selecting a particular response option is the same across groups, given the same loading on the corresponding factor. Measurement invariance is established if this property holds for all items (Kline, 2016; Millsap, 2011). Millsap and Yun-Tein (2004) identified thresholds, factor loadings, intercepts, and unique variances as parameters for testing measurement invariance. Group differences in factor variances or means do not affect the measurement properties of ordinal indicators (Kline, 2016). H. Wu and Estabrook (2016) suggested that the invariance of some parameters cannot be tested alone, e.g., the invariance of loadings cannot be tested in the absence of threshold invariance.

Differential item functioning

The concept of measurement invariance, or *lack of item bias*, was originally formulated in the context of IRT (Meredith & Teresi, 2006). On the connection between the concepts of measurement invariance – applying the strict terminology, factorial invariance – and lack of item bias, Meredith and Teresi (2006) noted that the two concepts are distinct but closely related. The lack of bias is defined by Meredith and Teresi (2006) as the invariance of the item parameters across groups of interest. When this invariance is not fulfilled, *differential item functioning* (DIF) occurs.

Issues with the traditional DIF analysis have been discussed extensively in the literature (see e.g., Bechger & Maris, 2015; Cuellar, 2022; Cuellar et al., 2021; Doebler, 2019; Yuan et al., 2021). In his recent dissertation, Cuellar (2022) discussed DIF methods in relation to ILSAs. He outlined the common practice to investigate DIF as: (1) a separate IRT model is fit for each group; (2) the estimated difficulties are compared; (2) the estimated abilities are used as matching criteria across groups. As Cuellar (2022) pointed out, this strategy is aimed to separate differences in abilities from differences in item parameters.

According to Cuellar (2022), there is a paradox in DIF analysis, which is twofold. Firstly, regarding ability, DIF implies a different performance of test-takers on an item given the same ability. The same ability means equal performance on the test as a whole *and* the non-DIF items combined. However, as he argued, if the respondents of one group have worse performance on the DIF item, they must have somewhat better performance on the others. Moreover, the difficulty of an item between two groups is not directly comparable because the difficulty also reflects the level of proficiency of the respondents. This phenomenon is known as the *identification problem* of IRT parameters (San Martín, 2016).

Secondly, regarding item parameters, Cuellar (2022) argued that if there are differences in the estimated item parameters between groups, the estimated abilities are not directly comparable, hence the matching variable is flawed. There are techniques, such as purification, anchoring, and refinement, which are aimed at obtaining a DIF-free test for estimating the abilities by removing some DIF items in a preliminary round of analysis. However, according to Cuellar (2022), there is a circularity problem with the reasoning in these procedures.

To handle the above-outlined issues in DIF analysis, Cuellar (2022) proposed a data visualization technique to explore the structure of DIF between two groups. His approach combined the differential item-pair functioning approach (Bechger & Maris, 2015) and the analysis of the structure of DIF (Doebler, 2019), i.e., patterns of measurement non-invariance.

Linking scales

The previous sections in this chapter concerned the cross-sectional aspect of international large-scale assessments. The second-phase IEA studies are also designed to maintain a system-level *trend measurement* of cognitive outcomes, and more recently, certain contextual or background scales as well. For this purpose, the surveys are *linked* to ensure that the scores are on the same scale. In this section, the framework for connecting cross-sectional measurement points is outlined.

The construction of trend scales in ILSAs is based on the adjustment of two or more tests. *Linking* is the process of making statistical adjustments to scores on tests that are different in content and/or difficulty, using the terminology of Holland and Dorans (2006), Linn (1993), and Mislevy (1992). Mazzeo and von Davier (2013) defined *linking scales* as the process of achieving a scale of results produced by a sequence of assessments, which maintains a stable, comparable meaning over time.

Linking designs

Linking two or more assessments in the IRT framework involves placing the item parameter estimates from the different test administrations on a common scale. The four linking designs defined by Hambleton et al. (1991) are single-group designs, equivalent-groups designs, common-persons designs, and anchor test designs. The authors discussed these designs from a practical point of view as follows.

Single-group designs

The same group of students completes the tests to be linked. This design might be impractical because the testing time would be too long for the individuals, hence, fatigue can affect the results. In addition, the practice effect caused by completing multiple tests may hinder the accuracy of parameter estimation.

Equivalent-groups designs

The tests to be linked are administered to equivalent but not identical, randomly selected groups of students. This design is more practical than the first one because of avoiding fatigue and practice effects.

Common-persons designs

There is a common group of students across the different groups who complete the tests to be linked. This design poses similar challenges to the common group of students as the single-group design; therefore, it is considered less practical.

Anchor-test designs

Different groups of students complete the tests to be linked. Each test includes a set of common items, which set is referred to as the anchor test. This design is very popular because of its feasibility, and it solves the previously mentioned shortcomings of the single-group, equivalent-group, or common-persons designs.

In the context of IEA ILSAs, the achievement tests have been administered to different groups of students at different time points and the assessments have maintained a set of common items between consecutive administrations. Within this anchor-test design, there are several possible methods in terms of parameter estimation to employ for linking the assessments.

Linking methods with anchor-test design

The parameter estimation based on the response data from the different tests may be done together or separately. The procedure during which the response data from the different test forms are pooled together for the parameter estimation is referred to as *concurrent calibration*. This method provides smaller standard errors, involves fewer assumptions than other IRT procedures, and good linking may be achieved with few common items (Wingersky & Lord, 1984).

In the case of separate parameter estimation, one way to link the tests is to fix the item parameters for the anchor items to those estimated on one of the tests when calibrating the items on the other tests. This process is referred to as *fixed-parameter calibration* by Kolen and Brennan (2014). Alternatively, the parameters can be estimated separately also for the common items and a *scale transformation* can be applied so that the distributions of the common item parameters match.

Some characteristics of the best equating practices that were described by Dorans et al. (2011) apply to linking scales. First, since the amount of collected data has a substantial effect on the utility of the results, large representative samples are favorable. Such data ensure that the statistical uncertainty (standard error) becomes much smaller than other sources of variation in the results. Second, the behavior of the anchor items needs to be evaluated across administrations. The assumption is that the item behavior is similar across student groups of the same grades, and it is tested via DIF methods. Finally, the performance on the anchor test needs to be highly correlated with the total test score.

Linking scales in TIMSS

The scaling and linking procedures carried out in the TIMSS assessments are thoroughly described in the various technical reports produced by the IEA (see e.g.,

Martin et al., 2020; Martin & Kelly, 1996b). The following is a summary of the procedures based on the type of construct.

Achievement

The TIMSS scaling procedure involves three steps: item parameter estimation, person scoring, and population modeling. The first step is performed by IRT concurrent calibration i.e., the item parameters for each TIMSS cycle are estimated based on the data from both the current and the previous assessment. In the first step, a 3PL model is used for multiple-choice items, a 2PL model for free-response items with two response options, and a generalized partial credit model (GPCM; Muraki, 1992) is used with polytomous free-response items, i.e., those with a maximum score point of more than one.

In the early cycles, the scaling was carried out on samples that were composed of equal-sized randomly selected subsamples from each of the participating countries (see e.g., Adams et al., 1996). More recently, the whole student samples have instead been used and weighted so that each country contributes equally to the item calibration (see e.g., Foy et al., 2020). The latter approach means that more information is used for item calibration, which ensures more precise parameter estimates but also is more computationally demanding.

The person scoring procedure involves drawing five plausible values from the IRT model-based distribution of scale scores for each test-taker. To enhance the reliability of the estimates of student scores, the third step in the TIMSS scaling is a process known as population modeling or conditioning. This means that all available background data are included in a principal component analysis, and components representing 90% of the variance in the data are included in a regression model for each educational system.

After the scaling is complete in each TIMSS assessment, the newly generated scale is linked to the TIMSS reporting metric. This is achieved in two steps. The first step is a linear transformation of the latent ability distribution of the new assessment data to match the distribution of that in the previous cycle. To match the distributions, the data from the common countries, i.e., trend countries across the two cycles are used. These trend countries contribute equally to the calculation of the transformation constants. The second step is to apply this linear transformation to the whole of the newly gathered assessment data. With these steps completed, the transformed PVs are placed on the reporting scale (see e.g., Foy et al., 2020).

Motivation

The TIMSS studies conducted between 1995 and 2007 combined information to form an index for each background questionnaire scale. The most recent TIMSS context

questionnaire scales (from 2011 onwards) are constructed with IRT scaling using the Rasch partial credit model (PCM; Martin et al., 2012, 2016; Masters, 1982; Yin & Fishbein, 2020).

The longitudinal scaling for the affective constructs started later in the history of IEA ILSAs, first appearing by linking TIMSS 2011 with TIMSS 2015 (Martin et al., 2016). Currently, certain context questionnaire scales that have maintained many of the same items across TIMSS cycles in 2011, 2015, and 2019, are linked through a two-step transformation process.

The first transformation places the TIMSS 2019 logit scale scores on the TIMSS 2015 metric by applying the mean/sigma method (Kolen & Brennan, 2014; Marco, 1977) to the two sets of common item parameters. These scales are estimated by the separate calibration of the TIMSS 2019 data and the TIMSS 2015 data. The mean and standard deviation of the estimates of the threshold parameters, i.e., the difference between item location and item step parameters, are used for all common items and all categories for each calibration. The second step is to transform the TIMSS 2015 logit scores to the TIMSS scale reporting metric (mean 10, standard deviation 2).

To change or not to change the measure

When the interest is in measuring trends, it is useful to consider the renowned maxim introduced by Beaton (1990) "to measure change do not change the measure" (p. 10). Beaton, in the report of the analysis of the so-called *1985-86 reading anomaly* in NAEP (Beaton & Zwick, 1990), highlighted the tension between maintaining continuity and introducing new measurement technologies or curriculum concepts when measuring educational trends.

The results in the average reading proficiency estimates for 1986 indicated sharp declines at ages nine and 17 from the 1984 estimates but a slight rise at age 13. Despite the intention of keeping the assessments as similar as possible in the early years of NAEP, Beaton (1990) pointed out that changes have occurred in measurement instruments by e.g., rearranging or reformatting assessment tasks. As mentioned earlier, IRT was introduced to the NAEP testing protocol in 1983. According to the assumption that item characteristics are independent of context, many of the common items, which were kept from the previous assessment, were placed in a different context in the 1986 assessments. Contrary to expectations, item context showed a substantial effect on the behavior of the common items, largely contributing to anomalous results (Beaton, 1990).

Beaton (1990) also stated that the "precise implementation of this dictum is, of course, impossible in actual practice" (p.10). In the fifth chapter of this thesis, the changes that have occurred across the IEA studies on mathematics and science are

summarized. It is important to point out that even if the test or questionnaire items have been unchanged over time, it may be asked whether, for instance, the measured constructs carry the same meaning over several decades. Student populations change over time and several other factors outside the control of survey administrators can influence comparability.

For example, in the Evaluation Through Follow-up (Utvärdering Genom Uppföljning) project in Sweden, the same cognitive tests have been administered to thirteen-year-old students. The project is based on large and nationally representative samples. Svensson (2008), when analyzing the long-term trends between 1961 and 2005, decided to only include results from students, who attended the normal grade level according to their age, i.e., sixth grade at the time of testing. Svensson pointed out that this decision affected the samples to a varying degree across time points, thus limiting the generalizability of the results. However, it improved the comparability of the scores.

The present investigation intended to resolve the tension between imperfect conditions for linking the first- and second-phase IEA studies and the interest in long-term educational trends. In the next chapter, the methodological choices to tackle the challenges imposed by changes across measurement points are outlined.

Chapter 4 Methodology

This chapter presents the methodological choices and ethical considerations of the empirical studies. An overview of the samples, constructs, and items is followed by a summary of the methods applied to link student outcomes, which are presented in detail in the empirical studies. Finally, validity issues are discussed. The following suggestions by the statistician George Box (1976) guided the methodological considerations throughout this thesis:

> Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. […] Since all models are wrong the scientist must be alert to what is importantly wrong. (p. 792)

Data sources

The empirical work in this thesis is based on data of the populations representing 13-year-olds (FIMS and SIMS), 14-year-olds (FISS and SISS), and eighth-grade students (TIMSS cycles). The data of the first-phase studies were processed differently compared to data from studies conducted later. This thesis benefitted from the work that has been done in the project titled *Center for Comparative Analysis of Educational Achievement* (COMPEAT). This project improved the conditions for secondary analysis by making the data and documentation from the early studies available online in updated formats[1]. Data and documentation for the TIMSS administrations were downloaded from the IEA Study Data Repository[2].

Ethical considerations

The principles of research integrity are of great priority for the present dissertation. These principles are defined by the European Science Foundation (2017) as follows:

> Reliability in ensuring the quality of research, reflected in the design, the methodology, the analysis and the use of resources. Honesty in developing, undertaking, reviewing, reporting and communicating research in a transparent, fair, full and unbiased way. Respect for colleagues, research participants, society,

[1] https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat

[2] https://www.iea.nl/data-tools/repository/timss

> ecosystems, cultural heritage and the environment. Accountability for the research from idea to publication, for its management and organization, for training, supervision and mentoring, and for its wider impacts. (p. 4)

In terms of procedures, previously collected data were used from the assessments of IEA in mathematics and science. Therefore, the source data have been already anonymized, and made free and publicly available. Consequently, authorization for this dissertation was not requested from the Central Ethical Review Board (see legislation in Swedish Research Council, 2017). Using the information on restricted items was permitted by the IEA. The cognitive scales established in Study II alongside the documentation of the linking are publicly available on the COMPEAT repository of this linking project[3].

Samples

The constituent studies in this thesis drew on data representing a large number of students. Study I focused on the four educational systems that participated in all mathematics studies: England, Israel, Japan, and the USA (for a complete account of the sample sizes see Majoros et al., 2021, Table 1). Study II used all data of the selected populations from the twelve countries that participated in FIMS, 20 in SIMS, 17 in FISS, and 23 in SISS. Detailed sample information is available in the documentation of Study II at the COMPEAT repository. Study III analyzed six educational systems that have participated in SIMS and all iterations of TIMSS between 1995 and 2015: England, Hong Kong, Hungary, Israel, Japan, and the United States (for a complete account of the sample sizes see Majoros et al., 2022, Table 1).

Constructs

The empirical studies focused on two types of constructs, cognitive, i.e., mathematics and science achievement, and affective, i.e., motivation for learning mathematics. Studies I and II were concerned with linking the mathematics and science cognitive tests.

The achievement test design in the IEA studies on mathematics and science is organized based on an analysis of the national curricula of the participating education systems. The curriculum frameworks of the studies include a system of categories, i.e., *content areas* and *cognitive domains*. The items comprising the tests represent specific content areas, e.g., algebra within mathematics or physics within science, and cognitive domains, e.g., reasoning. The proportions of items per content domain in

[3] https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/linking-projects/mathematics-and-science

each administration are shown in Study II (Majoros, 2022, Appendix A). The cognitive domains "define the sets of behaviors expected of students as they engage with the mathematics and science content" (Martin & Mullis, 2004, p. 8). The proportions of items per cognitive domain at each time point are shown in Study II (Majoros, 2022, Appendix A).

Study III explored linking affective, i.e., intrinsic- and extrinsic motivational items. In this thesis, motivation is distinguished by its source following the model proposed by Eccles and Wigfield (2002). When the source is external, i.e., individuals engage in an activity for instrumental reasons, e.g., receiving a reward, they are extrinsically motivated. When the source is internal, i.e., individuals engage because they enjoy the activity itself, they are intrinsically motivated. The constructs measured in the student questionnaires are shown in Table A1.

The SIMS items were selected from the attitude scales of *mathematics and myself*, *mathematics anxiety*, and *mathematics and society* (Kifer & Robitaille, 1989). The construct of *attitude towards mathematics* has been assessed throughout all TIMSS administrations. Since 2003, the IEA reports distinguish between the *enjoyment of* learning the subjects and *valuing* the subjects. These correspond to the constructs of intrinsic and extrinsic motivation, respectively. The student questionnaires of the TIMSS assessments are publicly available.

Mathematics and science achievement items

In the IEA ILSAs, the achievement tests have maintained a set of common items between consecutive administrations. The sets of common items serve as anchor tests between assessments. These items are referred to as *bridge* items in this thesis. These items are kept secure and permission for their restricted use needs to be sought from the IEA. The rest of the items in the achievement test are released to the public.

Several common items were identified bridging the first- and second-phase mathematics assessments in Studies I and II (for a complete overview of the number of unique and bridge items, see Majoros et al., 2021, Figure 1 and Majoros, 2022, Table 1.) The bridges across science test administrations were identified in Study II (for a complete account of the number of items, see Majoros, 2022, Table 2).

As for administering the test items, the procedures changed somewhat over time. In FIMS, students received three booklets each, containing 70 mathematics items in total (Thorndike, 1967). In SIMS, five different tests were produced. A matrix sampling scheme was applied with a core test and four rotated tests (Schmidt et al., 1992). The core test was the same for all students and comprised 40 items, administered alongside one of the other four tests. The other four tests involved 34 tasks each and were constructed through stratified randomization of the remaining items. In FISS, two test booklets consisting of 40 science items each were

administered to all students (Comber & Keeves, 1973). The matrix item sampling design in SISS involved a core test and two rotated test booklets. A total of 50 items were presented per student (Postlethwaite & Wiley, 1992).

In the TIMSS assessments, a version of matrix sampling has been applied which is referred to as the balanced incomplete block (BIB; Mazzeo et al., 2006) design. This design means that the entire assessment pool of mathematics and science items at each grade level is divided into blocks of items and the blocks are distributed into sets of booklets, with each student completing just one booklet. Booklets are randomly distributed to students within each sampled classroom. The sequence of the blocks within each booklet is balanced to account for potential position effects. The number of booklets has varied slightly over the years. Each item appears in two booklets, providing a mechanism for linking together the student responses across the population sample (see e.g., Martin et al., 2020).

All mathematics and science cognitive items that were identified as overlapping across assessments were found to be identical. The common affective items reviewed in the following section represent a variety of identical and similar items.

Mathematics motivation items

Beyond measuring cognitive achievement, IEA ILSAs employ contextual, or background questionnaires administered to school principals, teachers, and students to gather information on areas such as student attitude, or teacher instructional quality (see e.g., Yin & Fishbein, 2020). IEA studies have long assessed affective constructs, such as students' readiness to learn, enjoyment, confidence, and value of mathematics. The student motivational measures in the focus of this thesis appear in the questionnaires as Likert-type items.

The items explored in Study III correspond to indicators of intrinsic and extrinsic motivation toward mathematics (see Majoros et al., 2022, Appendices A and B). An example statement from the intrinsic motivation scales in SIMS is "I think mathematics is fun". An example statement from the same assessment related to extrinsic motivation is "It is important to know mathematics in order to get a good job".

Over time, there have been unique, identical, and similar motivational items across assessments. More specifically, the wording of the items has changed in some cases. In a few instances, this has also meant shifting from positively worded to negatively worded statements or the other way around (for the exact wording see Majoros et al., 2022, Tables 2 and 3).

The students have had four response options to choose from for all motivational items in all TIMSS cycles: *strongly agree*, *agree*, *disagree*, and *strongly disagree* (the wording refers to TIMSS 1995). In SIMS, students were presented with an additional middle

option: *undecided*. In FIMS, the attitude items only had three response options to express agreement, disagreement, or uncertainty about the declarative statements in the questionnaire.

Due to a large number of middle responses in SIMS (see Majoros et al., 2022, Tables 5 and 6), a decision was made to recode these responses to random answers between the options agree and disagree. There were some cases when a student selected the middle option for all items: 0.95% for the extrinsic and 0.71% for the intrinsic motivation scale. These cases were excluded from the analyses.

The number of overlapping items in the intrinsic motivation scales is presented in Table 3. Numbers in the same row represent common items between surveys. The number of items common with the preceding administration is shown in the *bridge* row. The total number of items in the item pool per administration is shown in the *total* row. The first bridge only consisted of two out of 11 SIMS items and five TIMSS 1995 items, but the ratio of bridge items has improved over time. Overall, the pooled intrinsic motivation scale comprised an item pool of 19 questions.

Table 3 Number of common and unique items in the intrinsic motivation scales

	SIMS	T95	T99	T03	T07	T11	T15
	11	2	2	3	2	3	3
		3	3		2	2	2
						1	1
							4
Bridge		2	5	2	2	4	6
Total	11	5	5	3	4	6	10

Note. This table shows item overlaps across the assessments over time. The TIMSS assessments are denoted by *T* and the last two digits of the year of the assessment cycle. Numbers in the same row represent common items between surveys. The number of items common with the preceding administration is shown in the *bridge* row. The total number of items in the item pool per administration is shown in the *total* row.

The number of overlapping items in the extrinsic motivation scales is shown in Table 4, in which the same logic applies as in Table 3. We can notice that similar or identical items have been maintained over time. For instance, three items out of the item pool of eight in the SIMS questionnaire were common in TIMSS 2015. Overall, the pooled extrinsic motivation scale consisted of 15 items.

Table 4 Number of common and unique items in the extrinsic motivation scales

	SIMS	T95	T99	T03	T07	T11	T15
	8	3	3	2	2	2	3
		4	4	1	1	2	2
				1	1	1	1
							2
Bridge		3	7	3	4	4	5
Total	8	7	7	4	4	5	8

Note. This table shows item overlaps across the assessments over time. The same logic applies as in Table 3.

Analytical methods

The present thesis was guided by two overarching research questions concerning the comparability of the outcomes and the impact of applying different linking approaches on the trend descriptions. Arguably, comparing surveys that were administered decades apart is challenging, and involving multiple countries "is in several aspects an exercise in comparing the incomparable" (Husén, 1983, p. 455). In this section, the analytical methods for addressing these questions and challenges are outlined.

Evaluating the comparability of the outcomes

This section is a brief overview of the methods addressing longitudinal and cross-sectional comparability that were applied in Studies I-III. The substantive basis of the three empirical studies lies in the evaluation of the extent of similarity across survey administrations. The degree of similarity across assessments to be linked determines the "utility and reasonableness" (Kolen & Brennan, 2014, p. 498) of linking. Kolen and Brennan (2014) proposed four criteria for evaluating similarity: inferences, populations, constructs, and measurement characteristics. Thus first, the goals need to be evaluated concerning the types of inferences drawn from the tests to be linked. Second, the alignment between the target populations of the assessments to be linked needs to be scrutinized. Third, the similarity of the measured constructs is to be evaluated. Finally, the measurement conditions, such as test length, test format, and administration need to be scrutinized.

After the substantive analysis of the similarity of the surveys, the comparability of bridge items over time was addressed. More specifically, bridge item behavior was investigated across assessments with the delta plot method (Angoff & Ford, 1973). The delta plot is a method to identify DIF among dichotomously scored items (Magis & Facon, 2014). The (transformed) proportion of correct answers (test items) or

responses indicating positive endorsement (questionnaire items) is compared between the reference group and the focal group. In this thesis, these groups refer to test-takers in two different administrations. Each item is represented as a point in the plot. The graphical representation also includes a regression line as a point of reference. If there is no DIF, these proportions should be located on this line. Items that are separated from the diagonal are flagged as DIF items. Following the suggestion by Magis and Facon (2014), the threshold for detection of DIF was derived by using a normality assumption on the delta points. The delta plot method for detecting DIF works under the CTT framework. This method in this thesis has been chosen for two main reasons. Firstly, the delta plot is a not computationally intensive method. Secondly, this is a relative DIF method, i.e., the items were evaluated in relation to all items comprising the bridge, which is one way to handle issues with traditional DIF analysis that are outlined in Chapter 3.

Lastly, the cross-cultural measurement invariance of the constructs was considered. The cognitive constructs were assumed to be invariant based on the numerous quality assurance processes applied in the assessments. The cross-cultural comparability of affective constructs was evaluated by applying MGCFA for each time point. The CFA approach was chosen based on the suggestion by Meade and Lautenschlager (2004) that CFA is theoretically preferable over IRT methods when the number of items is small. The questionnaire items were treated as categorical variables and the students were grouped by country. The first step was to identify the baseline model and test for configural invariance among countries. After establishing configural invariance, threshold invariance was tested, followed by invariance testing for factor loadings (Svetina et al., 2020; H. Wu & Estabrook, 2016).

Model fit was evaluated by absolute and relative fit indices. χ^2, the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) values served as absolute model fit indices. The relative fit was indicated by the comparative fit index (CFI) and the Tucker-Lewis index (TLI). When evaluating these results, some caution needed to be taken. First, the χ^2 values are sample size sensitive (Brown, 2015). Second, the presence of negatively worded items potentially causes one-dimensional CFA models to show a poor fit (see e.g., Marsh, 1996; Steinmann et al., 2021; Woods, 2006; Zhang et al., 2016). Finally, model fit values are influenced by many factors, such as estimation method or categorical/continuous specification and Shi and Maydeu-Olivares (2020) suggested using only the SRMR because it is more consistent across these factors.

Linking approaches

Several approaches for linking were explored in the constituent empirical studies. In the case of the cognitive scales, the empirical work started with substantially and

empirically evaluating the feasibility of linking the assessments. Hence, the first study may be viewed as a preparatory study for Study II. Study II placed the mathematics and science assessments on the TIMSS reporting scale. Study III is an exploratory investigation of linking affective scales, therefore, several different approaches were compared.

Study I

Using the pooled data of four countries that participated at each time point from FIMS to TIMSS 2015, the linking procedure involved three main steps. First, different IRT models were tested to select the best fitting one. The chosen IRT models were 2PL for dichotomous items, i.e., multiple-choice items and constructed-response items for one score point, and the GPCM for polytomous items, i.e., constructed-response items for two or more score points.

Second, the item parameters were estimated via concurrent calibration. Thus, the item parameters were on the same IRT scale using data from all studies and four countries. Senate weights that sum to 500 for each country's student data were applied (stratum weights in SIMS were rescaled to sum to 500), thus, each country contributed equally to the item calibration. There were no weight variables in the FIMS 1964 datasets; therefore, individuals within a country were weighted equally, to sum up to 500.

The third step was the person scoring with the use of item parameter estimates. Five PVs were drawn per student using the expected a-posteriori method. The estimated abilities were converted to scale scores; thus, each PV was transformed to a metric with a mean of 500 and a standard deviation of 100 points across time. The transformed scores were used to compute the mean mathematics achievement for the respective country per study following Rubin's (1987) rule of pooling.

Study II

This study compared two linking approaches for the mathematics scale. Firstly, the *four-country-all-time* (points) approach took advantage of the item parameters estimated by the method in Study I. The procedure started with the separate ability estimation for FIMS and SIMS, fixing the item parameters to the estimated values via the method in Study I. Then the distribution of the five PVs estimated for FIMS and SIMS was matched with the distribution of the reported TIMSS 1995 PVs. This was done by calculating transformation constants, similar to the TIMSS scale linking procedure, in two steps: (1) The means and standard deviations of the reported 1995 PVs, which are on the required scale, were matched with the means and standard deviations of the 1980 PVs for the same set of countries and years of schooling within countries, which are on an independent scale. Then all countries were put on this scale using the

same transformation constants. (2) The same procedure was done to match the distribution of the FIMS PVs with the SIMS PVs.

Secondly, the *first-second-time* approach involved the concurrent calibration of item parameters using the first and second ILSAs on mathematics with the bridge item parameters fixed to the values reported for TIMSS 1995. These item parameters were reported after a rescaling procedure in the 1999 assessment cycle (Martin et al., 2000). Then, similarly to the scaling procedure in the first step, the ability distribution of SIMS was matched with the reported TIMSS 1995 scale and then SIMS with FIMS.

When constructing the science achievement scale, the first-second-time approach was chosen for several reasons. First, the IRT models are the same as those used in the TIMSS procedures, i.e., the 2PL, 3PL, and GPCM. Second, this procedure is more economic although data from more countries (but fewer time points) were used. Comparing the amount of information, i.e., the number of item responses used for item calibration in the two approaches, on the one hand, the four-country-all-time concurrent calibration involves 893 items (1964-2015), while the first-second-time approach uses the items between 1964 and 1995, i.e., 373 items. On the other hand, the weighted number of item responses used for the link between SIMS and TIMSS 1995 is close to threefold in the first-second-time approach than those in the four-country-all-time method due to the larger number of countries (42) involved in the procedure.

Study III

Three methods were explored for constructing longitudinal affective scales: an IRT, a SEM, and a market-basket approach. Firstly, in the IRT approach, the GPCM model was found to fit the data best. Then the item parameter estimation was conducted by concurrent calibration of all items in all studies, thus the parameters for all tests are automatically put onto the same scale. The parameters of the anchor items were assumed identical in each sample. Third, person scores were estimated and transformed onto a scale with a mean of five and a standard deviation of one.

Secondly, a CFA model was fit for each motivation scale on a pooled sample composed of data from all countries and cycles. Strong invariance of the anchor items across countries and over time was assumed. Factor scores were estimated by applying maximum likelihood estimation with robust standard errors (MLR), while the items were treated as categorical variables. The factor scores then were transformed onto a scale with a mean of five and a standard deviation of one.

Finally, a market-basket approach was applied. The market-basket approach assumes that the items included in the assessment or survey define the construct. In this case, the assumption is that all the items across the time points, related to intrinsic and extrinsic motivation towards mathematics, define each construct and can be

considered as a market basket of representative items. The missing responses occur as a consequence of changes in the questionnaires across cycles.

A measurement model per country was employed to generate plausible responses that fill the missing responses following the procedure suggested by Zwitser et al. (2017). The measurement model was fit for each country separately to account for potential differences among countries. The measurement model was the GPCM model for consistency with the results from the IRT approach and the TIMSS procedure for linking contextual scales.

Using the item parameters estimated by fitting the measurement models, missing responses were imputed five times per respondent. Then individual sum scores were calculated, thereby estimating five plausible scores per student. Finally, the plausible scores were transformed onto a scale with a mean of five and a standard deviation of one.

Handling missing data

Given the previously described matrix sampling of items in the TIMSS achievement tests, most item responses are missing by design for each student. In this case, the missing data is considered *not administered*. However, missing responses may also result from a test-taker not answering an item. If the missing response is located before the end of a given test booklet, it is classified as *omitted*. In the TIMSS student achievement tests, if the omitted item is in a position close to the end of a test booklet, the missing response is coded as *not reached*. More precisely, an item is considered not reached when itself and the item immediately preceding it are not answered, and all items are incomplete in the remainder of that part of the booklet.

Not-administrated items were treated as missing responses, while omitted responses were treated as incorrect answers both when estimating item parameters and scoring. The not-reached items were treated as if they were not administrated i.e., missing for item calibration to avoid comparability issues between the assessment cycles (see Gustafsson & Rosén, 2006). In contrast, not-reached items were treated as incorrect responses when student proficiency scores were generated. This approach is in line with the procedures for handling missing responses in the TIMSS studies. In the datasets of the first-phase studies, the various types of missing data were not distinguished.

There are several approaches to handling missing data in SEM applications. The most widely preferred methods are maximum likelihood estimation and multiple imputations (Brown, 2015). These approaches make use of all the available data in contrast to e.g., listwise deletion. Under the methods in the SEM framework in this thesis, the full information maximum likelihood (FIML) estimation was applied,

which uses all available information from all observations to handle the missing data. For the responses missing by design, we applied a pattern function in Mplus 8.

Validity and validation

A situative perspective on validity, as proposed by (Mislevy, 2017) is applied to the context of the present thesis. The *socio-cognitive perspective* regards human thinking, acting, and learning in the physical and social world via socially constructed practices in contrast to perspectives focusing on individuals, in which the environment functions as merely the context to behavior and cognition (Greeno, 1998; Mislevy, 2017).

According to Newton and Shaw (2014), the history of validity theory and validation in educational and psychological measurement is characterized by richness and diversity. Since the first edition in 1954 of the "Standards of Educational and Psychological Testing" (American Educational Research Association et al., 2014), the validity chapter has been evolving but no consensus has been reached. In the present thesis, construct validity is regarded as a unified form of validity.

The *construct model of validity* was first proposed by Cronbach and Meehl (1955). Messick (1989) further elaborated on the concept of construct validity and extended the boundaries of validity beyond the meaning of test scores to include relevance and utility, values, and social consequences. After 2000, two distinct debates developed about the nature and the scope of modern validity theory. The debate on the nature of validity concerns the issue of whether validity should be a narrow scientific concept of score meaning. The debate on the scope of validity refers to whether the concept of validity should be expanded beyond construct validity and incorporate consequences. Overall, throughout the decades of debate on the characteristics of quality in educational and psychological measurement, it has been difficult to find an agreement about the need for a unified definition of validity.

Mislevy (2017) regarded the shift of focus from validity as a property of the tests to a property of the inferences from test scores as a consensus view. He extended Kelley's (1927) definition as "a test is valid for a given interpretation or use to the degree to which empirical evidence and theoretical rationales support reasoning *as if* 'it measures what it is purposed to measure"' (p.202). This definition focuses on situated interpretations and is regarded as constructive-realistic (Messick, 1989).

Validation involves the actions taken to evaluate the soundness of inferences based on different forms of assessment. Mislevy (2017) built the socio-cognitive perspective on the six aspects of validity proposed by Messick (1998): content, substantive, structural, generalizability, external, and consequential aspects. Based on these aspects, the validity issues of the present thesis are discussed in the following.

As secondary data analysis, the present research relies on the quality procedures and methodologies applied by the IEA. Therefore, the content, substantive, and structural aspects of validity are described here but serious threats against these validity aspects are not suggested. The content aspect of validity refers to content relevance, representativeness, and technical quality. Hence, the content line of validation evaluates the rationale for task situations. Similar to the TIMSS scale linking procedure, all items in each subject were calibrated together in this thesis. Treating the entire mathematics or science item pool as a single domain maximizes the amount of data in terms of content representativeness and item responses.

The substantive aspect of validity concerns the theoretical rationales for the observed consistencies in test responses. The substantive validation refers to evidence about whether the kinds of perception, cognition, and actions that comprise the construct are evoked by the students who are assessed. The structural aspect of validity refers to the adherence to the scoring and the construct domain structures. Therefore, the structural validation regards whether the scoring procedures are consistent with the intended constructs and whether the response patterns reflect patterns that tend to emerge at different levels of proficiency.

The generalizability strand of validity covers many aspects of score interpretation and score use. It concerns the extent to which the information of a particular assessment holds for inferences about other times, population groupings, cultures, settings, and task- and criterion situations. This aspect is central in the present thesis and evaluated through rigorous inquiry outlined earlier in this chapter.

The external aspect of validity includes evidence of relationships between test scores and other sources of information from the same test-takers. For instance, if several constructs are assessed with the same method and participants, the question to address is whether the scores show expected correlations. The evaluation of the external validity of the scales requires further research to map common constructs, e.g., socioeconomic status measured in the assessments that are linked in this thesis.

Finally, the consequential strand of validity regards the value implications of score interpretation. This strand can be viewed from a socio-cultural and a technical point of view. The socio-cultural aspect of consequential validity concerns whether the assessment practice shapes the lives of individuals and societies. Johansson (2016), in his review of the literature relating to the uses and consequences of ILSAs, found that much research suggests that ILSAs have unintended consequences that affect and influence educational policy. The results of ILSAs are presented in different ways outside the research community, fueling political discussion on for example school reforms (e.g., Braun & Singer, 2019). Johansson (2016) argued that the influences on educational policy are complex and interwoven. He suggested that a beneficial

consequence of ILSAs is the infrastructure they provide for studies in the social sciences for studying global trends and evolving systems in education.

The technical view of the consequential aspect is concerned with the alignment between intended consequences and assessment interpretations and uses. Oliveri et al. (2018) reviewed examples of unintended consequences from the aspect of misalignments between participating educational systems' goals and what testing organizations can provide. Such consequences include "the inappropriate use, misuse, overgeneralization, or lack of knowledge of a test's limitations that leads test users to make score-based inferences that are more ambitious than the testing program can withstand" (p.2). In the present thesis, the identified limitations of linking the first- and second-phase IEA studies are discussed in the next chapter and in the empirical studies to reduce such unintended consequences.

Chapter 5 Results and discussion

This chapter is an integrated summary and discussion of the results of the empirical studies. Limitations to the linking and implications for research are also discussed. The results of the three studies are presented according to the overarching research questions.

Comparability of the outcomes

This section reviews the results of the evaluation to address the first overarching research question: To what extent are the student outcomes comparable across the first- and second-phase IEA assessments on mathematics and science? Taking advantage of the scheme proposed by Kolen and Brennan (2014), four sub-questions are addressed in the following sections: (1) To what extent were the assessments used to draw similar *inferences*? (2) Were the assessments designed for the same target *populations*? (3) To what extent do the tests assess similar *constructs*? and (4) To what extent were the *test conditions* similar?

Inferences

The inferences that can be drawn from the IEA studies on mathematics and science are essentially the same in terms of data, generalization, and explanation. The assessments use a high-level inference approach (Ercikan & Roth, 2006; Gustafsson, 2008) to generate data by abstracting information over contexts and items. The IEA aimed to achieve generalizability to the population level. Finally, even though the purpose of these assessments in terms of explanation has changed, i.e., the focus has shifted from input factors to the output of education (Gustafsson, 2018), the initial explanatory purpose was nevertheless found not to be feasible due to the constraints of the cross-sectional design.

Populations

ILSAs typically define a cohort of the student population as the population of interest for a particular iteration of that survey, referred to as the *target population*. Two common approaches to establishing a target population in current ILSAs are to define the population based on a particular grade in school (e.g., in TIMSS) or by the age of

the student (e.g., in PISA). The populations that were attending secondary school, typically studying in the 7th-10th year of schooling or being 13-14 years old were selected in the present thesis. As Strietholt and Rosén (2016) pointed out, the IEA changed the definition of target populations from an age-based to a grade-based definition in the 1980s for all their studies of student achievement (for the target population definitions see Majoros, 2022, Table 4). Arguably, any sampling strategy changes result in a violation of the assumption of comparable samples.

In the report on the changes in achievement between the FIMS and SIMS studies, Robitaille and Taylor (1989) argued that the populations targeted in FIMS and SIMS should be considered equivalent because all students in the assessed educational systems around the age of 13 would be studying the same levels of mathematics. However, between the first-phase science assessments, as Keeves and Schleicher (1992) pointed out, there occurred some sampling deviations. On the one hand, in SISS, there were two options for the target populations as outlined in Table C.1. On the other hand, it was decided in most countries that intact classrooms were sampled. Australia, England, and Italy did not test intact classrooms.

To tackle the above-outlined changes and improve comparability across the samples, student responses were kept in the analyses based on grade level, i.e., the number of years of schooling. In practice, this meant keeping data from students who were in their 7th to 9th year of schooling for FIMS, and 7th to 10th year of schooling for FISS. After removing cases with missing responses to all items in the final scaling in Study II, 89.95% of the FIMS, 93.62% of the SIMS, 87.71% of the FISS, and practically 100% of the SISS sample were kept.

Constructs

Overall, the changes in the mathematics and science content domains over time are mostly terminological and organizational (see Majoros, 2022, Appendix A). One exception is the science content domain of *Earth science*, which was first introduced in SISS and then assessed throughout all time points. Since 2007, the TIMSS assessment frameworks concerning mathematics and science achievement have been consistent in terms of conceptualizing the content domains. If we compare the relative proportion of content domains, it has been fairly stable but some shift in the focus can be observed. For example, the statistics content of the mathematics test has become greater at the expense of arithmetic. In the science tests, the number of biology items has become more prominent over time, in contrast to the initial equal share of content domains. The relative proportion of the items addressing the different cognitive domains has been fairly consistent over time (see Majoros, 2022, Appendix A).

Concerning the affective constructs, the domains have been fairly consistent throughout the TIMSS iterations (see Table A1). Indicators related to the constructs of extrinsic and intrinsic motivation were identified in Study III. However, the items have changed over time (see Majoros et al., 2022, Appendices A and B). Due to the considerable changes in the affective scales, this part of the present thesis is exploratory.

Test conditions

All assessments included in this thesis were paper-based surveys. The early studies were conducted in somewhat longer testing sessions, due to administering more items per test-taker than in the TIMSS cycles. Another difference is that as mentioned earlier, since TIMSS 1995, the survey consists of both mathematics and science items.

To conclude this section, the substantive evaluation of the comparability across administrations shows that there has been a high level of stability concerning the inferences and measured constructs among the assessments. The changes in the sampling and test conditions introduced challenges to the linking. Some of these challenges were handled to achieve a sufficient degree of similarity across the assessments. The rest remain as limitations of the scales.

Bridge items

In Study I, to test the assumption about the behavior of common items, the delta plot method was applied to all seven bridges between adjacent time points. A total number of three items were flagged for DIF in the first two bridges, i.e., from FIMS to SIMS, and from SIMS to TIMSS 1995, respectively. No DIF items were detected in the rest of the bridges (for the delta plots, see Majoros et al., 2021, Figures 2a and 2b). The items showing DIF were excluded from the calibration. Furthermore, twelve non-anchor items were excluded due to missing answers in all countries. Overall, 893 items were included in the concurrent calibration.

In Study II, the delta plot method was applied for the six sets of bridges. These bridges consist of common items in the mathematics assessments between 1964-1980 (bridge 1), 1964-1995 (bridge 2), and 1980-1995 (bridge 3), and among the science surveys between 1970-1984 (bridge 4), 1970-1995 (bridge 5), and 1984-1995 (bridge 6). Two items in the first, one item in the third, and two items in the fourth bridge were flagged for DIF (for the delta plots, see Majoros, 2022, Appendix B). In the final, first-second-time linking procedure, these items were treated as unique items.

To test the assumption about the performance on the anchor test and the whole test, Pearson's correlations were calculated. These correlations were moderate or high: for FIMS, $r = .97$, $p < .001$ (bridge 1) and $r = .84$, $p < .001$ (bridge 2); for SIMS, $r =$

.88, $p < .001$ (bridge 1) and $r = .66$, $p < .001$ (bridge 3); for FISS $r = .92$, $p < .001$ (bridge 4) and $r = .69$, $p < .001$ (bridge 5); for SISS, $r = .86$, $p < .001$ (bridge 4) and $r = .80$, $p < .001$ (bridge 6).

In Study III, the delta plot method was applied for each bridge between consecutive time points. The tests were conducted for each country separately as well as with the pooled data and all these tests yielded no items flagged for DIF (for the delta plots, see Majoros et al., 2022, Appendices E and F).

Cross-cultural invariance

In Study III, the measurement invariance was tested across countries at each time point. The MGCFA results revealed that in SIMS, measurement invariance did not hold for Japan and all further analyses in this study were continued excluding data from this country. The threshold and loadings equality constraints yielded an acceptable model fit at most time points for the five-country multiple-group model (see Majoros et al., 2022, Appendix D).

Trend descriptions by linking approach

This section presents results from the empirical studies related to the second research question: How do different linking approaches influence the descriptions of the system-level trends? The results are discussed in terms of the measured constructs.

Mathematics achievement

The two approaches applied in Study II yielded very similar results despite the difference in the IRT modeling. This result indicates the robustness of the scales. The mathematics individual plausible scores (five per each test-taker) estimated in the four-country-all-time and the first-second-time approach showed strong correlations for all possible plausible value pairs.

The strong individual-level correlation translated well into the country-level aggregation in almost all cases with the remarkable exception of Japan (see Majoros, 2022, Figure 1). In the four-country-all-time approach, the Japanese mean score was significantly higher than in the first-second-time approach. In contrast, the individual scores correlate strongly, Pearson's r ranges between .92-.93, $p < .001$ in FIMS, while in SIMS $r = .91$, $p < .001$ for all possible plausible value pairs.

There were two main differences in the linking approaches. In the first-second-time approach, a guessing parameter was included in the IRT model for multiple-choice items, and more responses were available for the item calibration than in the four-country-all-time approach. The former suggests a possible cultural difference

concerning guessing, i.e., introducing guessing to the model resulted in a lower country mean for Japan than having no guessing.

Motivation for learning mathematics

As mentioned earlier, the five countries were treated as a single group both cross-sectionally and longitudinally in the CFA and IRT procedures. They were treated separately in the market-basket approach, but data were pooled into a single group model over time. The observed scales were constructed by computing the sum of the scores per person at each time point divided by the number of answered items. Then the standardized scores considering a mean of five and a standard deviation of one were calculated.

The three methods yielded similar results on the individual- as well as the country-levels. The correlations between individual scores were high across methods for both motivation constructs, ranging between 0.96 and 1. It is striking in the country-level trends that both models (CFA and IRT) with assumptions for strong cultural- and longitudinal- invariance resulted in very similar results to the observed scale in the extrinsic motivation scales (see Majoros et al., 2022, Figure 1). The same pattern may be observed in the case of the intrinsic motivation scales (see Majoros et al., 2022, Figure 2).

Nevertheless, we can observe how the two types of motivation constructs have changed over time in the five educational systems. It may be seen that these affective constructs did not change dramatically in Hong Kong over time and that the system-level results indicate that the overall level of intrinsic motivation is consistently higher. Interestingly, the country-level extrinsic motivation shows similar trends in Israel and England, with high motivational levels recently. Hungary's trend lines of the two types of motivation seem to almost mirror each other. The results of the United States are the most stable in this set of countries.

Another interesting finding is that the average intrinsic motivation in all countries is fairly close to the mean in recent years except for Hungary. There seems to be a larger variation on the extrinsic motivation scale. Considering that the comparability of the affective scales between countries is challenging, the market-basket approach could be a potential choice for putting more educational systems on the scales.

Potential applications of the scales

With the newly established scales, it is possible to examine long-term changes comparatively or within countries. System-level changes take time, therefore, evaluating school reforms requires long-term data. Furthermore, as mentioned earlier,

powerful statistical approaches to address causal research questions may be applied to system-level longitudinal data.

For instance, Strietholt et al. (2019) recently reviewed the international comparative literature on the impact of education policies on the socioeconomic achievement gap. The authors found that most of the existing research was descriptive, estimating correlations based on cross-sectional data. Further research into mapping indicators of socioeconomic background in the first-phase IEA surveys combined with the achievement scales could potentially contribute to this line of inquiry.

Another potential area to take advantage of the long-term scales lies in issues related to the global educational reform movement (Fuller & Stevenson, 2019; Sahlberg, 2016). Such related phenomena involve privatization, free school choice, school competition, or teacher education.

Sweden was selected to present an example of national long-term achievement trends established in this thesis. Sweden is one of the countries that have participated in most ILSAs administered by the IEA. The Swedish educational system is affected by large-scale education reforms. The research concerning the consequences of the school choice reform (see e.g., Fjellman et al., 2019; Yang Hansen & Gustafsson, 2016), which started in the early 1990s, may benefit from system-level comparable achievement scores. The weighted country means of mathematics achievement are shown in Figure 5, while those of science are in Figure 6.

Overall, the patterns are similar in these subjects. However, as shown in Figure 8, there was a considerable increase in the average mathematics achievement in seventh grade between 1980 and 1995, much larger than in the same time period for science achievement. Potential explanations include changes in the mathematics curriculum and the assessment frameworks in terms of content domains.

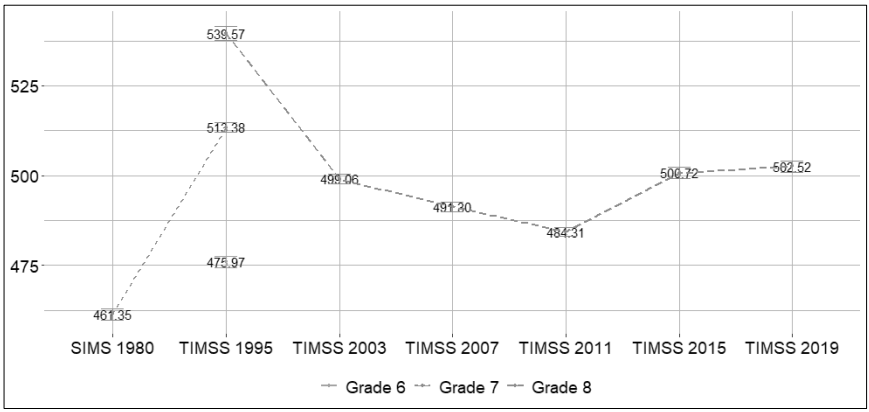


Figure 5 Swedish average mathematics achievement, 1980-2019

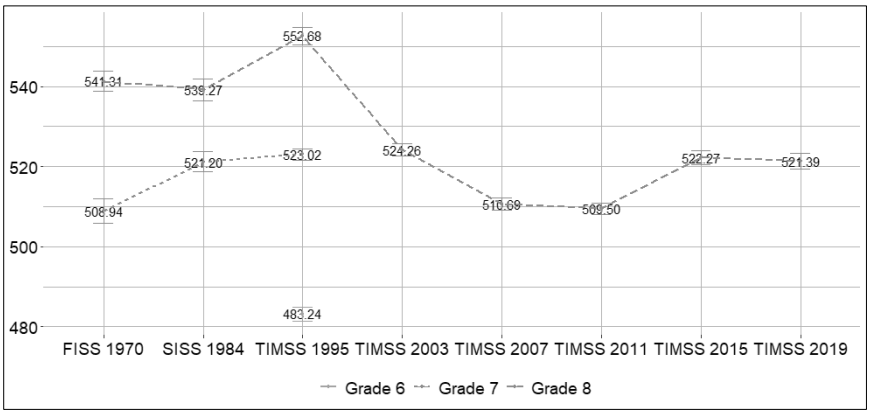


Figure 6 Swedish average science achievement, 1970-2019

Chapter 6 Concluding remarks

With this thesis, by linking the first- and second-phase IEA studies on mathematics and science for grade eight, a challenging endeavor has been taken due to the changes that occurred across administrations. The utility of linking the studies stems from the advanced econometric methods and country-level longitudinal modeling techniques that have already encouraged research involving ILSA outcomes that are on separate scales. The main purpose was to facilitate future country-level longitudinal studies that include the first-phase IEA studies by the means of comparable measures of mathematics and science achievement in eighth grade.

Such studies might shed light on explanations for changes in the educational outcomes of participating countries. However, drawing valid causal inferences from observational data is challenging (see e.g., Allardt, 1990; D. Rutkowski & Delandshere, 2016). Observational data in this context means data that were collected without having control over any treatment or grouping of students, such as ILSA data. Suggestions for advanced statistical methods for causal inferences based on ILSA data have been made by several researchers (see e.g., Gustafsson, 2008; Gustafsson & Nilsen, 2022; Robinson, 2013; Schlotter et al., 2014). Many of these methods were developed in the field of economics and economists frequently rely on ILSA data while monitoring educational systems.

Causal inference and ILSA data

A causal *effect* refers to the estimation of a causal relationship, whereas a causal *mechanism* sheds light on the mechanisms behind it and the conditions under which the causal relationship holds (Shadish et al., 2002). Without an explanation of the causal mechanism behind a causal effect, the utility and the generalizability of the effect are limited (Gustafsson & Nilsen, 2022). On the importance of studying explanations, Gustafsson (2008) argued that "the widely publicized descriptive results call for explanations, and if no explanations are offered by the researchers, explanations will be put forward by other stakeholders, such as the media and politicians" (p. 9).

However, a word of caution might be necessary regarding causal inferences. Despite the great progress that has been made concerning statistical methods for drawing causal inferences from observational data, Gustafsson and Nilsen (2022)

argued that the problem of *endogeneity* is an often-overlooked threat to validity. Endogeneity refers to situations in which the outcomes of an empirical study are influenced by the characteristics and actions of the individuals involved in the study beyond the independent variables.

As one approach, Gustafsson and Nilsen (2022) proposed a simple switch of the roles of what researchers typically see as dependent and independent variables. They used the example of the relationship between student achievement and class size. Although class size is often treated as an independent variable, their results showed that it was systematically related to students' home resources for learning, which in turn was related to achievement. This finding indicates that interpreting relations between student achievement and class size in causal terms is challenging.

Limitations

The final scales for the first-phase studies are publicly available on the COMPEAT website[4] along with the documentation of the scale linking. It should be noted that the sampling differences need to be considered when using the scales. For instance, Strietholt et al. (2013) developed a correction model to improve comparability across countries and IEA studies on reading in terms of age and schooling. Another approach to account for these differences between time and countries is to treat age and grade level as plausible explanatory variables.

Many factors influence the quality of linking. These factors include the degree of similarity across assessments, the stability of the constructs in terms of content, meaning, and context, and the number and behavior of bridge items. This thesis addressed these influences from a substantive as well as a measurement point of view. However, there may be a degree of uncertainty in terms of evaluating these aspects because the linking involves legacy data.

A possible limitation to the longitudinal achievement scales lies in the within-country comparability because of the target populations of the assessments. The shift in the IEA sampling strategy was tackled with as good approximations of homogenous samples as possible. In further analyses using the new scale scores, age and grade level can be treated as control variables.

The number of common items in the bridges connecting to TIMSS 1995 is another concern. The ratio of bridge items in the affective scales is much less than in the case of the cognitive scales. However, the concurrent calibration method provides the best approach to having only a few bridge items, as pointed out by Wingersky and Lord (1984).

---

[4] https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat/linking-projects/mathematics-and-science

Another limitation concerns the coding and treatment of different types of missing data in the achievement tests. In the first-phase studies, the not-reached type of missing responses was not distinguished. Therefore, those missing responses were treated as missing, unlike in the TIMSS scaling procedure. It would be possible to make this distinction and explore the influence on the results.

The affective scales analyzed in Study III until recently had not been designed for trend measurement. Modifications have occurred over time, e.g., the number of response options was changed from 1995. The way of handling the middle option in SIMS posed a limitation to the study. In addition, the linking methods in this study did not account for the differences in the motivation distributions over time since a single-group approach was applied.

Another limitation of the affective trend scales is that the standard errors of the means are underestimated. The reason for this is that because of the stratified multistage sampling design used in TIMSS, the simple random sampling assumed in the procedure for calculating standard errors of estimates did not apply (L. Rutkowski et al., 2010).

Finally, one of the most challenging remaining questions is whether changes in wording affect the internal relationships among motivational items (e.g., factor structure). Since non-identical sets of items were explored over time, the number of items varies at almost every time point, which makes the investigation of the effects of changes in item wording difficult.

Future research

An avenue for future research is to continue the exploration of long-term trends in the attitudes toward learning mathematics and science. There are considerably larger challenges with these outcomes than the achievement scales due to item-level changes. The market-basket approach employed in this thesis offers possibilities for linking with fewer assumptions of comparability than item-level linking.

Another interesting area is to explore the possibilities of linking the first- and second-phase IEA studies on science for the grade four population. Data are publicly available for the old studies. Furthermore, in the 1995 administration of TIMSS, the achievement tests of the younger and older populations were linked through anchor items (Martin & Kelly, 1996a). This particular design feature has rarely been used in previous research, so it would be interesting to revisit this linkage for investigating developmental changes.

Furthermore, it may be useful to map other contextual indicators from the first-phase studies, which could facilitate more complex investigations in, for instance, gender differences. The relative proportion of girls choosing a mathematical track in

upper secondary and higher education or professions related to science, engineering, mathematics, and technology (STEM) is still unreasonably low and unrelated to mathematics achievement in many countries. A growing body of research suggests that in wealthier and more gender-egalitarian countries, gender differences including attitudes and occupational preferences in STEM disciplines are often larger (see e.g., Breda et al., 2020; Stoet & Geary, 2020). These counterintuitive differences are often referred to as gender-equality paradoxes. These paradoxes are regarded to contradict the assumption of social role theory that the pressure on divergent social roles should be lowest in countries with more egalitarian gender roles, gender socialization, and socio-political gender equity, thereby decreasing psychological gender differences (Mac Giolla & Kajonius, 2019; Schmitt et al., 2008, 2017). However, current research on gender equality paradoxes based on ILSA data typically does not account for more complex group differences, e.g., based on socioeconomic background.

Swedish summary

Abstract

Syftet med avhandlingen var att utveckla procedurer som gör det möjligt för forskare att göra jämförelser av prestations- och motivationsskalor inom matematik och naturvetenskap i åk 8 över en lång tidsperiod, trots förändringar av instrument, populationer och procedurer mellan olika undersökningar. Data valdes ut från internationella storskaliga studier som genomförts av International Association for the Evaluation of Educational Achievement. Elevdata användes från ”Trends in International Mathematics and Science Study” (TIMSS) och dess fyra föregångare som genomfördes före 1995. Eftersom studierna har riktat sig till delvis olika populationer (13-åringar, 14-åringar, elever i åttondeklass), och eftersom begreppen har ändrats något mellan undersökningarna, syftade den aktuella avhandlingen till att: 1) utvärdera graden av jämförbarhet mellan dessa studier; 2) koppla de kognitiva testresultaten till TIMSS-skalan med hjälp av item response theory-modellering (IRT); 3) undersöka möjligheten att koppla motivationsskalorna i dessa undersökningar med olika IRT-procedurer och strukturella ekvationsmodeller. Studier som gjorts sedan 1960-talet visade på en hög nivå av stabilitet i undersökningarnas slutsatser och kognitiva begrepp. Motivationsbegreppen visade sig vara mindre stabila. Resultaten visade också att skalornas jämförbarhet har förbättrats över tid. Olika länkningsmetoder gav liknande resultat på land-nivå vad gäller trendbeskrivningar av prestationer och motivation. De länkningstekniker som diskuteras i denna avhandling kan tillämpas på andra storskaliga studier, där förändringar har skett mellan olika undersökningar. Med de skalor som fastställts i denna avhandling är det dessutom möjligt att undersöka långsiktiga förändringar i utbildningssystemen. Kraftfulla statistiska metoder kan tillämpas på dessa longitudinella data på systemnivå för att ta itu med forskningsfrågor som avser orsaksförhållanden.

Inledning

International Association for the Evaluation of Educational Achievement (IEA) har upprätthållit trendskalor för skolprestationer inom matematik och naturvetenskap sedan 1995. Innan dess genomförde IEA fyra internationella storskaliga undersökningar (ILSA) i dessa ämnen, men resultaten från dessa tidiga studier har inte

varit officiellt kopplade till TIMSS-skalorna. I denna avhandling benämns ILSA som genomfördes före 1995 som förstafas-studier, medan de som genomfördes efter 1995 benämns andrafas-studier (Gustafsson, 2008). Förstasfas-studierna i matematik och naturvetenskap finns förtecknade i Tabell 1.

Tabell 1 IEA ILSA av matematik och naturvetenskap genomförda i den första fasen

Bedömning	Tidpunkt för datainsamling	Antal deltagande utbildningssystem
First International Mathematics Study (FIMS)	1964	12
First International Science Study (FISS)	1970–71	17
Second International Mathematics Study (SIMS)	1980–82	20
Second International Science Study (SISS)	1983–84	24

IEA:s beslut att inte koppla samman studierna från de två faserna motiverades av de förändringar som har gjorts i instrument, populationer och procedurer mellan de olika undersökningarna (Martin & Kelly, 1996a). Tekniska och metodmässiga möjligheter vid den här tiden kom också att begränsa möjligheten att länka undersökningar. Allt sedan den första undersökningen har beslut fattats om exempelvis urval av uppgifter och provtagare, och hur uppgifter har formulerats. Dessa beslut innebär utmaningar för jämförbarheten och följaktligen för möjligheterna att koppla samman undersökningarna. Nya tekniska och metodmässiga framsteg gör det dock möjligt att ta itu med sådana utmaningar.

Tidigare forskning har visat att det är möjligt att koppla kognitiva resultat från de tidiga IEA ILSA till senare undersökningar med olika länkningsmetoder. Ett tillvägagångssätt har varit att länka samman skalor från undersökningar som inkluderar gemensamma uppgifter med hjälp av IRT-modellering (Item Response Theory). Afrassa (2005) och Strietholt och Rosén (2016) kopplade samman kognitiva resultat dels inom matematik dels avseende läsprestationer med detta tillvägagångssätt. Afrassas (2005) länkningsstudie förblev dock begränsad när det gäller att utvärdera jämförbarheten med TIMSS-skalan och omfattningen av de utbildningssystem som ingår i länkningen.

Ett annat tillvägagångsätt vad gäller länkning, vilket bygger på klassisk testteori (CTT), har använts för testresultat från olika regionala, nationella eller internationella undersökningar. Här har inte alla undersökningar överlappande uppgifter, varför länkningen måste genomföras under starkare antaganden avseende förmågefördelningarna (se till exempel Chmielewski, 2019; Hanushek & Woessmann, 2012).

Trendmätning av affektiva resultat började med 2011 års genomförande av TIMSS. Vissa skalor för enkäter som inkluderade gemensamma uppgifter i TIMSS

2011, TIMSS 2015 och TIMSS 2019 var kopplade till gemensamma mått (Martin et al., 2012, 2016; Yin & Fishbein, 2020). Så vitt jag vet finns det ingen tidigare forskning där syftet varit att utvidga dessa longitudinella affektiva skalor.

Man kan dra slutsatsen att de senaste metodologiska framstegen och den ökande betydelse som ILSA har inom utbildningssystemen, gör det värt att utforska de möjligheter som ligger i äldre data. Bidraget från denna avhandling är dubbelt. För det första kan de länkningstekniker som diskuteras i denna avhandling tillämpas på andra storskaliga studier, där förändringar har skett mellan olika undersökningar. För det andra kan de skalor som skapats i denna avhandling potentiellt få användning i framtida longitudinella studier.

Syfte

I avhandlingen undersöks nyare och äldre internationella storskaliga mätningar avseende matematik och naturvetenskap i syfte att länka dessa undersökningar för att skapa ett gemensamt mått på utbildningsresultat på systemnivå. Det främsta syftet med att koppla samman undersökningarna är att förse forskarna med jämförbara data om resultat i matematik och naturvetenskap i årskurs åtta, liksom motivation, över en lång tidsperiod. Skalorna i kombination med kraftfulla analysmetoder, som longitudinella modelleringstekniker på landnivå och avancerade ekonometriska metoder, möjliggör undersökning av förändringar i utbildningssystem. Exempelvis kan utbildningsreformer som får effekt på lång sikt utvärderas på nationell nivå. I det jämförande sammanhanget är longitudinella studier användbara för att utforska globala fenomen, såsom trender mot en ”världsläroplan” (Johansson & Strietholt, 2019; Rutkowski & Rutkowski, 2009) eller förändringar i ”det socioekonomiska prestationsgapet” (Broer et al. al., 2019; Chmielewski, 2019). Denna avhandling består av en integrerande uppsats och tre empiriska studier. Frågor om förändring, jämförbarhet, länkning och skalning undersöks i studierna. Denna avhandling styrs av två övergripande forskningsfrågor:

1. I vilken utsträckning är elevresultaten jämförbara i IEA:s undersökningar i första och andra fasen om matematik och naturvetenskap?
2. Hur påverkar olika länkningsmetoder beskrivningarna av trender på systemnivå?

Metod

Det empiriska arbetet i denna avhandling baseras på data från populationerna som representerar 13-åringar (FIMS och SIMS), 14-åringar (FISS och SISS) och elever i åttonde klass (TIMSS-studierna). Data från förstafas-studierna bearbetades

annorlunda jämfört med data från studier som genomfördes senare. Denna avhandling har dragit fördel av det arbete som har gjorts inom projektet ”Center for Comparative Analysis of Educational Achievement” (COMPEAT). Detta projekt har förbättrat förutsättningarna för sekundäranalys genom att göra data och dokumentation från de tidiga studierna tillgängliga online i uppdaterade format[5]. Data och dokumentation för TIMSS-studierna laddades ner från IEA Study Data Repository[6].

I IEA ILSA har proven behållit de gemensamma uppgifterna mellan studierna. Dessa uppgifter fungerar som ankartest mellan undersökningarna och benämns bryggуppgifter i denna avhandling. Studie I och II handlade om att koppla ihop kognitiva uppgifter i matematik och naturvetenskap. I studie III undersöktes möjligheterna att koppla samman frågor som mäter inre och yttre motivation.

Resultatens jämförbarhet

Detta avsnitt ger en kort översikt över de metoder för att undersöka jämförbarhet i longitudinella data och tvärsnittsdata som används i studierna I-III. Grunden för de tre empiriska studierna ligger i utvärderingen av graden av likhet mellan genomförandet av studierna. Graden av likhet mellan de undersökningar som ska kopplas samman bestämmer ”nyttan och rimligheten” (Kolen & Brennan, 2014, s. 498) av länkningen. Kolen och Brennan (2014) föreslog fyra kriterier för att utvärdera likhet: slutsatser, populationer, begrepp och mätegenskaper. Sålunda måste jämförbarheten i studiernas mål och syften utvärderas med avseende på vilka typer av slutsatser som dras från testerna som ska länkas. Sedan måste överensstämmelsen mellan målpopulationerna för de undersökningar som ska kopplas undersökas. Därefter måste likheten mellan de uppmätta begreppen utvärderas. Slutligen måste också jämförbarheten i mätförhållandena, såsom testlängd, testformat och administration, granskas.

Efter denna innehållsanalys, undersöktes bryggуppgifternas jämförbarhet över tid. Mer specifikt undersöktes beteendet hos bryggуppgifterna i de olika studierna med delta-plot-metoden (Angoff & Ford, 1973). ”Delta-plot”-metoden är en metod för att identifiera testuppgifter som fungerar olika för olika grupper (differential item functioning, DIF) i detta fall för dikotomt poängsatta uppgifter (Magis & Facon, 2014). Den (omvandlade) andelen korrekta svar (på en testuppgift) eller svar som indikerar positivt stöd (på en enkätfråga) för referensgruppen jämförs med fokalgruppen.

---

5 https://www.gu.se/en/center-for-comparative-analysis-of-educational-achievement-compeat

6 https://www.iea.nl/data-tools/repository/timss

Varje testuppgift/enkätfråga representeras som en punkt (en s.k. deltapunkt) i ett plotdiagram. Den grafiska representationen inkluderar även en regressionslinje som referenspunkt. Om det inte finns någon DIF återfinns dessa andelar korrekt/positiva svar på denna linje. Uppgifter som hamnar utanför regressionslinjen flaggas som DIF-objekt. Magis och Facon (2014) har föreslagit tröskelvärden för att fastställa när DIF för en uppgift föreligger genom att använda ett normalitetsantagande för deltapunkterna.

Delta-plot-metoden för att upptäcka DIF fungerar under CTT-ramverket. Valet av denna metod har gjorts av flera skäl. För det första är delta-plot-metoden inte en beräkningsintensiv metod. För det andra tar delta-plot-metoden, till skillnad från andra metoder, hänsyn till uppgifternas relativa beteende, vilket innebär att brygg-uppgifter utvärderas i förhållande till alla uppgifter som ingår i bryggan. Dock kvarstår vissa problem med de traditionella DIF-analysmetoderna, till exempel identifiering, som har diskuterats flitigt i litteraturen (se till exempel Bechger & Maris, 2015; Cuellar, 2022; Cuellar et al., 2021; Doebler, 2019; Yuan et al., 2021).

Slutligen bedöms mätinvariansen hos begreppen mellan länder. De kognitiva måtten har antagits vara invarianta eller ekvivalenta, d.v.s. att de har mättekniskt fungerat likadant för alla deltagande länder, med hänvisning till de många kvalitets-säkringsprocesser som tillämpades i undersökningarna. Den tvärkulturella jämförbarheten hos de affektiva måtten utvärderades med hjälp av flergrupps konfirmatorisk faktoranalys (MGCFA) för varje tidpunkt. Konfirmatorisk faktoranalys (CFA) valdes på grundval av synpunkten från Meade och Lautenschlager (2004) att CFA teoretiskt sett är att föredra framför IRT-metoder när antalet uppgifter är få. Frågorna i formuläret behandlades som kategorivariabler och eleverna grupperades efter land. Det första steget var att identifiera en baslinjemodell och pröva antagandet om konfigural invarians, d.v.s. att faktorstrukturen är densamma mellan länder. Efter fastställande av konfigural invarians prövades ekvivalensen i tröskelvärden mellan indikator och faktor, och därefter ekvivalensen i faktor-laddningar (Svetina et al., 2020; H. Wu & Estabrook, 2016).

Hur väl dessa enfaktors-modeller passade data utvärderades med absoluta och relativa anpassningsmått. $\chi 2$, ”root mean square error of approximation” (RMSEA) och ”standardized root mean square residual" (SRMR) fungerade som absoluta modellanpassningsindex. Den relativa passningen indikerades av det jämförande "comparative fit index" (CFI) och Tucker-Lewis-indexet (TLI). Vid användning av dessa mått behövde viss försiktighet iakttas. För det första är $\chi 2$-värdena känsliga för stickprovstorlek (Brown, 2015). För det andra kan förekomsten av negativt formulerade uppgifter få potentiellt endimensionella CFA-modeller att uppvisa dålig modellanpassning (se till exempel Marsh, 1996; Steinmann et al., 2021; Woods, 2006; Zhang et al., 2016). Slutligen påverkas modellanpassningsvärden av många faktorer,

såsom skattningsmetod eller kategorisk/kontinuerlig specification. Shi och Maydeu-Olivares (2020) har därför föreslagit att man endast använder SRMR eftersom det är ett mer konsekvent och gemensamt mått på modellpassning för samtliga faktorer.

Tillvägagångssätt vid länkning

Flera metoder för länkning av skalor undersöktes i de tre ingående empiriska studierna. När det gäller de kognitiva skalorna började det empiriska arbetet med en analys av om det var innehållsligt och empiriskt möjligt att länka samman de olika skalorna. Den första studien kan därför ses som en förberedande studie för Studie II. I Studie II transformerades provpoängen i matematik- och naturvetenskap till TIMSS trendskala. Studie III är en explorativ undersökning av möjligheten att koppla affektiva skalor, varför flera olika tillvägagångssätt jämfördes.

I studie I, som baserades på sammanslagna data från de fyra länder som deltagit vid varje tidpunkt från FIMS till TIMSS 2015, innefattade länkningsproceduren tre huvudsteg. Först prövades olika IRT-modeller för att välja modellerna med bäst anpassning. De IRT-modeller som valdes var en tvåparameters logistisk modell (2PL) för dikotoma uppgifter, d.v.s. flervalsuppgifter och öppna svars-uppgifter för en poäng; den generaliserade partiella kreditmodellen (GPCM) för polytoma uppgifter, d.v.s. uppgifter med produktiva svar för två eller fler poäng.

I steg två skattades uppgiftsparametrarna med s.k. samtidig kalibrering. På så vis placerades uppgiftsparametrarna från alla studier och för alla fyra länderna samtidigt på samma IRT-skala. För att få varje lands data att väga lika användes statistiska vikter. I IEA:s senare studier finns s.k. senatsvikter som uppgår till 500 för varje lands elevdata (dessa motsvarar stratumvikter i SIMS vilka omskalades till summan till 500), och sålunda bidrog varje land lika till uppgiftskalibreringen. I FIMS 1964 databas fanns inga viktvariabler, varför individer inom ett land fick vägas lika, för att summera till 500.

I steg tre användes de estimerade uppgiftsparameterarna för att räkna fram personpoäng. Därtill drogs fem plausibla värden (PV) per elev med den 'förväntade a-posteriori-metoden'. De uppskattade måtten omvandlades till skalpoäng; sålunda omvandlades varje PV till ett mått med ett medelvärde på 500 och en standardavvikelse på 100 poäng. De transformerade poängen användes för att beräkna den genomsnittliga matematikprestationen för respektive land per studie enligt Rubins (1987) sammanvägningsregel.

Studie II jämförde två länkningsmetoder för matematikskalan. Metoden *four-country-all-time* (points) använde de parametrar som skattades med metoden i Studie I. Proceduren började med separata skattningar av förmåga för FIMS och SIMS, med uppgiftsparametrarna fixerade till de skattade värdena från Studie I. Därefter matchades fördelningen av de fem skattade PV för FIMS och SIMS med fördelningen

av de rapporterade PV i TIMSS 1995. Detta gjordes genom att beräkna transformationskonstanter, liknande TIMSS länkningsprocedur, i två steg: (1) Medelvärdena och standardavvikelserna för de rapporterade PV från 1995, som ligger på den erforderliga skalan, matchades med medelvärdena och standardavvikelserna för 1980 års PV för samma uppsättning länder och skolår inom länder, som är på en oberoende skala. Sedan placerades alla länder på denna skala med samma transformationskonstanter. (2) Samma procedur gjordes för att matcha fördelningen av FIMS PV med SIMS PV.

Den andra metoden, *first-second-time* metoden, innebar samtidig kalibrering av uppgiftsparametrarna med användning av den första och andra ILSA för matematik med brygguppgifternas uppgiftsparametrar fixerade till de värden som rapporterades för TIMSS 1995. Dessa uppgiftsparametrar rapporterades efter en omskalnings-procedur i 1999 års utvärderingscykel (Martin et al., 2000). På samma sätt som skalningsproceduren i det första steget, matchades sedan förmågefördelningen av SIMS med den rapporterade TIMSS 1995-skalan och sedan SIMS med FIMS.

När man konstruerade skalan för naturvetenskap valdes first-second-time metoden av flera skäl. För det första är IRT-modellerna desamma som de som används i TIMSS-procedurerna, d.v.s. 2PL, logistisk modell med tre parametrar (3PL) och GPCM. För det andra är denna procedur mer ekonomisk även när data från fler länder (men färre tidpunkter) används. Om vi jämför mängden information, d.v.s. antalet uppgifter som används för kalibrering i de två tillvägagångssätten, kan vi å ena sidan notera att den parallella kalibreringen av four-country-all-time omfattar 893 uppgifter (1964–2015), medan first-second-time metoden använder uppgifterna mellan 1964 och 1995, det vill säga 373 uppgifter. Å andra sidan är det viktade antalet svar som används för länkningen mellan SIMS och TIMSS 1995 nära tre gånger så stort i first-second-time metoden jämfört med four-country-all-time-metoden på grund av det större antalet länder (42) som är involverade i förfarandet.

I studie III prövades tre olika ansatser för att konstruera longitudinella skalor för affektiva mått på motivation: IRT, CFA och marknadskorgsmetoden. Med IRT-ansatsen fann man att GPCM-modellen passade data bäst. Därefter utfördes skattningen av uppgiftparametrar med samtidig kalibrering av alla motivationsfrågor i alla studier, och på så sätt placerades parametrarna för alla uppgifter automatiskt på samma skala. Parametrarna för brygguppgifterna antogs vara identiska i varje prov. I ett tredje steg skattades personpoäng vilka omvandlades till en skala med ett medelvärde på fem och en standardavvikelse på ett.

För det andra anpassades en CFA-modell till varje motivationsskala i ett poolat urval bestående av data från alla länder och undersökningsomgångar. Stark invarians av ankaruppgifter mellan länder och över tid antogs. Faktorpoäng beräknades med ”maximum likelihood estimation with robust standard errors” (MLR), medan

uppgifterna behandlades som kategorivariabler. Faktorpoängen omvandlades sedan till en skala med ett medelvärde på fem och en standardavvikelse på ett.

Slutligen tillämpades en marknadskorgs-strategi. Marknadskorgsmetoden förutsätter att de uppgifter som ingår i undersökningen definierar begreppet. I det här fallet är antagandet att alla uppgifter över samtliga tidpunkter, som är relaterade till inre och yttre motivation avseende matematik, definierar varje begrepp och kan betraktas som en marknadskorg av representativa uppgifter. Uteblivna svar uppstår som en konsekvens av förändringar i frågeformulären över studierna.

En mätmodell per land användes för att generera plausibla svar som kompletterar de saknade svaren enligt en procedur som föreslagits av Zwitser et al. (2017) Mätmodellen anpassades till data för varje land separat för att ta hänsyn till potentiella skillnader mellan länder. Mätmodellen var GPCM-modellen för att få överensstämmelse med resultaten från IRT-metoden och TIMSS-proceduren för att länka kontextuella skalor.

Med hjälp av uppgiftsparametrarna som skattades genom att anpassa mätmodellerna till data, imputerades uteblivna svar fem gånger per respondent. Därefter beräknades individuella summapoäng, varigenom fem PV per elev uppskattades. Slutligen omvandlades de fem PV till en skala med ett medelvärde på fem och en standardavvikelse på ett.

Sammanfattande resultat och diskussion

Jämförbarhet av resultaten

Den innehållsmässiga utvärderingen av jämförbarheten mellan undersökningarna visar på en hög grad av stabilitet vad gäller slutsatser och uppmätta begrepp bland studierna. Förändringar i provtagnings- och testförhållandena skapade utmaningar för länkningen. En del av dessa utmaningar hanterades för att uppnå en tillräcklig grad av likhet mellan studierna, medan det finns kvar vissa begränsningar hos skalorna.

Brygguppgifter

I studie I tillämpades delta-plot-metoden på alla sju bryggorna mellan intilliggande tidpunkter för att testa antagandet om beteendet hos vanliga uppgifter. Tre uppgifter flaggades för DIF i de två första bryggorna, det vill säga från FIMS till SIMS respektive från SIMS till TIMSS 1995. Inga fall av DIF upptäcktes i resten av bryggorna. Uppgifter som visade DIF exkluderades från kalibreringen. Vidare uteslöts tolv uppgifter på grund av uteblivna svar i alla länder. Totalt ingick 893 uppgifter i den samtidiga kalibreringen.

I studie II användes delta-plot-metoden för de sex uppsättningarna av bryggor. Dessa bryggor ingår i matematikuppgifterna mellan 1964–1980 (brygga 1), 1964–1995 (brygga 2) och 1980–1995 (brygga 3), och bland de naturvetenskapliga undersökningarna mellan 1970–1984 (brygga 4), 1970–1995 (brygga 5) och 1984–1995 (brygga 6). Två uppgifter i den första, en uppgift i den tredje och två uppgifter i den fjärde bryggan flaggades för DIF. I den sista, first-second-time länkningen, behandlades dessa som unika uppgifter.

För att testa antaganden om hur väl ankartestet och hela testet fungera, beräknades Pearsons korrelationer. Dessa korrelationer var måttliga eller höga: för FIMS, $r = 0{,}97$, $p < 0{,}001$ (brygga 1) och $r = 0{,}84$, $p < 0{,}001$ (brygga 2); för SIMS, $r = 0{,}88$, $p < 0{,}001$ (brygga 1) och $r = 0{,}66$, $p < 0{,}001$ (brygga 3); för FISS $r = 0{,}92$, $p < 0{,}001$ (brygga 4) och $r = 0{,}69$, $p < 0{,}001$ (brygga 5); för SISS, $r = .86$, $p < .001$ (brygga 4) och $r = .80$, $p < .001$ (brygga 6).

I studie III användes delta-plot-metoden för varje brygga mellan på varandra följande tidpunkter. Testerna utfördes för varje land separat, såväl som med den sammanslagna informationen. Inget test identifierade objekt som flaggats för DIF.

Kulturell invarians

I studie III testades mätinvariansen över länder vid varje tidpunkt. MGCFA-resultaten för SIMS avslöjade att mätinvarians inte höll för Japan och alla ytterligare analyser i denna studie fortsattes exklusive data från detta land. Tröskel- och invarianstesten gav en acceptabel modellpassning vid de flesta tidpunkter för flergruppsmodellen med fem länder.

Trendbeskrivningar genom länkning

I detta avsnitt presenteras resultat från de empiriska studierna relaterade till den andra forskningsfrågan: Hur påverkar olika länkmetoder beskrivningarna av trender på systemnivå? Resultaten diskuteras i termer av de uppmätta begreppen.

Matematik

Det fanns två huvudsakliga skillnader i de tillvägagångssätt som tillämpades i studie II. I first-second-time metoden inkluderades en gissningsparameter i IRT-modellen för flervalsuppgifter, och fler svar var tillgängliga för uppgiftskalibreringen än i four-country-all-time metoden. Länkmetoderna gav mycket likartade resultat trots dessa skillnader. Detta resultat indikerar skalornas robusthet. De individuella PV (fem för varje provtagare) som beräknades i four-country-all-time och first-second-time metoden visade starka korrelationer för alla möjliga rimliga värdepar. Den starka

korrelationen på individnivå motsvarades väl av aggregeringen på landnivå i nästan alla fall med undantag för Japan.

Motivation för att lära sig matematik

De fem länder som ingick i Studie II behandlades som en enda grupp både tvärsnittsmässigt och longitudinellt i CFA- och IRT-procedurerna. De behandlades separat i marknadskorgmetoden, men data slogs samman till en enda modell över tid. De observerade skalorna konstruerades genom att beräkna summan av poängen per person vid varje tidpunkt dividerat med antalet besvarade frågor. Sedan beräknades de standardiserade poängen för att ge ett medelvärde på fem och en standardavvikelse på ett. De tre metoderna gav liknande resultat på såväl individ- som landnivå. Korrelationerna mellan individuella poäng var höga mellan metoderna för båda motivationsbegreppen, mellan 0,96 och 1. Trenderna på landnivå visade liknande mönster över länkmetoderna.

Begränsningar

De slutliga skalorna är offentligt tillgängliga på COMPEAT-webbplatsen tillsammans med dokumentation av skalorna. Jag vill betona att urvalsskillnaderna måste beaktas vid användning av skalorna. Strietholt et al. (2013) utvecklade exempelvis en korrigeringsmodell för att förbättra jämförbarheten mellan länder och IEA-studier om läsning i termer av ålder och skolgång. Ett annat tillvägagångssätt för att ta hänsyn till dessa skillnader mellan tid och länder är att behandla ålder och betygsnivå som rimliga förklaringsvariabler.

Många faktorer påverkar kvaliteten på länkningen. Dessa faktorer inkluderar graden av likhet mellan studierna, stabiliteten hos begreppen vad gäller innehåll, betydelse och sammanhang, och antalet och beteendet hos bryggguppgifterna. Avhandlingen behandlar dessa påverkansfaktorer och potentiella källor till bias ur en innehållsligt och en mätbar synvinkel. Det kan dock finnas en viss osäkerhet när det gäller att utvärdera dessa aspekter eftersom länkningen involverar äldre data.

En möjlig begränsning i de longitudinella prestationsskalorna ligger i jämförbarheten inom länder på grund av olika målpopulationer för undersökningarna. Förskjutningen i IEA:s provtagningsstrategi angreps med så bra approximationer av homogena prov som möjligt. I ytterligare analyser med de nya skalpoängen kan ålder och betygsnivå behandlas som kontrollvariabler.

De relativt få gemensamma uppgifterna i bryggorna som ansluter till TIMSS 1995 är ett annat problem. Förhållandet mellan bryggguppgifter i de affektiva skalorna är mycket mindre oroande än i fallet med de kognitiva skalorna. Den samtidiga

kalibreringsmetoden ger dock den bästa metoden för att hantera situationer med ett fåtal bryggуppgifter, vilket påpekats av Wingersky och Lord (1984).

En annan begränsning gäller kodningen och behandlingen av olika typer av bortfall av data i de kognitiva testerna. I studierna i den första fasen särskildes inte den icke-nådda typen av saknade svar. Därför behandlades de icke nådda uppgifterna som saknade, till skillnad från i TIMSS-proceduren där uppgifterna behandlades som icke givna. Det skulle vara möjligt att göra denna distinktion och utforska inverkan på resultaten.

De affektiva skalorna som analyserades i Studie III har tills nyligen inte utformats för trendmätning. Modifieringar har skett över tid; till exempel ändrades antalet svarsalternativ från 1995. Sättet att hantera mellanalternativet i SIMS utgjorde en begränsning för studien. Dessutom tog länkmetoderna i denna studie inte hänsyn till skillnaderna i motivationsfördelningarna över tid.

En annan begränsning i de affektiva trendskalorna är att medelvärdenas medelfel är underskattade. Anledningen till detta är att på grund av den stratifierade samplingsdesignen i flera steg som används i TIMSS, gäller inte antagandet om det enkla slumpmässiga urvalet som antogs i proceduren för att beräkna medelfelen (L. Rutkowski et al., 2010).

Slutligen är en av de mest utmanande återstående frågorna om ändringar i frågeformuleringarna påverkar de interna relationerna mellan motivationsuppgifterna, till exempel vad gäller faktorstruktur. Eftersom icke-identiska uppgifter undersöktes över tid, varierar antalet uppgifter vid nästan varje tidpunkt, vilket gör undersökningar av effekter av ändringar i uppgifternas ordalydelse utmanande.

Slutsatser

Med de skalor som etablerats i denna avhandling är det möjligt att undersöka långsiktiga förändringar genom jämförande studier eller genom longitudinella studier inom länder. Förändringar på systemnivå tar tid, och kräver därför utvärdering av reformer med hjälp av långsiktiga data. Dessutom kan, som tidigare nämnts, kraftfulla statistiska metoder användas för att hantera forskningsfrågor kring orsaker med hjälp av longitudinella data på systemnivå.

Strietholt et al. (2019) granskade exempelvis nyligen den internationella jämförande litteraturen om utbildningspolitikens inverkan på det socioekonomiska prestationsgapet. Författarna fann att det mesta av den befintliga forskningen var beskrivande och beräknade enkla korrelationer baserade på tvärsnittsdata. Ytterligare forskning om indikatorer för socioekonomisk bakgrund i den första fasen av IEA-undersökningarna skulle, i kombination med prestationsskalorna, kunna bidra till att utveckla denna undersökningslinje.

Ett annat potentiellt område som kan dra nytta av de longitudinella skalorna gäller frågor relaterade till den globala utbildningsreformrörelsen (Fuller & Stevenson, 2019; Sahlberg, 2016). Sådana fenomen omfattar bland annat privatisering, fritt skolval, och lärarutbildning.

Det är också av stort intresse att fortsätta utforskningen av långsiktiga trender i attityder till att lära sig matematik och naturvetenskap. Det finns betydligt större utmaningar med dessa resultat än prestationsskalorna på grund av förändringar på uppgiftnivå. Marknadskorgsansatsen som används i denna avhandling erbjuder möjligheter till länkning med färre antaganden om jämförbarhet än länkning på uppgiftsnivå.

Ett annat intressant område är att utforska möjligheterna att koppla samman IEA-studierna i första och andra fasen om naturvetenskap för årskurs fyra-populationen. Data är offentligt tillgängliga för de gamla studierna. Vidare var i 1995 års TIMSS-studie prestationstesterna för de yngre och äldre populationerna kopplade genom brygguppgifter (Martin & Kelly, 1996a). Det kan vara intressant att åter pröva denna koppling för att undersöka utvecklingsförändringar.

Vidare kan det vara användbart att kartlägga andra kontextuella indikatorer från förstafasstudierna, vilket skulle kunna underlätta mer komplexa undersökningar av till exempel könsskillnader. Andelen kvinnor som väljer ett matematiskt spår inom gymnasie- och högre utbildning eller STEM-relaterade yrken är fortfarande orimligt låg och inte relaterad till matematikprestationer i många länder. Men aktuell forskning om jämställdhetsparadoxer baserade på ILSA-data tar vanligtvis inte hänsyn till mer komplexa gruppskillnader, till exempel baserade på socioekonomisk bakgrund.

References

Adams, R. J., Wu, M., & Macaskill, G. (1996). Scaling methodology and procedures for the mathematics and science scales. In M. O. Martin & D. L. Kelly (Eds.), *Third international mathematics and science study technical report* (Vol. 2, pp. 111–146). TIMSS & PIRLS International Study Center, Boston College.

Afrassa, T. M. (2005). Monitoring mathematics achievement over time: A secondary analysis of FIMS, SIMS and TIMS: a Rasch analysis. In Alagumalai, Curtis, David D. & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 61–77). Springer.

Allardt, E. (1990). Challenges for comparative social research. *Acta Sociologica*, *33*(3), 183–193. https://doi.org/10.1177/000169939003300302

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Angoff, W., & Ford, S. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*(2), 95–106.

Beaton, A. E. (1990). Introduction. In A. E. Beaton & R. Zwick (Eds.), *The effect of changes in the national assessment: Disentangling the NAEP 1985-86 reading anomaly* (pp. 1–13).

Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement*, *29*(2), 163–175. https://doi.org/10.1111/j.1745-3984.1992.tb00372.x

Beaton, A. E., Rogers, A. M., Gonzalez, E., Hanly, M. B., Kolstad, A., Rust, K. F., Sikali, E., Stokes, L., & Jia, Y. (2011). *The NAEP Primer.* U.S. Department of Education, National Center for Education Statistics.

Beaton, A. E., & Zwick, R. (Eds.). (1990). *The effect of changes in the national assessment: Disentangling the NAEP 1985-86 reading anomaly* (Issue ETS-17-TR-21). National Assessment of Educational Progress.

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*(2), 317–340. https://doi.org/10.1007/s11336-014-9408-y

Bos, K. T. (2002). *Benefits and limitations of large-scale international comparative achievement studies: The case of IEA's TIMSS study* [Doctoral dissertation, University of Twente]. https://ris.utwente.nl/ws/files/6081834/thesis_K_Bos.pdf

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Braun, H. I., & Singer, J. D. (2019). Assessment for monitoring of education systems: International comparisons. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), 75–92. https://doi.org/10.1177/0002716219843804

Breda, T., Jouini, E., Napp, C., & Thebault, G. (2020). Gender stereotypes can explain the gender-equality paradox. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(49), 31063–31069. https://doi.org/10.1073/pnas.2008704117

Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic inequality and educational outcomes: Evidence from twenty years of TIMSS* (Vol. 5). Springer. https://doi.org/10.1007/978-3-030-11991-1

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second edition). The Guilford Press.

Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review*, *84*(3), 517–544. https://doi.org/10.1177/0003122419847165

Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries: An empirical study* (Vol. 1). Almqvist & Wiksell.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Csíkos, C., Pásztor, A., Rausch, A., & Szitányi, J. (2020). A matematikai nevelés kutatásának aktuális irányzatai [Recent trends of research on mathematics education]. *Magyar Tudomány*. https://doi.org/10.1556/2065.181.2020.1.3

Cuellar, E. (2022). *Making sense of DIF in international large-scale assessments in education* [Doctoral dissertation]. University of Amsterdam.

Cuellar, E., Partchev, I., Zwitser, R., & Bechger, T. (2021). Making sense out of measurement non-invariance: How to explore differences among educational systems in international large-scale assessments. *Educational Assessment, Evaluation and Accountability*, *33*(1), 9–25. https://doi.org/10.1007/s11092-021-09355-x

Doebler, A. (2019). Looking at DIF from a new perspective: A structure-based approach acknowledging inherent indefinability. *Applied Psychological Measurement*, *43*(4), 303–321. https://doi.org/10.1177/0146621618795727

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21–42). Springer.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Ercikan, K., & Roth, W.-M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher*, *35*(5), 14–23. https://doi.org/10.3102/0013189X035005014

European Science Foundation. (2017). *The european code of conduct for research integrity* (Revised edition). All European Academies.

Fjellman, A.-M., Yang Hansen, K., & Beach, D. (2019). School choice and implications for equity: The new political geography of the Swedish upper secondary school market. *Educational Review*, *71*(4), 518–539. https://doi.org/10.1080/00131911.2018.1457009

Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures* (pp. 12.1–12.146).

Fuller, K., & Stevenson, H. (2019). Global education reform: Understanding the movement. *Educational Review*, *71*(1), 1–4. https://doi.org/10.1080/00131911.2019.1532718

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: An introduction* (2nd ed.). SAGE Publications.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *The American Psychologist*, *53*(1), 5–26. https://doi.org/10.1037/0003-066X.53.1.5

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal*, *7*(1), 1–17. https://doi.org/10.2304/eerj.2008.7.1.1

Gustafsson, J.-E. (2018). International large scale assessments: Current status and ways forward. *Scandinavian Journal of Educational Research*, *62*(3), 328–332. https://doi.org/10.1080/00313831.2018.1443573

Gustafsson, J.-E., & Nilsen, T. (2022). Methods of causal analysis with ILSA data. In T. Nilsen, A. Stancel-Piątak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education*. Springer International Publishing. https://doi.org/10.1007/978-3-030-38298-8_56-1

Gustafsson, J.-E., & Rosén, M. (2006). The dimensional structure of reading assessment tasks in the IEA reading literacy study 1991 and the Progress in International Reading Literacy Study 2001. *Educational Research and Evaluation*, *12*(5), 445–468. https://doi.org/10.1080/13803610600697179

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publications.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth*, *17*(4), 267–321. https://doi.org/10.1007/s10887-012-9081-x

Härnqvist, K. (1975). The international study of educational achievement. *Review of Research in Education*, *3*(1), 85–109. https://doi.org/10.3102/0091732X003001085

Hernández-Torrano, D., & Courtney, M. G. R. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-Scale Assessments in Education*, *9*(17), 1–33. https://doi.org/10.1186/s40536-021-00109-1

Heyneman, S. P., & Lee, B. (2013). The impact of international studies of academic achievement on policy and research. In L. Rutkowski, M. von Davier, & D. Rutkowski

(Eds.), *Handbook of international large-scale assessment* (pp. 37–72). CRC Press. https://doi.org/10.1201/b16061-5

Heyneman, S. P., & Lykins, C. (2008). The evolution of comparative and international education statistics. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 105–127). Routledge.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (pp. 187–220). Praeger Publishers.

Howson, G. (1999). The value of comparative studies. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 165–188). Falmer.

Husén, T. (1979). An international research venture in retrospect: The IEA surveys. *Comparative Education Review*, *23*(3), 371–385. https://doi.org/10.1086/446067

Husén, T. (1983). Are standards in U.S. schools really lagging behind those in other countries? *The Phi Delta Kappan*, *64*(7), 455–461.

Husén, T., & Postlethwaite, T. N. (1967). Intentions and background to the project. In T. Husén (Ed.), *International study of achievement in mathematics* (pp. 25–34). Almqvist & Wiksell.

Johansson, S. (2016). International large-scale assessments: What uses, what consequences? *Educational Research*, *58*(2), 139–148. https://doi.org/10.1080/00131881.2016.1165559

Johansson, S., & Strietholt, R. (2019). Globalised student achievement? A longitudinal and cross-country analysis of convergence in mathematics performance. *Comparative Education*, *55*(4), 536–556. https://doi.org/10.1080/03050068.2019.1657711

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202.

Keeves, J. P., & Schleicher, A. (1992). Changes in science achievement: 1970-84. In J. P. Keeves (Ed.), *The IEA study of science III: Changes in science education and achievement: 1970 to 1984* (pp. 263–290). Pergamon Press.

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Co.

Kifer, E., & Robitaille, D. F. (1989). Attitudes, preferences and opinions. In D. F. Robitaille & R. A. Garden (Eds.), *The IEA study of mathematics II: Contexts and outcomes of school mathematics* (pp. 178–208). Pergamon Press.

Kirsch, I., Lennon, M., von Davier, M., & Gonzalez, E. (2013). On the growing importance of international large-scale assessments. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments* (pp. 1–11). Springer Netherlands. https://doi.org/10.1007/978-94-007-4629-9_ 1

Klemenčič, E., & Mirazchiyski, P. V. (2018). League tables in educational evidence-based policy-making: Can we stop the horse race, please? *Comparative Education*, *54*(3), 309–324. https://doi.org/10.1080/03050068.2017.1383082

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford Press.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (pp. 155–186). Praeger Publishers.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3. ed.). Springer.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1). https://doi.org/10.1207/s15324818ame0601_5

Mac Giolla, E., & Kajonius, P. J. (2019). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology*, *54*(6), 705–711. https://doi.org/10.1002/ijop.12529

Magis, D., & Facon, B. (2014). DeltaPlotR: An R package for differential item functioning analysis with Angoff's delta plot. *Journal of Statistical Software*, *59*(Code Snippet 1). https://doi.org/10.18637/jss.v059.c01

Majoros, E. (2022). *Linking the first- and second-phase IEA studies on mathematics and science* [Manuscript submitted for publication].

Majoros, E., Christiansen, A., & Cuellar, E. (2022). Motivation towards mathematics from 1980 to 2015: Exploring the feasibility of trend scaling. *Studies in Educational Evaluation*, *74*, 101174. https://doi.org/10.1016/j.stueduc.2022.101174

Majoros, E., Rosén, M., Johansson, S., & Gustafsson, J.-E. (2021). Measures of long-term trends in mathematics: Linking large-scale assessments over 50 years. *Educational Assessment, Evaluation and Accountability*, *33*(1), 71–103. https://doi.org/10.1007/s11092-021-09353-z

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139–160.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*(4), 810–819. https://doi.org/10.1037/0022-3514.70.4.810

Martin, M. O., Gregory, K. D., & Stemler, S. E. (Eds.). (2000). *TIMSS 1999 technical report*. TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., & Kelly, D. L. (Eds.). (1996a). *Third international mathematics and science study technical report: Design and development* (Vol. 1). TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., & Kelly, D. L. (Eds.). (1996b). *Third international mathematics and science study technical report: Implementation and analysis* (Vol. 2). TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., & Mullis, I. V. S. (2004). Overview of TIMSS 2003. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 3–21). TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., Arora, & Alka. (2012). Creating and interpreting the TIMSS and PIRLS 2011 context questionnaire scales. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/methods/t-context-q-scales.html

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 15.1–15.312). TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center, Boston College.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 681–699). Praeger Publishers.

Mazzeo, J., & von Davier, M. (2013). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment*. CRC Press. https://doi.org/10.1201/b16061-13

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361–388. https://doi.org/10.1177/1094428104268027

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11), S69–S77.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan Publishing Co, Inc and American Council on Education.

Messick, S. (1998). Alternative modes of assessment, uniform standards of validity. In M. D. Hakel (Ed.), *Beyond multiple choice* (pp. 59–74). Lawrence Erlbaum Associates Publishers.

Messick, S., Beaton, A. E., & Lord, F. M. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era*. Educational Testing Service.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge. https://doi.org/10.4324/9780203821961

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. https://doi.org/10.1207/S15327906MBR3903_4

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. https://doi.org/10.1007/BF02294457

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. https://eric.ed.gov/?id=ED353302

Mislevy, R. J. (2017). *Sociocognitive foundations of educational measurement*. Routledge.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. SAGE Publications.

OECD. (1999). *Measuring student knowledge and skills: A new framework for assessment*. OECD Publishing.

OECD. (2019). *TALIS 2018 technical report*. OECD Publishing. http://www.oecd.org/education/talis/TALIS_2018_Technical_Report.pdf

Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). *Bridging validity and evaluation to match international large–scale assessment claims and country aims* (Issue RR-18-27). https://doi.org/10.1002/ets2.12214

Olsen, R. V. (2005). *Achievement tests from an item perspective: An exploration of single item data from the PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science* [Doctoral dissertation, University of Oslo]. https://doi.org/10.5617/nordina.457

Pedhazur, E. J., & Pedhazur Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach*. Lawrence Erlbaum Associates.

Plomp, T. (1998). The potential of international comparative studies to monitor the quality of education. *Prospects*, *28*(1), 45–59. https://doi.org/10.1007/BF02737779

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Robinson, J. P. (2013). Causal inference and comparative analysis with large-scale assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 521–545). CRC Press.

Robitaille, D. F., & Taylor, A. R. (1989). Changes in patterns of achievement between the first and second mathematics studies. In D. F. Robitaille & R. A. Garden (Eds.), *The IEA study of mathematics II: Contexts and outcomes of school mathematics* (pp. 153–177). Pergamon Press.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education*, *4*(1). https://doi.org/10.1186/s40536-016-0019-1

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, L., & Rutkowski, D. (2009). Trends in TIMSS responses over time: Evidence of global forces in education? *Educational Research and Evaluation*, *15*(2), 137–152. https://doi.org/10.1080/13803610902784352

Sahlberg, P. (2016). The global educational reform movement and its impact on schooling. In K. E. Mundy, A. Green, B. Lingard, & A. Verger (Eds.), *The handbook of global education policy* (pp. 128–144). John Wiley & Sons, Ltd.

San Martín, E. (2016). Identification of item response theory models. In *Handbook of item response theory: Statistical tools* (Vol. 2, pp. 127–150). Chapman and Hall/CRC.

Schlotter, M., Schwerdt, G., & Woessmann, L. (2014). Econometric methods for causal evaluation of educational policies and practices: A non-technical guide. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 95–126). Waxmann.

Schmidt, W. H., Wolfe, R. G., & Kifer, E. (1992). The identification and description of student growth in mathematics achievement. In L. Burstein (Ed.), *The IEA study of mathematics III* (pp. 59–99). Pergamon Press.

Schmitt, D. P., Long, A. E., McPhearson, A., O'Brien, K., Remmert, B., & Shah, S. H. (2017). Personality and gender differences in global perspective. *International Journal of Psychology*, *52*(S1), 45–56. https://doi.org/10.1002/ijop.12265

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, *94*(1), 168–182. https://doi.org/10.1037/0022-3514.94.1.168

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement*, *80*(3), 421–445. https://doi.org/10.1177/0013164419885164

Steinmann, I., Sánchez, D., van Laar, S., & Braeken, J. (2021). The impact of inconsistent responders to mixed-worded scales on inferences in international large-scale assessments. *Assessment in Education: Principles, Policy & Practice*, 1–22. https://doi.org/10.1080/0969594X.2021.2005302

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, *161*(3844), 849–856.

Stoet, G., & Geary, D. C. (2020). Corrigendum: The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, *31*(1), 110–111. https://doi.org/10.1177/0956797619892892

Strietholt, R., Gustafsson, J.-E., Hogrebe, N., Rolfe, V., Rosén, M., Steinmann, I., & Hansen, K. Y. (2019). The impact of education policies on socioeconomic inequality in student achievement: A review of comparative studies. In Volante & Melchior (Eds.), *Socioeconomic inequality and student outcomes. Education Policy & Social Inequality* (Vol. 4, pp. 17–38). Springer Singapore. https://doi.org/10.1007/978-981-13-9863-6_ 2

Strietholt, R., & Rosén, M. (2016). Linking large-scale reading assessments: Measuring international trends over 40 years. *Measurement: Interdisciplinary Research and Perspectives*, *14*(1), 1–26. https://doi.org/10.1080/15366367.2015.1112711

Strietholt, R., Rosén, M., & Bos, W. (2013). A correction model for differences in the sample compositions: The degree of comparability as a function of age and schooling. *Large-Scale Assessments in Education*, *1*(1), 1. https://doi.org/10.1186/2196-0739-1-1

Svensson, A. (2008). Har dagens tonåringar sämre studieförutsättningar? En studie av förskjutningar i intelligenstestresultat från 1960-talet och framåt. *Pedagogisk Forskning I Sverige*, *13*(4), 258–277.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using Mplus and the lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111–130. https://doi.org/10.1080/10705511.2019.1602776

Swedish Research Council. (2017). *Good research practice*. Swedish Research Council.

Thorndike, R. L. (1967). The mathematics tests. In T. Husén (Ed.), *International study of achievement in mathematics* (pp. 90–108). Almqvist & Wiksell.

Travers, K. J., & Weinzweig, A. I. (1999). The second international mathematics study. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 19–29). Falmer.

Truman, D. B. (1959). The American system in crisis. *Political Science Quarterly*, *74*(4), 481. https://doi.org/10.2307/2146419

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*(3), 347–364. https://doi.org/10.1177/014662168400800312

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. https://doi.org/10.1007/s10862-005-9004-7

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Wu, M. (2010). *Comparing the similarities and differences of PISA 2003 and TIMSS* (No. 32; OECD Education Working Papers). OECD Publishing. https://doi.org/10.1787/5km4psnm13nx-en

Yang Hansen, K., & Gustafsson, J.-E. (2016). Causes of educational segregation in Sweden – school choice or residential segregation. *Educational Research and Evaluation*, *22*(1–2), 23–44. https://doi.org/10.1080/13803611.2016.1178589

Yin, L., & Fishbein, B. (2020). Creating and interpreting the TIMSS 2019 context questionnaire scales. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures* (pp. 16.1–16.331). TIMSS & PIRLS International Study Center, Boston College.

Yuan, K.-H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: QQ plots and graphical test. *Psychometrika*, *86*(2), 345–377. https://doi.org/10.1007/s11336-021-09746-5

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the need for cognition scale. *PloS One*, *11*(6), e0157795. https://doi.org/10.1371/journal.pone.0157795

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*, *82*(1), 210–232. https://doi.org/10.1007/s11336-016-9543-8

Longitudinal measurement of country-level student achievement in an international context has become increasingly influential in monitoring educational systems. The first international large-scale surveys on mathematics and science conducted by the International Association for the Evaluation of Education Achievement (IEA) were administered in 1964 and 1970-71. The outcomes of these and the following studies administered before 1995 have not been officially linked to the Trends in International Mathematics and Science Study (TIMSS) trend scales. In this thesis, these older and recent mathematics and science assessments for grade eight are linked to a common scale. The goal of linking the assessments is to provide researchers with comparable data of grade-eight mathematics and science achievement and motivation scales over a long time period.

The thesis includes three empirical studies. Issues of change, comparability, linking, and scaling are investigated in the studies. Study I evaluated the feasibility of constructing comparable trend measures of mathematics achievement as assessed by the IEA from 1964 onwards. Study II compared different linking approaches and linked the early mathematics and science achievement scales with the TIMSS trend scales. Study III explored the feasibility of establishing long-term student motivational scales using eighth-grade data from the Second International Mathematics Study in 1980 and the 1995-2015 iterations of TIMSS.

Erika Majoros has conducted Ph.D. studies in Education at the University of Gothenburg and has previously worked as a rehabilitation specialist for visual impairment. Her main research interests are in the areas of large-scale educational assessments, psychometrics, and affective educational outcomes.

ISBN 978-91-7963-109-3
9789179631093