Improving the modelling of response variation in international large-scale assessments

Annemiek Punter

IMPROVING THE MODELLING OF RESPONSE VARIATION IN INTERNATIONAL LARGE-SCALE ASSESSMENTS

Annemiek Punter

Graduation Committee:

Chairman	Prof. Dr. T.A.J. Toonen
Promotor	Prof. Dr. C.A.W. Glas
	Prof. Dr. T.J.H.M. Eggen
Assistant promotor	Dr. M.R.M. Meelissen
Members	Prof. Dr. L.A. van der Ark
	Prof. Dr. R.J. Bosker
	Dr. R.C.W. Feskens
	Dr. D. Hastedt
	Dr. J.W. Luyten
	Prof. Dr. B.P. Veldkamp

Improving the modelling of response variation in international large-scale assessments
PhD thesis, University of Twente, Enschede, the Netherlands
ISBN: 978-90-365-4686-7
DOI: 10.3990/1.9789036546867
Printed by Ipskamp Printing, Enschede, the Netherlands

IMPROVING THE MODELLING OF RESPONSE VARIATION IN INTERNATIONAL LARGE-SCALE ASSESSMENTS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. T.T.M. Palstra,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday 19 December 2018 at 16.45 hours

by

Renate Annemiek Punter
born 16 October 1987
in Tietjerksteradeel, the Netherlands

This dissertation has been approved by:

Promotor	Prof. Dr. C.A.W. Glas
Promotor	Prof. Dr. T.J.H.M. Eggen
Assistant Promotor	Dr. M.R.M. Meelissen

Acknowledgements

Four years ago, my PhD-trajectory started at the Department of Research Methodology, Measurement and Data Analysis (OMD) at the University of Twente. Now the final product of my work is complete and I am deeply appreciative for the opportunity that I was granted in writing this thesis and all the support that I received throughout the process. This doctoral thesis would not have been possible without it.

I owe many thanks to my supervisors. To Cees Glas, who enthused me for psychometrics and trained me, sharing interesting anecdotes along the way, and who foresaw the possibility of me writing a thesis before I did. To Martina Meelissen for introducing me to the "IEA family". Working together on the TIMSS and ICILS projects has been a great pleasure and I am grateful for the opportunity to write a thesis using data from these studies. To Theo Eggen for offering valuable perspectives on the combination of the technical complexities and more practical relevance of the studies, feedback that helped improve the research a lot.

It has been a pleasure to work on my research at the department of OMD: working with great colleagues at a green campus. I am especially thankful to the other PhD-students for sharing struggles, victories, and many lunch walks. And of course to the secretariat, for prioritising the social aspects within the department and the valuable advice provided regarding practical matters along the way.

A special thanks to Emmelien van der Scheer and Marieke van Geel for their willingness to join my defence as my paranymphs.

I want to thank my friends and family for their continued interest in the progress of my thesis and for all their loving support. From keeping me sane by pacing me around the running track to encouraging remarks over a cup of tea – it has all been much appreciated.

I am profoundly grateful to my parents, Jelke and Nieke, for all their love and support. *Tige tank!*

Mark, what a tremendous blessing it is to have you by my side. Getting to know you and building our life together during these past four years has been the best counterbalance to the stresses of academic life and a great amplification to all its highs.

Thank you all.

Annemiek Punter
Zwolle, November 2018

Contents

Chapter 1
Introduction

International large-scale assessments (ILSAs) play a major role in the evaluation of educational systems. Resulting from these ILSAs is the availability of high-quality data on student achievement and contextual factors, which together provide great opportunities for more theory-oriented educational effectiveness research. To ensure the validity of analyses based on these data, particularly relating to subpopulation invariance, efforts must be made to evaluate response behaviour across subpopulations of interest. In this thesis, ILSA data are used to both contribute to educational research and introduce advanced methodologies to handle validity issues. This chapter discusses the context and common themes of the studies presented in this thesis.

1.1 International Large-Scale Assessments

This thesis is concerned with psychometric modelling of data from ILSAs in education. ILSAs were set up to study educational achievement and its determinants in different countries so countries could learn from the experiences of others (Husén, 1979). This led to the establishment of the International Association for the Evaluation of Educational Achievement (IEA), which conducted the first ILSAs in 1960 (Johansson, 2016). In the past six decades many assessments were undertaken, with many of them under auspices of the IEA. The Organisation for Economic Co-operation and Development (OECD), responsible for the widely-known Programme for International Student Assessment (PISA), also contributed to the development of ILSAs by serving as a forum for international comparative educational research (Wendt, Bos, & Goy, 2011).

The ILSA projects are characterized by the assessment of both student achievement in several domains and contextual factors at the system, school, classroom and student level. The studies are sample-based with clearly defined populations based on student age or grade level and they use comparable types of instruments and procedures for the test administrations. An important goal of these projects is to report country comparisons of student achievement at the national level. Also, the achievement data

are scaled on a common scale across cycles. This enables countries to not only monitor their performance relative to other countries but also over time. In addition to the achievement tests, contextual data is collected by means of curriculum, student, teacher, school and home questionnaires. Data from these questionnaires provide additional information on the current state of the educational system, but moreover, allow for analyses on how these factors are related to student achievement. Once the international reports are released, the data is made publicly available for further research. Researchers can use the data for extensive analyses to explore issues in educational research, utilizing the rich, high quality data from standardized measures. The frequent use of ILSA data for secondary analyses is well illustrated by reviews of studies using TIMSS (Drent, Meelissen, & Van der Kleij, 2013), PIRLS (Lenkeit, Chan, Hopfenbeck, & Baird, 2015) and PISA data (Hopfenbeck et al., 2018).

1.1.1 TIMSS, PIRLS and ICILS

In this thesis, data from three such ILSAs are used: the International Computer and Information Literacy Study (ICILS), Progress in Reading and Literacy Study (PIRLS) and Trends in International Mathematics and Science Study (TIMSS).

Evolving from early ILSAs such as the First International Mathematics and Science Study (FIMSS; Husén, 1967), TIMSS has the longest tradition and is IEA's most well-known ILSA, with 49 participating countries in recent cycles (Mullis, Martin, Foy, & Hooper, 2016). Every four years it assesses students in Grade 4 and 8 on their math and science skills. TIMSS is often said to be "curriculum-based", because it uses a curriculum model consisting of the intended, implemented and attained curriculum. These three aspects represent, respectively, what students are expected to learn according to countries' curriculum policies, the actual teaching in classrooms, and, finally, the achievement level of the students and their attitudes regarding the subjects (Mullis & Martin, 2013). To measure these levels, the set of instruments consists of a curriculum, school, teacher, student and, since 2011, home questionnaire in addition to the math and science test. Up until TIMSS-2015 the test was administered using paper-based booklets, though the transition to digital assessment is planned for upcoming cycles (Mullis & Martin, 2017).

Similar to TIMSS regarding sampling design is PIRLS. Since 2001, PIRLS assesses the reading literacy of students in Grade 4. The reading test is centred around two reading

purposes: reading for literary experience and reading to acquire and use information (Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009). PIRLS also collects background data on national policies regarding reading curricula, school climate and resources, and how classroom instructions take place. Since its start, PIRLS administers a student home questionnaire to provide a more complete picture of the context in which students learn to read. Among other things, the questions pertain to homework activities, home-school involvement and parents' early literacy activities with their child. PIRLS is also transitioning from paper-based reading booklets to a computer-based assessment, with the first ePIRLS assessments conducted in 2016 (Mullis, Martin, Foy, & Hooper, 2017). PIRLS has a five-year-cycle and the main data collection of PIRLS coincided with the main data collection of TIMSS in 2011.

ICILS was first administered in 2013 and had its second edition in 2018. It assessed the computer and information literacy skills of students in the eighth grade across 21 participating countries in 2013 (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014). This study is the first ILSA that measures students' acquisition of computer and information literacy. Contrary to the curriculum based TIMSS, ICILS aims to assess applied knowledge and skills. It uses purpose-designed software for the authentic computer-based student assessment and questionnaire. To gather more contextual data, questionnaires are also administered to teachers, school ICT-coordinators, school principals and national research centres.

1.2 Validity

Although the position of ILSAs in the field of education is widely established, the relevance, outcomes and effects of ILSAs are not without dispute. In particular, the publication of the country-by-country performance rankings, also known as league tables, have received considerable scrutiny (Johansson, 2016). A raised concern is for example the potential, unintended consequence of isomorphism of educational systems as countries may become much alike in their curricula with the aim of scoring high on ILSAs (Spring, 2008). At a more fundamental level, critique has been directed at the validity of the international results themselves. For example, Kreiner and Christensen (2014) raised concerns on methodological issues concerning the scaling model in PISA, stating that the resulting country ranking is not robust. They state that the used Rasch model poorly fits the data. This, obviously, elicited additional studies and commentaries on this issue (see for example, Jehangir, 2015). From 2012 onward,

the Rasch model is no longer used as the main item response analysis model for PISA and more advanced models have replaced it (OECD, 2017).

Aspects of validity are often related to Messick (1989, p. 5), who states that "validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment". Messick's framework of consequential validity has construct validity as its cornerstone and progresses to address issues of relevance, value implications and social consequences on the level of policy making based on ILSA outcomes.

As ILSAs are cross-national measurements, the question of construct validity extends from "is the construct of interest measured?" to "is the intended construct measured the same and with equal precision across participating educational systems?" (Wendt et al., 2011). Making valid comparisons across populations and (sub)populations is fundamental for ILSA outcomes as a starting point for policy making. But this is also essential for secondary research purposes that involve modelling relations between constructs across (sub)populations. The rigorous quality control on the translation and adaptation of test materials for use in different settings, the standardized test administrations and equivalently defined populations and samples, provide support for the validity of cross-national measurement.

Nevertheless, to best investigate the validity of a cross-national measurement, the response patterns should also be studied with a measurement model. This can be done to ensure that the precautions in test construction and administration have indeed resulted in items that show the same measurement properties across groups, that is, to ensure measurement invariance. Potential differences in item response behaviour for students of equal ability can complicate inferences regarding proficiency differences between countries or specific student populations. A lack of measurement invariance, characterized by these differences in response behaviour is called differential item functioning (DIF).

The issue of valid measurement in secondary analyses of both test and questionnaire data across (sub)populations, is central to this thesis. Attention is directed at the modelling of item responses across (sub)populations, particularly at dealing with potential DIF. The framework of item response theory (IRT; see, for example, Van

der Linden, 2016) offers ways to address this question and is at the heart of the methodologies throughout this thesis. The focus on this latent modelling is done from a validity perspective to assess whether items function well across (sub)populations. But also as a way to link cognitive theory to response models, potentially leading to more insights into the cognitive functioning underlying test results.

1.3 Item Response Theory

Scaling of ILSA assessment data, as well as some scales in the contextual questionnaires, is done within the framework of IRT. In this framework, the constructs of interest (e.g. math ability) is regarded as a latent construct that cannot be directly observed but can be studied through responses to a set of items. IRT models the probability of a specific item response depending on both item and person characteristics. The use of IRT in ILSA is motivated by the opportunity to link new cycles to previous scales via knowledge on the item characteristics, and handling booklet rotation. A booklet rotation system, in which students are administered a sample of items, allows for a larger total set of items to increase domain coverage. IRT also serves the psychometric goal, already stated at the start of ILSAs, namely "to see whether some indications of the intellectual functioning behind responses to short answer tests could be deduced from an examination of the patterning of such responses for many countries" (as stated in Wendt et al., 2011). Wendt et al. provide a general overview of IRT methodology and software used in recent ILSAs.

1.3.1 Generalized Partial Credit Model

Constructed responses to test items are often scored using partial credit scoring. Also items in questionnaires are often polytomously scored. In both cases this means that the scores on an item can be indexed j ($j = 1,..., M_i$). To handle polytomous data, several IRT models are available, such as the graded response model (Samejima, 1969), the sequential model (Tutz, 1990), and the generalized partial credit model (GPCM, Muraki, 1992). Since the response curves of these models are hard to discern based on empirical data (see for instance Verhelst, Glas, & de Vries, 1997), the choice for either model is not fundamental. In this thesis the GPCM is adopted, because this model is commonly used in ILSA for free response items requiring partial credit scoring (e.g. in TIMSS; Yamamoto & Kulick, 2000).

In the GPCM, one latent variable (θ_n) is assumed to underlie response behaviour. Furthermore, a parameter related to an item's discriminating ability (α_i) is defined, which relates to the steepness of the curve, as well as response category parameters (β_i), providing information on the salience of response alternatives. It is a generalization of the partial credit model (Masters, 1982), in which the discrimination parameters for all items are constrained to be equal. In the GPCM, the probability of a student n ($n = 1, \ldots, N$) scoring in item category j on item i, denoted as $X_{nij} = 1$, is written as

$$P(X_{nij} = 1 \mid \theta_n) = \frac{\exp(j\alpha_i\theta_n - \beta_i)}{1 + \sum_{h=1}^{M_i} \exp(h\alpha_i\theta_n - \beta_{ih})}, \text{ for } j = 0,\ldots,M_i.$$

1.3.2 Multidimensionality

In the GPCM described above, the latent dimension measured by the items is assumed to be unidimensional. However, it might also be hypothesized that there are multiple dimensions addressed by the item set. To incorporate a latent structure consisting of multiple abilities, multidimensional IRT (MIRT; Reckase, 2009) models can be fitted to the data. This can both be done in a confirmatory fashion, with clear hypotheses regarding the structure formulated a priori, and in a more exploratory fashion. Multidimensional modelling can serve the objective to model underlying cognitive structure in greater detail but can also serve as a way to handle validity issues, as will be discussed later in this chapter.

In the MIRT model, the probability of a correct response depends on a multidimensional vector of ability parameters. The relative importance of these ability dimensions for the specific items is modelled by an item-specific loading for each dimension. We distinguish between within-item and between-item multidimensional models. In a between-item multidimensional model, each item pertains to only one of the latent dimensions. Compared to analysing the different scales separately, this approach offers the advantage that the test structure is explicitly taken into account and estimates for the correlation between the latent dimensions are provided by estimating the covariance matrix of the latent variables (Adams, Wilson, & Wang, 1997). In within-item dimensional models, an individual item can load on multiple dimensions. The choice for either model depends on the research objective. In this

thesis, both types of models are used: between-item models in Chapters 2 and 6, and a special application of a within-item model is used in Chapters 3 and 5.

1.3.3 Estimation procedures

IRT models can be estimated using a frequentist approach by means of maximum likelihood estimation. This approach seeks to find parameter values that make the observed data the most likely outcome. To concurrently estimate the item parameters and the mean and covariance matrix of the distribution of the person ability parameters, marginal maximum likelihood (MML; see, for example, Bock & Aitkin, 1981) can be applied. MML maximizes a likelihood function that is marginalized with respect to these ability parameters, with certain restrictions in place to ensure identification of the model. The MML framework provides a solid theoretical underpinning for tests of item and person fit and provides a well-established approach to parameter estimation problems, with the desirable properties of being unbiased estimators as sample size increases. However, highly dimensional models can lead to computational difficulties in an MML framework.

Alternatively, the models can be handled in a Bayesian framework (see, for example, Fox, 2010), which seeks to find the most probable parameter values given the data and some prior knowledge. In this framework, parameters are treated as random variables and both item and person parameters are estimated concurrently. Estimates of the parameters come from the multivariate conditional distribution of the parameters (i.e. the posterior distribution), which naturally incorporates uncertainty from one subset of random variables or parameters into inferences for another subset.

In this thesis, both frameworks are used. Motivation for either approach can come from substantive reasons but also from more practical considerations. For complex models for which MML estimation becomes especially burdensome because of the complex likelihood function, Bayesian estimation can be a useful alternative using Markov chain Monte Carlo computational procedures (MCMC, Béguin & Glas, 2001, Fox & Glas, 2001). For discussions on both frameworks see for example Albert (1992).

1.3.4 Differential Item Functioning

In the framework of IRT, the detection and modelling of DIF can be approached from several angles as is demonstrated in this thesis. One is viewing this item-

respondent interaction as related to item properties and modelling DIF using virtual item parameters that are allowed to vary across groups, or regressing item parameters on item properties, as far as information on such properties are available (Glas, 1998; Glas & Jehangir, 2014).

DIF can also be regarded as relating to differences in the multidimensional ability distributions of different student populations. Scores modelled by a unidimensional model, may actually represent a composite of abilities (Ackerman, 1992). When groups of students differ in these latent abilities, but only a single score is reported, DIF can occur (Roussos & Stout, 1996). According to the multidimensional model for DIF of Shealy and Stout (1993), DIF is due to (a) an item being sensitive to both the construct the item intends to measure and a secondary, confounding nuisance dimension, and (b) a difference exists between subpopulations on the secondary construct, given their proficiency on the targeted construct. Walker and Beretvas (2001) demonstrate the inclusion of more dimensions as intentional, contributing to a more authentic multidimensional representation of the construct of interest. They have argued that since test developers are confronted with items that show DIF for no apparent reason (Angoff, 1993), DIF methodology needs to be considered that hypothesizes substantive reasons for the occurrence of DIF beforehand, preferably in a multidimensional framework.

1.4 Research Objectives

This thesis comprises of several studies that centre around analyses on ILSA data using the framework of IRT for modelling the constructs of interest. Each study is driven by a substantive question from the field of educational research, where the unique properties of ILSA data, such as standardized measurements across countries, and the use of advanced IRT modelling may contribute in an innovative way. In addition, the studies aim to provide new approaches to study validity issues, particularly DIF. Models in this dissertation will be estimated both within a frequentist and Bayesian framework.

1.5 Outline of the Thesis

This thesis continues in Chapter 2 with the modelling of computer and information literacy based on data from ICILS-2013. The international report modelled the construct as unidimensional and showed that in most countries girls outperformed

boys in computer information literacy (Fraillon et al., 2014). In this chapter, the test data from nine European countries is modelled according to a between-item multidimensional IRT model to see if this proposed model fits the data better than the internationally reported unidimensional model. An additional focus is on gender differences in the multidimensional construct. Finally, equal test functioning across gender and across the nine countries is investigated.

In Chapters 3 and 4, attention is directed towards the role of parental involvement in student's reading literacy. The inconsistent results in the effects of parental involvement on student achievement may be caused by differences between educational systems and cultural differences, or by the great variation in the methods used to assess student achievement and parental involvement across studies. Chapter 3 describes how data from the PIRLS-2011 is used to develop a suitable psychometric framework to model country-specific differences in the multifaceted parental involvement construct, at both the item and scale level. Cultural differential item functioning (CDIF) in five constructed parental involvement components is identified using a Lagrange multiplier test statistic and modelled using (a) random item parameters, (b) fixed item parameters for strongest cases of CDIF, and (c) an application of the GPCM related to the bi-factor model (Gibbons & Hedeker, 1992).

In Chapter 4, the relation between the parental involvement components and student reading literacy is explored across a large number of countries. Student reading literacy is regressed in latent multilevel models on dimensions of parental involvement, where these dimensions were scaled using item response theory both with and without corrections for country-item interactions as described in Chapter 3. Results on both the relation between parental involvement and reading literacy, and the influences of potential CDIF are discussed.

In Chapters 5 and 6, the focus is on the functioning of the TIMSS mathematics items for second language learners, i.e. students not speaking the test language at home, and the role of item reading demand. Chapter 5 presents an IRT model that combines multiple approaches to detect and model DIF in large-scale assessments. Two generalizations of the GPCM are described which combine into the overall model. The first generalization is a bi-factor application, related to the one in Chapter 3. The second generalization is a model where item parameters are regressed on student and

item characteristics. The modelling steps are illustrated on data from the TIMSS-2015 mathematics test from four European countries, with reading demand classifications as the item properties of interest for students not speaking the test language at home.

In Chapter 6, data from the combined administration of TIMSS-2011 and PIRLS-2011 is used to further study the functioning of TIMSS mathematics items for second language learners, while controlling for their reading skills. This is done by estimating several latent regression models concurrently with response models on both the TIMSS and PIRLS response data. Contrary to research in Chapters 2, 3 and 4, where estimation is done in the framework of MML, the more complex models in Chapters 5 and 6 are estimated using a Bayesian framework.

In Chapter 7, the thesis concludes with a reflection on the results from the different studies presented in Chapters 2 to 6. Although some chapters are related, each chapter is written so that it can be read independently of other chapters. Consequently, some overlap between the chapters may be present.

Chapter 2
Gender Differences in Computer and Information Literacy: An Exploration of the Performances of Girls and Boys in ICILS-2013[1]

Abstract

IEA's International Computer and Information Literacy Study (ICILS) 2013 showed that in the majority of the participating countries, 14-year-old girls outperformed boys in Computer and Information Literacy (CIL): results that seem to contrast with the common view of boys as having better computer skills. This study used the ICILS data to explore whether the achievement test used in this study addressed specific dimensions of CIL and if so, whether the performances of girls and boys on these subscales differ. We investigated the hypothesis that gender differences in performance on computer literacy items would be slightly in favour of boys, whereas gender differences in performance on information literacy items would be slightly in favour of girls. Furthermore, it was examined whether such differences varied across European countries and if item bias was present. Data was analysed using a confirmatory factor analysis model, i.e., a multidimensional item response theory model, for the identification of the subscales, the explorations of gender and national differences, and possible item bias. To a large extent the results support our postulated hypothesis and shed new light on the commonly assumed disadvantaged position of girls and women in modern information society.

---

[1]Based on Punter, R. A., Meelissen, M. R. M., & Glas, C. A. W. (2016). Gender differences in computer and information literacy: An exploration of the performances of girls and boys in ICILS 2013. *European Educational Research Journal, 16*(6), 762-768.

2.1 Introduction

In the 1980s and 1990s, the low participation of girls and women in computer science courses and computer-related professions, as well as the implementation of educational computer use in primary and secondary education, resulted in many studies exploring the differences between girls and boys in computer access, use, abilities and attitudes (Volman & Van Eck, 2001; Cooper, 2006; Meelissen, 2008). The research in this area mainly focused on the perceived gender gap in computer attitudes such as liking computers, perceived usefulness of computers, self-confidence in computer use, and anxiety in using computers.

Several theories attempted to explain the disadvantages of girls in computer attitudes and competencies. For instance, the 'socialization theory' is based on the assumption that girls and boys are taught by their environment (parents, peers, media and school) to value computers differently. Computers are assumed to be unattractive to females because of computers' 'male image' caused by a past association with mathematics, science and technology (e.g. Charlton, 1999). As a result of the perceived masculinity of computers, boys feel more encouraged to explore various uses of computers, thereby increasing their knowledge and confidence (Ertl & Helling, 2011). Research has shown, for example, that girls' computer use is often limited to schoolwork while boys tend to also use computers much more for leisure activities (BECTA, 2008).

The 'attribution theory' was used to describe another effect of the perceived masculinity of computers (Volman, 1997). While working with computers, girls tend to blame themselves for mistakes and attribute success to external causes, such as the simplicity of the task or luck (the 'outsider repertoire'). Boys tend to find external causes in the case of failure and boast about their successes in using computers (the 'expert repertoire'). This gender-specific behaviour could also explain gender differences often found in self-efficacy; that is, students' own assessment of their success in performing computer-related tasks. Girls feel less confident about their computer competencies and tend to underestimate their abilities, while boys tend to overestimate their achievements (Meelissen, 2008).

According to some scholars, teachers also had a role in this confidence gap (e.g. Janssen Reinen & Plomp, 1993; Volman & Van Eck, 2001). There were not enough computer-literate female teachers who could function as role models for girls, and it

was assumed that teachers were not aware of their sometimes gender-biased instruction. For example, Volman (1994) observed how secondary school teachers in The Netherlands were often inclined to help girls by demonstrating how to perform the computer tasks. Boys, on the other hand, were often encouraged to find out for themselves and, as a result, became more confident in their abilities.

Most studies from the 1980s and 1990s confirmed the gender gap in attitudes and (perceived) competencies, especially among secondary school students (Volman & Van Eck, 2001; Cooper, 2006). For example, in 1992, the Computers in Education (COMPED) study showed that in most participating countries, boys outperformed girls in functional knowledge and skills in information technology, in primary, lower secondary and upper secondary schools (Janssen Reinen & Plomp, 1993).

However, when new uses of computers such as the Internet became available, the gender gap seemed to lessen, although this gap lessened least for females' participation in computer science courses and computer related professions (Lau & Yuen, 2015). Based on a review of studies between 1995 and 2007, Meelissen (2008) concluded that the disadvantage of girls in terms of computer attitudes had become less self-evident. Not all studies showed significant gender differences in attitudes. Where differences were found, girls often did not show negative attitudes, but showed (slightly) less positive attitudes towards computers. However, the results were inconclusive because the diversity of the scales used to measure attitudes made the comparison of research results difficult. The new opportunities that ICT had to offer, made it more complex to define 'computer use', and to measure computer attitudes and computer competencies. Furthermore, very few studies in that period focused on measuring actual computer competencies of students in relation to gender and the few studies that were conducted showed no gender differences (Meelissen, 2008; Kuhlemeier & Hemker, 2007; Hargittai & Shafer, 2006).

In today's society in which the Internet and the social use of smartphones and tablets are part of students' everyday life, it has become even more doubtful if computers are still perceived as a 'male domain' and if girls are still less confident and less experienced in using ICT (Tømte, 2011; Wong & Cheung, 2012). Furthermore, information literacy has become a very important part of computer competencies. Searching, evaluating and processing information is closely connected with reading

literacy skills (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014). The international large-scale assessment studies PISA (Programme for International Student Assessment) and PIRLS (Progress in International Reading and Literacy Study) showed that in almost every part of the world, girls outperform boys in reading literacy (OECD, 2010; Mullis, Martin, Foy, & Drucker, 2012). In PISA-2009 for example, 15-year-old girls outperformed boys in every participating country by roughly the equivalent of an average school year's progress (OECD, 2010).

Some recent studies have confirmed that the disadvantage of girls in computer attitudes and computer competencies is disappearing. Some research results even indicate that females now have more positive computer attitudes than males. For example, a recent study in the US measuring computer attitudes among eighth-grade students found that girls were more positive about computers than boys were (Hohlfeld, Ritzhaupt, & Barron, 2013). In this study, the same four-item 'Attitude Towards Computers' self-report scale was used as in PISA-2009. In PISA-2009, however, the computer attitude of 15-year-old boys was, according to this scale, still more positive than that of girls in all European countries participating in PISA but Spain (OECD, 2011). In the study of Hohlfeld et al. (2013) girls also rated their ICT-skills higher than boys. A study among Taiwanese Grade 8 students showed no gender differences in students' self-efficacy in using the Internet, but girls were more positive about their self-efficacy in online communication than boys were (Tsai & Tsai, 2010). The gender differences found in their study were related to the type of ICT use: boys were more exploration-oriented Internet users and girls more communication-oriented Internet users.

Studies reporting an advantage of girls are not limited to computer attitudes or self-efficacy in computer use. Nowadays, as technologically more advanced testing is possible, more studies use performance-based digital tests consisting of test items in a genuine, virtual test environment, which also resulted in a shift from testing 'knowing of' to 'showing how'. The first international large-scale assessment study using a performance-based digital test was IEA's International Computer and Information Literacy Study (ICILS) 2013 (Fraillon et al., 2014). Computer and information literacy (CIL) is defined as: "an individual's ability to use computers to investigate, create, and communicate in order to participate effectively at home, at school, in the workplace, and in society" (Fraillon, Schulz, & Ainley, 2013, p.17). It turned out that in most of

the 21 countries or regions, 14-year-old girls significantly outperformed boys in ICILS-2013. In line with the ICILS results, Flemish sixth-grade girls also outperformed their male classmates in both technical ICT skills, such as retrieving a file from a specific location or open an attachment, and so-called higher-order ICT competencies, such as delivering information in an email, in a computer-based assessment (Aesaert & Van Braak, 2015).

However, the conclusion that the traditional gender gap in computing has reversed may be premature, as some studies still report no gender differences or show differences that are still in favour of boys. In contrast to the studies mentioned above, no gender differences in digital competencies (that is, digital judgements, acquiring and processing digital information, and producing digital information), were found in a Norwegian study among upper secondary students (Hatlevik & Christophersen, 2013). Secondary school girls and boys in the Netherlands of various age groups also showed the same level of ability in their information and strategic Internet skills (Van Deursen & Van Diepen, 2013). Gui and Argentin (2011) assessed Italian secondary students' digital literacy by developing an assessment covering the following three areas: (1) theoretical skills, including answering knowledge-based questions; (2) operational skills, the ability to use computer applications and navigate efficiently; and (3) evaluation skills, the skills in information evaluation practices. The test in this study showed no differences between girls' and boys' performance in operational and evaluation skills. However, girls were outperformed by boys in theoretical skills. Gui and Argentin (2011) concluded that female students are as skilled as male students in common online activities, but might experience difficulty when confronted with unexpected technical problems or outcomes.

Evidence that the gender gap has not fully counterbalanced can also be found within the ICILS results themselves. In ICILS-2013 no gender differences were found for basic ICT self-efficacy, but the scores of boys on the advanced ICT self-efficacy scale were higher than the girls' scores in the participating countries (Fraillon et al., 2014). These outcomes contrast with the results of the actual assessment in CIL, which was in favour of girls. The differences between the results of the advanced ICT self-efficacy scale and the CIL assessment may be explained by the attribution theory mentioned earlier. In the case of advanced computer skills, girls may still tend to underestimate their abilities while boys may tend to overestimate their abilities. It also

suggests that using a self-efficacy scale to measure students' computer competencies may give a misleading indication of students' real abilities in computing (Aesaert & Van Braak, 2015).

In summary, recent studies suggest that there is no longer a digital divide in favour of boys. Regarding computer use, boys and girls may have different interests, but that does not necessarily mean that these differences result in advantages or disadvantages for either girls or boys. Regarding computer attitudes, there seems to be more and more evidence that the gender gap is closing. However, regarding self-efficacy in ICT use and ICT-competencies the results remain mixed.

One of the reasons for these mixed results may be the use of self-reported self-efficacy scales and assessments in such studies. The concept of 'computer' has become much more complicated due the many different uses of and devices for ICT, compared to the 1980s when gender differences concerning computers were first researched. As a consequence, studies in this area come with a great variety of ways to name, define and measure computer competencies (Ilomäki, Paavola, Lakkala, & Kantosala, 2016). This makes it difficult to compare studies and to draw general conclusions about the role of gender in (perceived) computer competencies, especially conclusions across countries. Furthermore, there is often limited information available about the validity of the instruments, such as the possible gender bias in the test items. For example, some test items may function differently for girls and boys, which complicates the comparisons of results between genders.

If measurement issues are handled properly and there is no sign of differential item functioning, international comparative research can reveal systematic differences in computer and information literary and provide valuable pointers to look where these differences originate from. The differences might be explained by cultural, economic and/or educational differences, potentially resulting in different experiences of students with ICT and therefore varying levels of competencies. Signalling these cross-country differences serves as an important step for the evaluation of educational systems.

This study focuses on the relation between gender and computer competencies. As a result of its scale, its extensive conceptualisation of computer competencies and the opportunity to model across countries, the assessment data of ICILS-2013 was used

for our explorations. Although the CIL test was initially based on two dimensions (strands): (a) collecting and managing information; and (b) producing and exchanging information, the results were only reported on a single, unidimensional scale (Fraillon et al., 2014). The results showed a high correlation between the observed scores on these two dimensions. Furthermore, the mean achievement of students across countries varied little when data from both dimensions were analysed separately.

In this study we propose an alternative classification of the ICILS test items based on a content review by experts. Next, we will compare the performances of girls and boys in European countries on these three new dimensions. The dimensions will be entered as three latent factors in a confirmatory factor analysis (see, for example Muthén & Muthén, 1998-2012), or, completely analogously, as three latent variables in a multidimensional IRT model (see for instance, Reckase, 2009).

In line with the study of Gui and Argentin (2011) we assume that the gender gap differs for different dimensions of CIL. As girls outperformed boys in the overall CIL achievement in ICILS-2013, we are specifically interested to see in which dimensions the gender differences are most prominent. It is hypothesized that boys will have a disadvantage in items referring to competencies such as evaluating and sharing information, but an advantage in items measuring technical skills.

As a last step, item bias, i.e. differential item functioning by culture and gender, is investigated to further evaluate the validity of the IRT model.

The four research questions are:

1. Is a three-dimensional representation of computer and information literacy appropriate for the ICILS data, i.e., to what extent does the data fit a three-dimensional IRT measurement model in terms of model fit, correlation structure, and item loadings?
2. In which dimensions of computer and information literacy are gender differences most prominent?
3. To what extent are these differences consistent across European countries participating in ICILS-2013?
4. To what extent is the validity of the multidimensional IRT measurement model threatened by gender-related item bias and cultural item bias?

2.2 Methodology

2.2.1 Dataset and Item Classification

By assessing 14-year old students from a representative sample in 21 countries, ICILS-2013 provides representative data within and across countries (Fraillon et al., 2014). Between 138 and 318 schools were randomly selected in each country. Twenty students were then randomly selected from all students in the target grade (usually Grade 8) in each sampled school.

Four computer-based modules were developed in ICILS. Each 30-minute module had a theme (for example, organising a school trip) and consisted of a number of small discrete tasks or questions followed by a large final task. The modules comprised 62 tasks and questions. Some allowed for dichotomous scoring (0 score points for no credit, 1 for full credit); others allowed for partial credit scoring (0 score points for no credit, 1 for partial credit, 2 for full credit). The test modules combined comprised 81 score points. The items were administered according to a balanced module rotation, meaning each student completed two modules randomly allocated from the set of four computer-based modules.

In each country the responses were coded by trained scorers. To assess the reliability of scoring, 20 percent of the responses were scored independently by two scores. Items with too low reliability (i.e. below 75 percent) were left out of further analyses.

This study explores the data from all European countries in ICILS-2013 with a response rate above 80 percent at school level. This high level of response of the representative sample ensures representative findings for the population of 14-year-olds in regular education. The average achievement test scores and the extent of gender differences of the countries are presented in Table 2.1.

Table 2.1
Achievement Test Scores for the Selected Countries in this Study

Country	Mean scale score (*SE*)	Gender differences in favour of girls (*SE*)
Czech Republic	553 (2.1)	12 (2.7)
Poland	537 (2.4)	13 (3.7)
Norway	537 (2.4)	23 (3.5)
Netherlands	535 (4.7)	20 (4.9)
Germany	523 (2.4)	16 (3.8)
Slovak Republic	517 (4.6)	13 (4.1)
Croatia	512 (2.9)	15 (3.5)
Slovenia	511 (2.2)	29 (3.6)
Lithuania	494 (3.6)	17 (3.4)

Note. Test scores are on the international scale with a mean of 500 and a standard deviation of 100 (Fraillon, Schulz, Friedman, Ainley, & Gebhardt, 2015).

Only items with satisfying scaling properties for all European countries according to the study's technical report (Fraillon, Schulz, Friedman, Ainley, & Gebhardt, 2015) were considered, resulting in a dataset of 45 items and a sample size of 25133 students. Although, based on literature, it was expected to find dimensions strongly related to the computer literacy and information literacy, there were no dimensions specified at forehand. A content review of the items by two experts resulted in several proposed categorizations of the items into multiple dimensions. Based on the experts' further consultation, the three-dimensional categorization was considered most suitable. The experts then assigned each item to one of the three dimensions with respect to its relevance, resulting in complete agreement about the categorization of the items according to the three dimensions. These dimensions can be illustrated by the following example. Suppose a student wants to send out a birthday invitation to his friends by email. An essential step is to know how to login to an email account, where to place the email addresses of the intended guests, where to put the text, how to import illustrations, and so on. These skills belong to the first dimension: applying technical functionality. The first dimension relates closely to the more traditional "computer literacy" as it entails knowing "which buttons to push". Knowing what information needs to be in the invitation email and putting email addresses in "bcc" to prevent the unnecessary sharing of addresses, are elements that show some reflection on the information the student uses or produces and relate to the second dimension: evaluating and reflecting on information. This second dimension relates closely to the more traditional "information literacy". The third dimension, sharing or communicating information, refers to preparing an information product. In the example, this could be choosing a suitable font size, colour and pictures to make the invitation inviting. This third dimension was strongly incorporated in the large tasks at

the end of each module in the ICILS test, where an information product had to be created, for example a poster. The three dimensions of CIL are further described in Table 2.2. Of the 45 items, 14 relate to the first dimension, 19 to the second and 12 to the third dimension.

Table 2.2
Three-Dimensional Structure of Computer and Information Literacy

Dimension	Description	Example in ICILS test
Applying technical functionality	Know how to get something done using the technology	Navigate to a website using a link provided in an email
Evaluating and reflecting on information	Evaluate information, reflect on using and sharing information	Indicate how an element of a potential trick email shows that the email may be a phishing email.
Sharing or communicating information	Prepare an information output product / share information	Choose an appropriate layout of text and images for an informative poster.

2.2.2 Model Estimation and Model Fit

A multidimensional IRT model, i.e. a confirmatory factor analysis model, was used to validate the proposed three-dimensional structure of the ICILS test data (research question 1). To justify the use of this rather extensive IRT model, the model was first compared to more simple IRT models: the unidimensional partial credit model (PCM; Masters, 1982) and the unidimensional generalized partial credit model (GPCM; Muraki, 1992). The PCM is the most parsimonious model and was also used by the ICILS consortium (Fraillon et al., 2015). The model assumes that only one latent variable is needed to explain response behaviour. The GPCM is an extended version of the PCM, which is also unidimensional but includes not only item location parameters β to characterize the overall score level of an item, but also an item discrimination parameter α which represents the extent to which the item correlates with the latent variable. In the terminology of factor analysis, this parameter represents a factor loading. In the three-dimensional GPCM, it is assumed that three correlated latent variables are needed to explain response behaviour. Comparisons were made based on the log-likelihood, AIC and BIC fit indices, with smaller values of the indices indicating a better model fit. More complex models result in lower values of these indices. However, they do, obviously, come at the cost of more complexity. Therefore, a more complex model (say, the GPCM) is only preferred over the simpler model (say, the PCM) when the difference in fit indices is judged substantial.

The models were estimated across the nine European countries with gender groups within each country (research questions 2 and 3). The models gave insight into the distribution of the respondents on the latent variables. That is, the estimates included the mean achievement scores for boys and girls in the countries of interest, the correlation structure of the three latent dimensions and the extent to which the items loaded on their specific dimensions.

Finally, two methods of investigation were applied to evaluate gender-related item bias and item bias across countries (research question 4): one based on the difference between observed mean item scores in the gender and country groups and their expected values under the three-dimensional GPCM; the other on comparing parameter estimates obtained for the subgroups.

Parameter estimation and evaluation of model fit was done in the framework of marginal maximum likelihood (MML; Bock, Gibbons, & Muraki, 1988) using the public domain software package MIRT/LEXTER (Glas, 2010).

2.3 Results

A model comparison based on fit indices is presented in Table 2.3 and provides a first indication of the appropriateness of a three-dimensional IRT model (research question 1). The first row pertains to a comparison of the PCM with the GPCM. Both models were estimated with a distinct normal latent variable distribution for each combination of gender and country. That is, every gender group within each country had a specific mean and variance on the latent scale. The item parameters were the same for all countries. The columns labelled 'df' and '-2LL' give the degrees of freedom and the value of the likelihood-ratio statistic, respectively. The AIC and BIC statistics are based on the likelihood-ratio statistic, but they penalize over-parameterized models and large sample sizes. Note that the value of the likelihood-ratio statistic (11657374 with 44 degrees of freedom) is highly significant and the PCM is clearly rejected in favour of the GPCM. The AIC and BIC do not substantially lower the value of the likelihood-ratio statistic, so the conclusion is not altered. From the first row of Table 2.3, it thus becomes clear that the GPCM provides a significantly better fit to the data than the PCM. The second row pertains to testing the unidimensional GPCM against the multidimensional version. Again, the simpler model was rejected, although the differences in fit indices were smaller than in the case of testing the PCM against the

GPCM. As such, this does not mean that the three-dimensional GPCM fits the data perfectly. In a subsequent testing step, an important aspect of the model will be evaluated, namely the absence of item bias. However, first the parameter estimates of the three-dimensional GPCM are presented and discussed.

Table 2.3
Differences in Fit Indices for Model Comparisons (N = 25133)

Compared Models	Δ *df*	Δ -2LL	Δ AIC	Δ BIC
1-dim GPCM vs. 1-dim PCM	44	11657374	11657284	11656918
3-dim GPCM vs. 1-dim GPCM	122	1358490	1358246	1357254

Note. GPCM refers to the generalized partial credit model, PCM to the partial credit model.

MML entails the concurrent estimation of three sets of parameters: the correlation structure, the means of the gender groups within countries and the item parameters. These estimates will be presented in that order, starting with the correlation structure between the three latent dimensions. The correlations between the latent dimensions and the respective standard errors are presented in Table 2.4 for each combination of a country and a gender group. Within countries, differences that are significant at the 1% level are marked with an asterisk. For the majority of the countries no clear country or gender effects on the correlation structure were manifest. From the table it becomes clear that in each country Dimensions 2 and 3 have the highest correlation, whereas Dimension 1 correlates the least with Dimension 3. The dimensions correlate most strongly for Slovenia (estimates range from .656 to .982) and the weakest for Norway (estimates range from .636 to .869). It is notable that the correlation between Dimensions 2 and 3 is extremely high in a number of countries. For Croatia, Germany and the Netherlands there is also a significant difference in this correlation between boys and girls. Significant gender differences are found for all correlations for Germany, with girls showing higher correlations between the dimensions.

Table 2.4
Correlations and Standard Errors Between the CIL Dimensions for Boys and Girls Across Countries

Country	Gender	*N*	Correlation			*SE*		
			Dimensions			Dimensions		
			1,2	1,3	2,3	1,2	1,3	2,3
Croatia	Boys	1447	0.737	0.627	0.989*	0.036	0.030	0.012
	Girls	1399	0.752	0.703	0.932*	0.029	0.024	0.016
Czech Republic	Boys	1507	0.714	0.713	0.974	0.029	0.024	0.013
	Girls	1554	0.780	0.678	0.980	0.027	0.026	0.014
Germany	Boys	1127	0.693*	0.642*	0.878*	0.030	0.033	0.018
	Girls	1098	0.773*	0.733*	0.942*	0.025	0.024	0.013
Lithuania	Boys	1412	0.715	0.620*	0.965	0.031	0.031	0.017
	Girls	1344	0.798	0.740*	0.996	0.035	0.026	0.005
Netherlands	Boys	1149	0.767	0.638	0.879*	0.021	0.027	0.013
	Girls	1045	0.776	0.674	0.936*	0.026	0.025	0.012
Norway	Boys	1212	0.772	0.693	0.844	0.031	0.033	0.020
	Girls	1219	0.725	0.636	0.869	0.035	0.036	0.016
Poland	Boys	1500	0.792	0.681	0.854	0.022	0.024	0.017
	Girls	1370	0.835	0.710	0.875	0.026	0.026	0.016
Slovak Republic	Boys	1516	0.779	0.671	0.906	0.023	0.022	0.011
	Girls	1477	0.779	0.700	0.922	0.021	0.021	0.012
Slovenia	Boys	1925	0.786	0.656	0.980	0.020	0.023	0.010
	Girls	1814	0.797	0.676	0.982	0.027	0.027	0.014

Note. Dimension 1 pertains to applying technical functionality, Dimension 2 to evaluating and reflecting on information, and Dimension 3 to sharing or communicating information. * $p < .01$.

Table 2.5 shows the estimates of the means on the scales for the three CIL dimensions and the sample size of each gender group within a country and addresses the second and third research question. Note that data from Slovenian girls were used to identify the origin of the subscales, with the mean of the latent scale fixed to zero and the variance fixed to 1, so they serve as a reference group. For all other combinations of gender and country, the standard deviations were free and varied between 1.051 and 2.184. The gender differences are the largest and significant across all countries for Dimension 3. Therefore, on sharing or communicating information girls outperform boys. Furthermore, girls outperform boys significantly in evaluating and reflecting on information (Dimension 2) in all countries. On Dimension 1, applying technical functionality, significant gender differences in favour of girls were found in five of the nine countries under review.

Table 2.5
Estimated Means and Standard Errors on Latent Dimensions of CIL for Boys and Girls Across Countries

Country	Gender	*N*	Mean			*SE*		
			Dimensions			Dimensions		
			1	2	3	1	2	3
Croatia	Boys	1447	-0.093*	-0.278*	-0.258*	0.063	0.038	0.042
	Girls	1399	0.140*	0.077*	0.114*	0.058	0.038	0.044
Czech Republic	Boys	1507	0.548	0.786*	0.792*	0.060	0.048	0.034
	Girls	1554	0.597	1.024*	1.063*	0.060	0.045	0.036
Germany	Boys	1127	0.603	0.295*	-0.054*	0.073	0.056	0.042
	Girls	1098	0.734	0.577*	0.132*	0.073	0.055	0.042
Lithuania	Boys	1412	-0.524*	-0.327*	-0.825*	0.053	0.040	0.037
	Girls	1344	-0.307*	-0.063*	-0.591*	0.049	0.042	0.032
Netherlands	Boys	1149	1.613*	0.261*	0.073*	0.076	0.056	0.053
	Girls	1045	1.890*	0.712*	0.655*	0.076	0.062	0.058
Norway	Boys	1212	0.986	0.163*	-0.087*	0.062	0.049	0.045
	Girls	1219	0.968	0.578*	0.535*	0.061	0.051	0.046
Poland	Boys	1500	0.376	0.285*	0.232*	0.066	0.046	0.044
	Girls	1370	0.336	0.664*	0.651*	0.062	0.047	0.049
Slovak Republic	Boys	1516	0.138*	-0.011*	0.031*	0.063	0.048	0.050
	Girls	1477	0.321*	0.207*	0.338*	0.064	0.046	0.048
Slovenia	Boys	1925	-0.457*	-0.505*	-0.424*	0.046	0.036	0.030
	Girls	1814	0.000*	0.000*	0.000*	0.000	0.000	0.000

Note. Dimension 1 pertains to applying technical functionality, Dimension 2 to evaluating and reflecting on information, and Dimension 3 to sharing or communicating information. Scores are on scales with means fixed to zero and variances fixed to 1 for the reference group (Slovenian girls). * $p < .01$.

Table 2.6 gives the estimates of the item parameters, together with the sizes of the sample that responded to each item. The items were administered according to a module rotation design, so these sample sizes are much lower than the total sample size of 25133 students. The item parameters provided in Table 2.6 are grouped per dimension, i.e. per latent scale. The discrimination parameters α indicate how well an item discriminates with respect to the overall CIL proficiency, where a high value also tends to indicate a higher information value for the item. These parameters can also be interpreted as factor loadings, that is, as coefficients of the regression of the item score on the latent scale. The final column for each scale gives the average of the β parameters for polytomous items, also called the item difficulty parameter. For polytomously scored items, the number of β parameters is equal to the number of score categories minus one, and their average can serve as an indication of the (average) score level on the item. Together, the α and average β parameters give an

indication of the contribution of the items to the reliability of a scale: the higher the α parameter and the closer the β is to the mean of a group on the latent scale, the greater the contribution.

In general, the discrimination parameters for items measuring Dimension 1, applying technical functionality, are the smallest and α estimates for Dimension 3, sharing or communicating information, the largest. This indicates that items for Dimension 3 discriminate best with respect to the overall CIL proficiency and are more likely to show a high item information value, whereas Dimension 1 discriminates the least. Looking at the item difficulty parameters, it becomes clear that, on average, Dimension 1 contains items with the lowest β estimates, indicating more easy items. For Dimension 2, evaluating and reflecting on information, the average of the β parameters is the highest, indicating that overall this dimension contains harder items.

At the item level, Table 2.6 shows that for the first dimension, item H03ZA has the highest α estimate. This item asks students to open the internet browser application from the taskbar. For the second dimension, the items B09EL and S08BL show the highest α estimates. These items require students to include the specified information in a webpage and to find required information on website pages, respectively. The largest α value within the third dimension, as well as overall, is item B09GL where the task is to create a balanced layout for a webpage. Examples of items with relatively low α estimates are items A01ZMZ, and A05ZA; identify who received an email by "cc" and to open webmail from a hyperlink. Item A05ZA also has a low β estimate, indicating that the item is easy. The hardest item in the test, showing a β estimate of 6.77, is item B04ZM, which asks the student to select a suitable website navigation structure for a given content on a webpage.

Table 2.6

Parameter Estimates in the Three-Dimensional Generalized Partial Credit Model for Items According to Dimension

Dimension 1: Applying technical functionality					Dimension 2: Evaluating and reflecting on information					Dimension 3: Sharing or communicating information				
Item	N	M_i	α	$\overline{\beta}$	Item	N	M_i	α	$\overline{\beta}$	Item	N	M_i	α	$\overline{\beta}$
A01ZMZ	12543	4	0.26	-4.84	A03ZM	12441	1	0.66	-4.32	A10AL	12057	2	0.63	-0.91
A02ZA	9727	1	0.36	-1.83	A06AC	11428	1	0.50	3.28	A10BF	12057	3	0.14	0.05
A05ZA	11867	1	0.27	-8.86	A06BC	10877	1	0.47	2.28	A10DL	12057	2	0.47	-0.24
A08ZA	12002	1	0.61	-9.28	A06CC	10556	1	0.46	4.88	A10JL	12057	1	0.61	-0.18
A09ZA	12537	2	0.34	-0.55	A07ZC	12215	1	0.51	-0.58	B09AL	12314	2	0.83	0.37
H01ZA	10325	1	0.57	-1.86	A10HL	12057	2	0.49	0.24	B09BL	12314	2	0.90	-0.88
H02ZA	12542	2	0.41	1.33	H05ZC	12138	1	0.41	3.82	B09FL	12314	2	1.13	0.41
H03ZA	12270	1	0.88	-1.97	H06ZC	12217	1	0.47	1.77	B09GL	12314	1	1.54	-0.45
B03ZA	8688	1	0.54	-2.09	H07HL	11529	2	0.37	-1.17	S08AL	11727	1	0.95	-0.80
B06ZA	11704	1	0.34	-8.23	H07IL	11529	1	0.42	-0.38	S08DL	11727	3	0.36	0.95
B07BA	12526	1	0.43	-6.65	B01ZC	12023	1	0.37	-3.88	S08EL	11727	2	0.72	1.16
B07CA	12526	1	0.42	-1.84	B02ZC	11878	2	0.26	-2.78	S08GL	11727	2	0.44	-0.04
S03ZA	8734	1	0.26	-2.43	B04ZM	11612	1	0.27	6.77					
S04BA	12386	1	0.47	-1.27	B05ZA	12210	1	0.48	0.14					
					B08ZZ	12422	5	0.28	-2.77					
					B09EL	12314	1	1.53	-0.69					
					S04AA	12386	1	0.58	1.68					
					S08BL	11727	1	1.22	0.05					
					S08CL	11727	2	0.63	-0.10					

Note. The item names correspond to the names used in the international databases of ICILS-2013 (available at www.iea.nl). N indicates the number of responses to the item, M_i to the maximum score points for the item.

The investigation of the fit of the three-dimensional model continued by evaluating the presence of item bias, that is, differential item functioning across gender and countries (research question 4). The motivation was that such items might confound the comparisons of boys and girls and of countries. Item bias was investigated by applying two methods. The first method was based on residuals that are calculated as the differences between observed mean item scores in subgroups, such as gender groups and countries, and their expected values under the three-dimensional GPCM. The second method was by comparing parameter estimates obtained in gender and country subgroups. The first method exists in many forms; the present application is based on the version by Glas and Jehangir (2014) and implemented in the MIRT/LEXTER software package (Glas, 2010). This method was applied in two forms. In the first form, the residuals were calculated for every combination of a gender group and a country, and their absolute values were then averaged over these 18 subgroups. For each dimension, these residuals are tabulated in Table 2.7 under the label 'R1'. High values in this column suggest an item functions differently for one or more groups of boys or girls in a country, that is, that there is item bias. In the second form, the residuals within these 18 subgroups were broken down further to create subgroups of respondents with low, medium and high total scores, resulting in a total of 54 subgroups. These residuals are summarized in Table 2.7 under the label 'R2' and can indicate if there is differential item functioning for subgroups of boys and girls. For example, whether certain items function differently for high achieving girls in particular. The residuals are scaled to lie between 0.0 and 1.0; 0.10 is often used as a critical value. Note that only a few items exceeded this critical value, so the overall conclusion based on this analysis is that item bias was not prevalent in this data set. Results from estimating the three-dimensional model excluding items with residuals larger than .10 showed no remarkable differences in group means or correlational structure. This further supports the conclusion that results were not influenced by cultural or gender item bias, based on this analysis.

Item bias was also investigated by a second method based on comparing parameter estimates obtained in gender and country subgroups. Therefore, the three-dimensional GPCM was estimated separately for each country under review to obtain an indication of the extent to which the parameters estimates were country-specific. The estimates for the Czech Republic and Slovenia showed the least agreement (Figure 2.1), with a correlation of .74. The strongest agreement was between estimates for Norway and

Table 2.7
Residual Analysis for the Three-Dimensional Generalized Partial Credit Model for Items According to Dimension

Dimension 1: Applying technical functionality			Dimension 2: Evaluating and reflecting on information			Dimension 3: Sharing or communicating information		
Item	R1	R2	Item	R1	R2	Item	R1	R2
A01ZMZ	0.06	0.08	A03ZM	0.03	0.03	A10AL	0.10	0.10
A02ZA	0.05	0.06	A06AC	0.07	0.07	A10BF	0.07	0.10
A05ZA	0.02	0.02	A06BC	0.07	0.08	A10DL	0.14	0.15
A08ZA	0.00	0.00	A06CC	0.04	0.05	A10JL	0.05	0.06
A09ZA	0.12	0.14	A07ZC	0.08	0.08	B09AL	0.08	0.08
H01ZA	0.05	0.05	A10HL	0.14	0.15	B09BL	0.08	0.08
H02ZA	0.16	0.17	H05ZC	0.05	0.05	B09FL	0.09	0.10
H03ZA	0.01	0.03	H06ZC	0.11	0.11	B09GL	0.07	0.07
B03ZA	0.05	0.05	H07HL	0.10	0.11	S08AL	0.05	0.06
B06ZA	0.01	0.02	H07IL	0.04	0.05	S08DL	0.17	0.18
B07BA	0.02	0.02	B01ZC	0.05	0.05	S08EL	0.09	0.10
B07CA	0.06	0.07	B02ZC	0.07	0.08	S08GL	0.13	0.14
S03ZA	0.10	0.11	B04ZM	0.02	0.06			
S04BA	0.03	0.04	B05ZA	0.04	0.05			
			B08ZZ	0.16	0.18			
			B09EL	0.04	0.04			
			S04AA	0.05	0.06			
			S08BL	0.03	0.04			
			S08CL	0.03	0.05			

Note. The item names correspond to the names used in the international databases of ICILS-2013 (available at www.iea.nl). Item and response category labels are provided in the international databases. R1 refers to the average of the residuals over all gender-by-country subgroups. R2 refers to the average of the residuals over low, medium and high scoring groups per gender per country.

Slovak Republic (see Figure 2.2, correlation .97). However, the overall correspondence provides no support for strong cultural differential item functioning and therefore supports the use of the three-dimensional GPCM across the nine countries to make country comparisons.

To assess the extent to which the parameter estimates for the three-dimensional GPCM differ for boys and girls within each country, the model was estimated for boys and girls country by country. A graphical representation of the estimates for boys and girls indicated that the estimates corresponded very well overall, with the greatest correspondence in Poland (Figure 2.3; correlation of .97), but the least in the Netherlands (Figure 2.4; correlation of .82). This indicates that there is little or no gender differential item functioning in the items.

Figure 2.1
Parameter Estimates for Czech Republic and Slovenia.

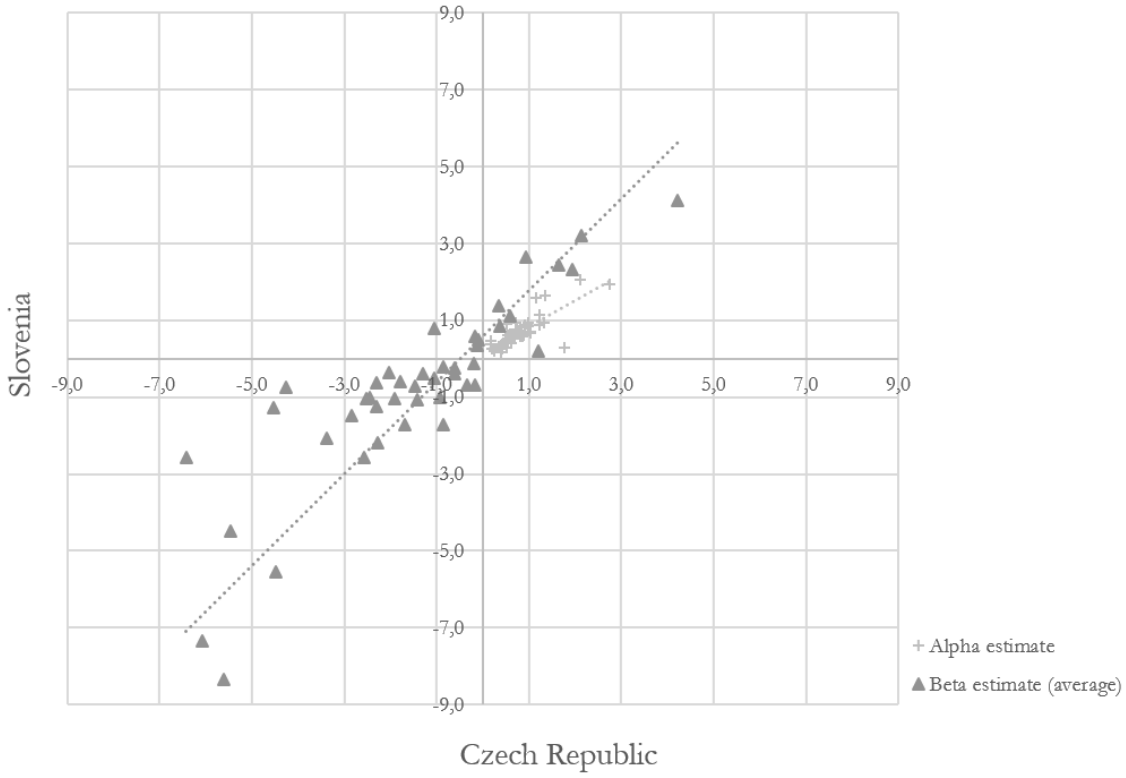


Figure 2.2
Parameter Estimates for Norway and Slovak Republic.

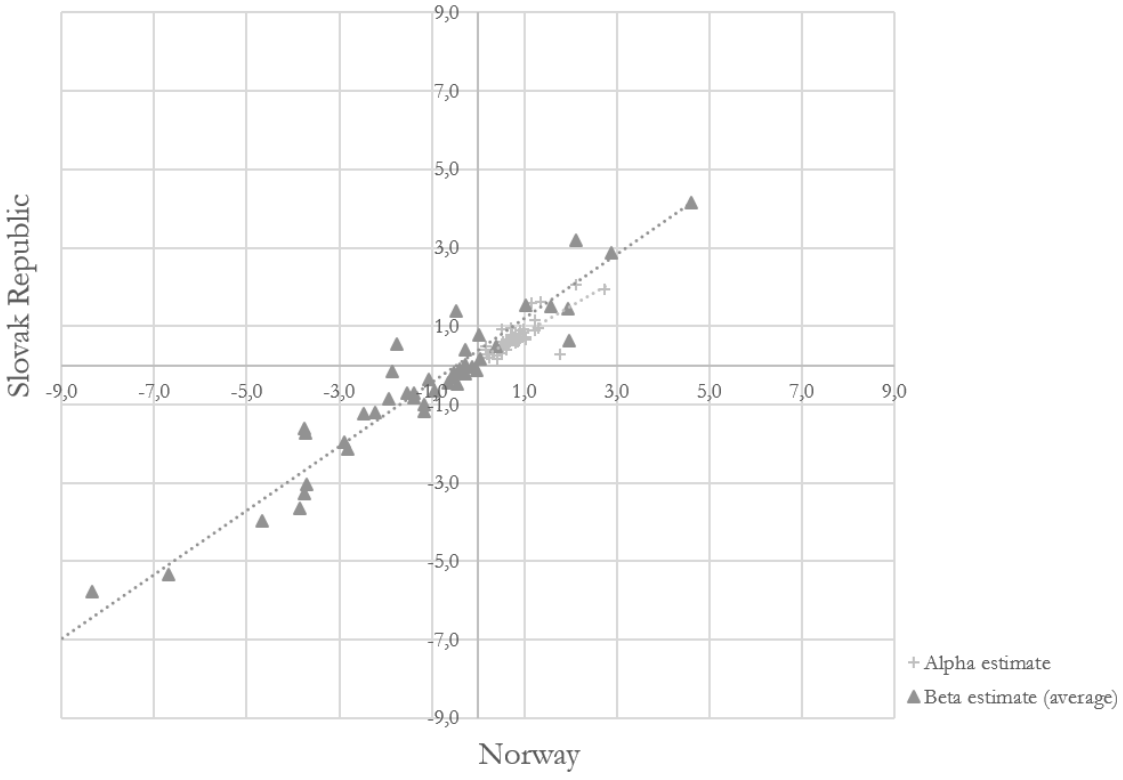


Figure 2.3
Parameter Estimates for Boys and Girls in Poland.

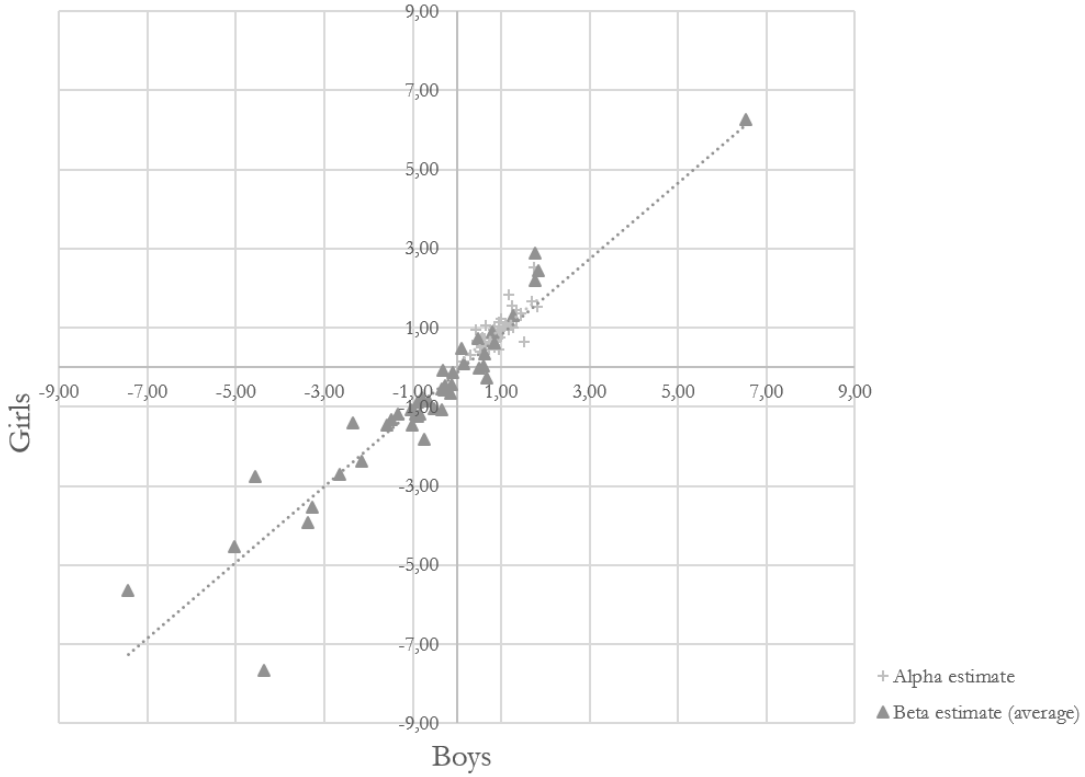
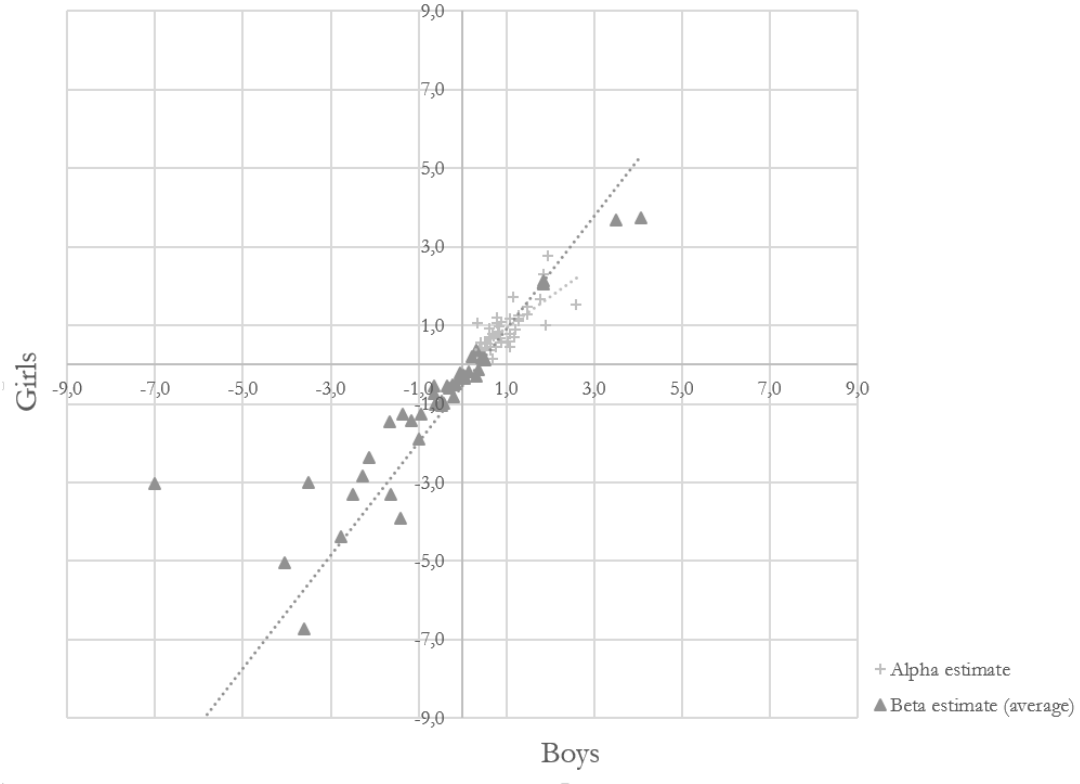


Figure 2.4
Parameter Estimates for Boys and Girls in the Netherlands.



2.4 Discussion

The main purpose of this research was to explore whether the ICILS-2013 achievement test addressed the various dimensions of CIL and if so, whether the performances of girls and boys on these subscales differed across the participating European countries. The dimensional structure proposed consisted of three dimensions: (a) applying technical functionality, (b) evaluating and reflecting on information; and (c) sharing or communicating information. Four research questions guided the study and these are discussed below.

1. *Is a three-dimensional representation of computer and information literacy appropriate for the ICILS data, i.e., to what extent does the data fit a three-dimensional IRT measurement model in terms of model fit, correlation structure, and item loadings?*

The estimated three-dimensional IRT model was found to fit the data from ICILS-2013 better than the unidimensional GPCM and PCM. The correlation structure showed high correlations between the three dimensions, and this was consistent across countries and genders. However, the clearest distinction was between Dimension 1 at the one hand and Dimensions 2 and 3 on the other, which corresponds with the distinction between the technical skills and information literacy known from the literature. The relative high correlations between the constructs can be interpreted as an affirmation of the integration of computer skills and information literacy into one competence: CIL. However, we believe multiple dimensions are still underlying this construct and assessing performance on these separate scales is useful to see where CIL differences originate from and which aspects could be targeted to remedy groups falling behind on CIL. Based on these findings, future research on the topic of CIL should consider using this dimensionality in the process of test development and scale analyses.

The third dimension, being able to make an information output product, was judged by the experts as a separate category of items in the ICILS test, not to be categorized in a two-dimensional model. However, it did show very high correlations with the second dimension. The item parameter estimates indicated that items on the scale for Dimension 3 are likely to show the highest information values for the CIL construct. Therefore, although the third dimension could not be clearly distinguished from Dimension 2, the items allocated to this scale do seem highly informative for measuring the CIL construct.

The scale for applying technical functionality, on the other hand, had the lowest discrimination parameters and contained, on average, items with low β estimates, indicating more easy items. Future research should broaden the scope of item difficulties on this scale, as the literature suggests that boys and girls may perform differently on items requiring basic ICT skills and tasks requiring higher-order ICT competencies (e.g. Aesaert and Van Braak, 2015; Gui and Argentin, 2011). The current ICILS data did not enable such a comparison.

Students scored, on average, lower on Dimension 2 than on Dimension 1, suggesting room for improvement in evaluating and reflecting on information. It can be argued that in order to prepare students for participation in the current information society, information skills need more attention in education in these European countries. For example, the ICILS-2013 teacher survey asked teachers to what extent evaluating the credibility of digital information and validating the accuracy of digital information was emphasized in their teaching (Fraillon et al., 2014). Results indicate that teacher means were below the ICILS average in the majority of European countries. The non-European countries Australia, Canada and Chile reported a much stronger emphasis in the classrooms on these topics.

2. *In which dimensions of CIL are gender differences most prominent?*

The postulated hypothesis that girls have an advantage in performance on items in the more information-oriented dimensions (Dimensions 2 and 3) and that for performance on items assessing computer literacy (Dimension 1) the advantage is reversed or even non-existent, was supported. Girls indeed outperformed boys in most countries on the second dimension (evaluating and reflecting on information) and no significant gender differences were found for Dimension 1 (applying technical functionality). The gender gap in this latter area, identified by previous research (e.g. Janssen Reinen & Plomp, 1993) therefore no longer seems to exist. However, to conclude that the gender gap in ICT is truly closed, would require further research on gender differences in affective factors; for example, by investigating self-efficacy measured in CIL, as the confidence of girls in various types of computer use is just as relevant for their academic progress and professional careers as their actual skills. Findings from the student survey in ICILS-2013 suggest the gap is not closed for these affective factors. Boys scored substantially higher on the self-efficacy scale for higher-level ICT skills (Fraillon et al., 2014).

The largest differences in favour of girls were on Dimension 3: sharing and communicating information. Although the conceptualisation of the third dimension deserves further attention, since items in this dimension were statistically not clearly identified as a separate subscale, the items showed good discrimination and gender differences are most prominent on these items. This finding may be explained by girls being more communication-oriented users of ICT, as reported by Tsai and Tsai (2010). Another possible explanation may be found in the overall advantage of girls in reading literacy (OECD, 2010; Mullis et al., 2012). The items for Dimension 3 often consisted of relatively extensive written instructions. More importantly, the ability to read, interpret and process the presented web-based information was essential to correctly complete these tasks. In future PIRLS, PISA or ICILS assessments, it would be interesting to combine the assessment of reading literacy and CIL and to examine gender differences in the relation between these two subjects.

In addition to this cross-sectional study, longitudinal studies and ICILS cycles can provide valuable additional information on the development of gender differences in CIL over time. Especially when a multidimensional perspective of CIL is taken.

3. *To what extent are these differences consistent across European countries participating in ICILS?*
The extent of gender differences on the separate scales varied from country to country, but no extreme difference was observed for one or several of the European countries under review. Although the number of countries in this study is limited, it suggests that the observed gender differences in CIL may be a common phenomenon, at least for European countries. The results thus provide evidence of general gender differences in CIL, but do not clearly point to, for example, educational systems with unique gender differences in CIL. A broader comparison among more countries participating in ICILS would be valuable future research on this topic.

4. *To what extent is the validity of the multidimensional IRT measurement model threatened by gender-related item bias and cultural item bias?*
We investigated item bias between genders and among countries as these item biases might affect the comparison of boys and girls and of countries. Both methods used, failed to uncover strong evidence for item bias across gender groups or countries. Therefore, results support the use of one model across the nine countries and confirm the validity of the group comparisons based on this model.

In conclusion, the results of this study support our postulated hypothesis on gender differences in CIL and the dimensionality of the construct. The findings shed a new light on the often-assumed disadvantaged position of girls and women in todays' information society, and provides directions for future studies on the topic with regard to dimensionality and item fit. However, before it can be convincingly concluded that the gender gap has closed or even reversed, more research is needed; for example, by assessing more challenging tasks for the computer literacy dimension and also by further investigating the gender differences in self-efficacy. In this study we specifically explored gender differences and found evidence for the assumption that the abilities of girls and boys differ between the dimensions of CIL. To deepen our understanding of gender differences in CIL - or possible differences between other groups of students, such as groups with different home environments, experiences or learning styles - we regard it important to acknowledge that CIL is not a unidimensional construct and that further exploration and conceptualisation of dimensions of CIL and its relationship with reading literacy is necessary. For curriculum developers, developers of learning materials, teacher training institutes, schools and teachers this information would be very valuable as ICILS-2013 convincingly showed that CIL should receive more attention in the intended and implemented curriculum of all participated countries, as most of the tested students turned out to be not a "digital native" at all (Fraillon et al., 2014).

Chapter 3
Modelling Parental Involvement and Reading Literacy: Handling Country-Specific Differences in Parental Involvement[2]

Abstract

Parental involvement is seen as a malleable factor in the student's home situation to enhance student achievement, though found effects are not univocal. The inconsistent results may be caused by differences between educational systems and cultural differences, or by the great variation in the methods used to assess student achievement and parental involvement across studies. This chapter describes how data from the Progress in International Reading and Literacy Study 2011 is used to develop a suitable psychometric framework to model country-specific differences in the multifaceted parental involvement construct, both at the item and scale level. Country-specific differences, i.e. cultural differential item functioning (CDIF), in five constructed parental involvement components were identified using a Lagrange multiplier test statistic and modelled by the generalized partial credit model (GPCM) with (1) random item parameters, and (2) fixed item parameters for the 10 and 20 percent strongest cases of CDIF. Also, (3) a bi-factor application of the GPCM was introduced. All models clearly and consistently supported the identification of CDIF. Results did vary over the methods as the first two methods focus on identification of uniform CDIF and the bi-factor application on non-uniform CDIF. In the next chapter the relationship between the parental involvement scales, taking CDIF into account, and student achievement is evaluated.

---

[2]Based on Chapters 1, 3, 4, and 6 from: Punter, R. A., Glas, C. A. W., & Meelissen, M. R. M. (2016). *Psychometric Framework for Modeling Parental Involvement and Reading Literacy*. Cham: Springer.

3.1 Introduction

Although the role of parental involvement, defined by Hill et al. (2004, p. 1491) as "parents' interactions with schools and with their children to promote academic success", in student achievement (and reading literacy in particular) is widely acknowledged, research findings regarding its effect differ considerably (Punter, Glas, & Meelissen, 2016a). The literature review on parental involvement and student achievement by Punter et al. pointed to several explanations for the inconsistent findings, such as the multidimensional nature of parental involvement that has led to a variety of definitions and measurements, making comparisons of findings across studies troublesome. Owing to this large variation in the methods and perspectives used to measure parental involvement and student achievement in these studies, it is difficult to establish whether the inconsistency in the results is caused by differences between educational systems and cultures, or by the method applied and the instruments used. Empirical research is therefore required into the measurement of student achievement and indicators of the parental involvement, and how comparisons can be made between educational systems (countries), to find out to what extent and under which conditions, parental involvement influences student achievement. In-depth analyses of large-scale international comparative data, such as that contained in the Progress in International Reading and Literacy Study (PIRLS) may provide a valuable addition to the research into parental involvement.

PIRLS is an international large-scale assessment study (ILSA) collecting data in over 40 countries on reading literacy achievement of students in Grade 4. In addition to the reading assessment, questionnaires are administered to collect a range of contextual factors that affect students' learning, among which parental involvement. The administration of questionnaires to students, parents as well as principals and teachers enables the measurement of parental involvement from these different perspectives. In addition, the use of the same instruments across countries paves the way for international comparisons. In the present study, the available indicators of parental involvement in the PIRLS-2011 data (Mullis, Martin, Foy, & Drucker, 2012) were matched to the framework for parental involvement by Punter et al. (2016a, p.10), resulting in multiple scales on parental involvement. However, in ILSAs such as PIRLS the extent to which the data of different countries can be usefully compared may be limited. Despite the high-quality demands for the translation of the

instruments and the conditions of administration in each participating country, cultural differences could influence the international validity of the indicators measured. Particularly, item-by-country interactions may be present in the item parameters, indicating cultural differential item functioning (CDIF).

The research objective in the present chapter is to explore the extent to which there are any cultural differences (differences between countries) in the components that measure dimensions of parental involvement in PIRLS-2011 and to provide and compare ways to model these differences. To address this, five extracted item sets, each related to one of the above-mentioned dimensions of parental involvement, were scaled using item response theory (IRT) models and studied for CDIF. The detection and modelling of CDIF was done by applying random and fixed item parameters. Cultural differences in the construct were also studied using a bi-factor IRT model. Such a bi-factor model gives an indication of the extent to which the scale loads on the intended latent variable, for example a particular parental involvement component, and the extent to which it loads on a country-specific component. The applied methods are presented in this chapter and results are compared to evaluate their usefulness in identifying and modelling CDIF.

For the purpose of developing a psychometric framework, including CDIF, to assess the relation between parental involvement and reading literacy, this chapter starts out by presenting a framework to identify and model CDIF in the parental involvement construct in multiple ways. In the next chapter (Chapter 4) the relation between parental involvement and student achievement in reading literacy is further explored, using the scales constructed in this chapter.

3.2 Method

3.2.1 Parental Involvement and Countries in PIRLS-2011

Our analysis of the PIRLS-2011 data (Mullis, Martin, Foy, & Drucker, 2012) was guided by an analytic framework (Table 3.1), based on the dimensions of parental involvement in literature identified by Punter et al. (2016a). The framework matched the dimensions and perspectives to the available indicators of parental involvement in the PIRLS-2011 data. The first dimension they discerned from literature, i.e. home-based involvement from the perspective of parents, was split into two components or indicators: early literacy activities and help with homework. The early literacy activities

component is measured by the PIRLS home questionnaire. In the international reports of PIRLS-2011, early literacy activities is the only component reported as a scale, with Cronbach's alphas ranging from .70 (Czech Republic, Hungary, Italy and Oman) to .88 (Romania), indicating high reliability (Martin & Mullis, 2012). Although the international report does not report the scale statistics on the items regarding parental help with homework, this component could be used in our analyses as well. A total of eight items asks about these practices in the home questionnaire. To consider the dimension of school-based involvement and home-school communication from the parent's perspective, three relevant items were selected from the home questionnaire. The number of items for this indicator is low, but the items do seem highly relevant to this context. The student's perception of parental involvement and the school's practices on parental involvement are measured by five items in the student questionnaire and 15 items in the school questionnaire, respectively.

Table 3.1
Components and Items for Modelling Parental Involvement Using PIRLS-2011 Data

Component	Source	Question	Item number per component	Item in international datasets	Number of response categories
1 Early literacy activities before beginning primary school (home-based involvement)	Home questionnaire	*Before your child began primary school, how often did you or someone else in your home do the following activities with him or her?*			
		Read books	1	ASBH02A	3[a]
		Tell stories	2	ASBH02B	3[a]
		Sing songs	3	ASBH02C	3[a]
		Play with alphabet toys	4	ASBH02D	3[a]
		Talk about things you had done	5	ASBH02E	3[a]
		Talk about what you had read	6	ASBH02F	3[a]
		Play word games	7	ASBH02G	3[a]
		Write letters or words	8	ASBH02H	3[a]
		Read aloud signs and labels	9	ASBH02I	3[a]
2 Help with homework (home-based involvement)	Home questionnaire	*How often do you or someone else in your home do the following things with your child?*			
		Discuss my child's schoolwork with him/her	1	ASBH09A	4[b]
		Help my child with his/her schoolwork	2	ASBH09B	4[b]
		Make sure my child sets aside time to do his/her homework	3	ASBH09C	4[b]
		Ask my child what he/she learned in school	4	ASBH09D	4[b]
		Check if my child has done his/her homework	5	ASBH09E	4[b]
		Help my child practice his/her reading	6	ASBH09F	4[b]

(continued)

Table 3.1 (continued)

Component	Source	Question	Item number per component	Item in international datasets	Number of response categories
		Help my child practice his/her math skills	7	ASBH09G	4[b]
		Talk with my child about what he/she is reading	8	ASBH09H	4[b]
3 School practices on parental involvement, parent perspective (home-school communication)	Home questionnaire	*What do you think of your child's school?*			
		My child's school includes me in my child's education	1	ASBH10A	4[c]
		My child's school should make a greater effort to include me in my child's education	2	ASBH10B	4[c]
		My child's school should do better at keeping me informed of his/her progress	3	ASBH10E	4[c]
4 Parental involvement, student perspective (home-based involvement)	Student questionnaire	*How often do the following things happen at home?*			
		My parents ask me what I am learning in school	1	ASBG07A	4[b]
		I talk about my schoolwork with my parents	2	ASBG07B	4[b]
		My parents make sure that I set aside time for my homework	3	ASBG07C	4[b]
		My parents check if I do my homework	4	ASBG07D	4[b]
		Do you read for any of the following reasons?			
		My parents like it when I read	5	ASBR09C	4[c]
5 School practices on parental involvement, school perspective (home-school communication, school-based involvement)	School questionnaire	*How often does your school do the following for parents concerning individual students?*			
		Inform parents about their child's learning progress	1	ACBG11AA	4[d]
		Inform parents about the behaviour and well-being of their child at school	2	ACBG11AB	4[d]
		Discuss parents' concerns or wishes about their child's learning	3	ACBG11AC	4[d]
		Support individual parents in helping their child with schoolwork	4	ACBG11AD	4[d]
		How often does your school ask parents to do the following?			
		Volunteer for school projects, programs, and trips	5	ACBG11BA	4[d]
		Serve on school committees	6	ACBG11BB	4[d]
		How often does your school do the following for parents in general?			
		Inform parents about the overall academic achievement of the school	7	ACBG11CA	4[d]
		Inform parents about school accomplishments	8	ACBG11CB	4[d]

(continued)

Table 3.1 (continued)

Component	Source	Question	Item number per component	Item in international datasets	Number of response categories
		Inform parents about the educational goals and pedagogic principles of the school	9	ACBG11CC	4[d]
		Inform parents about the rules of the school	10	ACBG11CD	4[d]
		Discuss parents' concerns or wishes about the school's organization	11	ACBG11CE	4[d]
		Provide parents with additional learning materials	12	ACBG11CF	4[d]
		Organize workshops or seminars for parents on learning or pedagogical issues	13	ACBG11CG	4[d]
		How would you characterize each of the following within your school?			
		Parental support for student achievement	14	ACBG12E	5[e]
		Parental involvement in school activities	15	ACBG12F	5[e]

Note. The datasets are described in detail in Foy and Drucker (2013).
[a] Category labels are: 0 - Often, 1 - Sometimes, 2 - Never or almost never.
[b] Category labels are: 0 - Every day or almost every day, 1 - Once or twice a week, 2 - Once or twice a month, 3 - Never or almost never.
[c] Category labels are: 0 - Agree a lot, 1 - Agree a little, 2 - Disagree a little, 3 - Disagree a lot.
[d] Category labels are, after recoding: 0 - More than three times a year, 1 – Two to three times a year, 2 - Once a year, 3 - Never.
[e] Category labels are: 0 - Very high, 1 - High, 2 - Medium, 3 - Low, 4 - Very low.

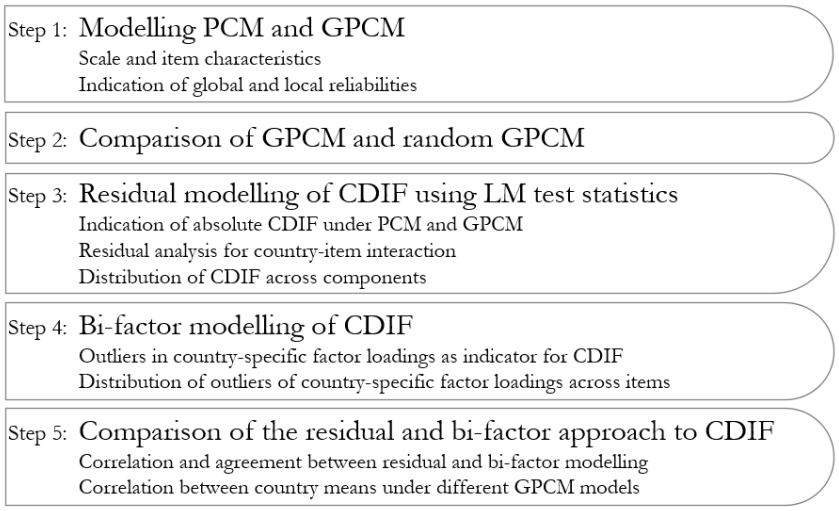
For this study, we initially considered the data from 43 countries participating in PIRLS-2011; countries for which the average achievement in student reading literacy was not reliably measured were excluded (Mullis, Martin, Foy, & Drucker, 2012). However, two countries, England and the USA, did not administer the home questionnaire and were therefore not included in the scale analyses for components 1 to 3, using items from the home questionnaire.

3.2.2 Modelling Steps

Figure 3.1 illustrates the process of identifying and modelling CDIF in this study. Details on the models are provided in the next section. The analyses started by modelling the five scales using the partial credit model (PCM; Masters, 1982) and the generalized partial credit model (GPCM; Muraki, 1992) to see how well the scales perform by means of global and local reliability. Global reliability of a scale refers to the ratio of the systematic and the total variance in the θ-parameter within a country, local reliability to the extent to which different θ-values can be distinguished from each other. In step two, the first global indication of CDIF is provided by comparing

the GPCM to a GPCM with random item parameters. More specific indications of CDIF are obtained by the residual approach in step three. This includes residual analyses to identify strong cases of CDIF and the application of a so-called split item approach to the 10% and 20% strongest CDIF cases. In step four, the five scales are modelled using the bi-factor application of the GPCM, providing information on the cultural differences based on the country-specific factor loadings. The final step of analyses is comparing the methods of modelling CDIF.

Figure 3.1
Steps in Modelling Cultural Differential Item Functioning.



3.2.3 Description of the Models

We describe the procedure for the bi-factor model, combined with the PCM and GPCM as IRT models, since these two models were the ones we selected for the present study. However, the approach also applies to other IRT models, such as the graded response model (Samejima, 1969), the sequential model (Tutz, 1990), and other versions of these models with random item parameters instead of fixed item parameters.

The bi-factor model used in this study was in two parts: a measurement model (i.e., an IRT model) and a structural model. The measurement model pertains to a polytomously-scored response of a student n to an item i. The possible item scores range from 0 to m_i and the score of student n on item i is denoted by the variables x_{nij} ($j = 0,\ldots, m_i$) where $x_{nij} = 1$ if the response is in category j and zero otherwise. Note that m_i has an index i, which indicates that the maximum score of items can differ.

In the bi-factor GPCM, the probability of scoring in category j ($j = 0,\ldots, m_i$) is given by

$$p_{ij}(\theta_n) = P(x_{nij} = 1 \mid \theta_n, a, b) = \frac{\exp\left(\sum_{h=1}^{j} a_{i0}\theta_{n0} + a_{ig(n)}\theta_{ng(n)} - b_{ih}\right)}{1 + \sum_{k=1}^{m_i} \exp\left(\sum_{h=1}^{k} a_{i0}\theta_{no} + a_{ig(n)}\theta_{ng(n)} - b_{ih}\right)}, \qquad (3.1)$$

where θ_{n0} is the score of a student n on the latent scale pertaining to all countries, $\theta_{ng(n)}$ is the score on a country-specific latent dimension, and the index $g(n)$ indicates the country to which student n belongs. Further, a_{i0} and $a_{ig(n)}$ are the factor loadings of item i on these two dimensions, and b_{ih} ($h = 1,\ldots, m_i$) is the item location parameter. The location parameter b_{ih} is the position on the latent scale, where it is assumed that summations such as $h=1$ to 0 result in zero. The unidimensional GPCM lacks the country-specific dimensions $\theta_{ng(n)}$ and the associated factor loadings $a_{ig(n)}$. Further, the PCM is obtained by fixing all item parameters a_{i0} to 1.

The formula for the response probability and subsequent derivations can be simplified by introducing the re-parametrization $d_{ij} = \sum_{h=1}^{j} b_{ih}$ and by defining $a_{ig}^t\theta_n$ as the inner product of the vectors $\left(a_{i0}, a_{ig(n)}\right)$ and $\left(\theta_{n0}, \theta_{ng(n)}\right)$, respectively. Thus, Eq. (3.1) becomes

$$p_{ij}(\theta_n) = \frac{\exp\left(ja_{ig}^t\theta_n - d_{ij}\right)}{1 + \sum_{k=1}^{m_i} \exp\left(ka_{ig}^t\theta_n - d_{ik}\right)} \qquad (3.2)$$

The θ_0-dimension is the general dimension that pertains to all countries and is the basis for the comparison of the countries. The θ_g-dimensions are the country-specific dimensions, and the factor loadings on these dimensions give an indication of country-by-item interaction. It is assumed that within each country g, the dimensions θ_0 and θ_g have a bi-variate normal distribution $N(\theta_{n0}, \theta_{ng}; \boldsymbol{\mu}_g, \Sigma_g)$. For the two-dimensional

country mean $\boldsymbol{\mu_g} = (\mu_{g0}, \mu_g)$, it assumes that the mean on the second dimension is fixed at zero, that is $\mu_g = 0$. The covariance matrix is given by $\Sigma_g = \begin{bmatrix} \sigma_g^2 & 0 \\ 0 & 1 \end{bmatrix}$.

In the unidimensional GPCM and PCM, the latent student parameters θ_0 have a univariate normal distribution with a mean μ_g and a variance σ^2_g. Finally, random item parameters are obtained by introducing independent multivariate normal distributions on the parameters for each item (for further details, see De Jong, Steenkamp, & Fox, 2007).

The present application of the bi-factor model is not standard, but an extension of the basic model (Gibbons & Hedeker, 1992). The technical details on the estimation equations, expressions for the covariance matrix of the estimates, and tests of model fit are provided in Punter, Glas, and Meelissen (2016b).

3.2.4 Detection and Modelling of Differential Item Functioning

In the third modelling step, this study uses a Lagrange multiplier (LM) test statistic (Rao, 1947; see also, Aitchison & Silvey, 1958). LM tests have been previously applied to IRT frameworks (Glas, 1999; Glas & Falcón, 2003; Glas & Dagohoy, 2007). Glas and Jehangir (2014) already showed the feasibility of the method to identify CDIF and modelled it using country-specific item parameters using PISA data, although in the slightly simpler framework of one-dimensional IRT models. Our primary interest is not in the actual outcome of the LM test, because due to the very large sample sizes in educational surveys even the smallest model violation, that is, the smallest amount of differential item functioning (DIF), will be significant. The reason for adopting the framework of the LM test is that it clarifies the connection between the model violations, and observations and expectations used to detect DIF. Further, because it produces comprehensible and well-founded expressions for model expectations, the value of the LM test statistic can be used as a measure of the effect size of DIF, and the procedure can be easily generalized to a broad class of IRT models.

To define the test and the associated residuals, we define a background variable

$$y_{nc} = \begin{cases} 1 \text{ if person } n \text{ belongs to country } c, \\ 0 \text{ if person } n \text{ does not belong to country } c. \end{cases}$$

The LM test is based on adding item parameter δ_i to the model to allow for country-specific differences in response probabilities and targets the null-hypothesis of no DIF, namely the null-hypothesis where $\delta_i = 0$. The LM test statistic is computed using the MML estimates of the null-model, where δ_i is not estimated. The test is based on evaluation of the first-order derivatives of the marginal likelihood with respect to δ_i evaluated at $\delta_i = 0$ (see Glas, 1999). If the first-order derivative in this point is large, the MML estimate of δ_i is far removed from zero, and the test is significant. If the first-order derivative in this point is small, the MML estimate of δ_i is probably close to zero and the test is not significant. The actual LM statistic is the squared first-order derivative divided by its estimated variance, and it has an asymptotic chi-squared distribution with one degree of freedom. However, as already discussed, the primary interest is not so much in the test itself, but in the information it provides regarding the fit between the data and the model.

The LM approach can be outlined using the bi-factor GPCM; the special cases for the unidimensional PCM and GPCM are obtained if the restrictions denoted above are invoked. The probability of a response is given by a generalization of the bi-factor GPCM, namely, $p_{ij}(\theta_n) = \dfrac{\exp\left(j a_{ig}^t \theta_n - d_{ij} + j \sum_c y_{nc} \delta_{ic} \right)}{1 + \sum_{k=1}^{m_i} \exp\left(k a_{ig}^t \theta_n - d_{ik} + k \sum_c y_{nc} \delta_{ic} \right)}$.

For one so-called reference country, the covariate y_{nc} is equal to zero. This country serves as a baseline where the bi-factor GPCM with item parameters *a* and *b* holds. In the other *C-1* countries, the covariates y_{nc} are equal to 1. It can be shown (see Glas, 1999) that the test statistic is based on the residuals

$$\frac{\sum_{n=1}^{N} \sum_{j=1}^{m_i} y_{nc} j X_{ij}}{\sum_{n=1}^{N} y_{nc}} - \frac{\sum_{n=1}^{N} \sum_{j=1}^{m_i} y_{nc} j E\left(p_{ij}(\theta_n) \mid x_n; \lambda \right)}{\sum_{n=1}^{N} y_{nc}}, \tag{3.3}$$

for $c = 1, \ldots, C-1$, where λ is a vector of all the parameters in the null model, i.e. the item parameters and the parameters of the population distributions. Dividing this residual by the number of respondents $\Sigma_n y_{nc}$ produces residuals that are the

differences between the observed and expected average item-total score in country c = *1,…, C-1*. The residual gauges so-called uniform DIF, in other words, the residual indicates whether the item total function $\Sigma_j jp_{ij}(\theta)$ is shifted for the item, namely whether there is item-by-country interaction.

The LM statistic for the null-hypothesis $\delta_{ic} = 0$ (c = *1,…, C-1*) is a quadratic form in the *(C-1)*-dimensional vector of residuals and the inverse of their covariance matrix (for details, see Glas, 1999). It has an asymptotic chi-squared distribution with *C-1* degrees of freedom. A special case of this procedure is obtained if one country serves as the focal country and all other countries serve as reference. Then the model under the alternative hypothesis has only one additional parameter, δ_i, and the associated LM statistic has an asymptotic chi-squared distribution with one degree of freedom.

Items that show the worst misfit, based on their value of the LM statistic and residuals, are given country-specific item parameters. From a practical point of view, defining country-specific item parameters is equivalent to defining an incomplete design where the DIF item is split into a number of virtual items, and where each virtual item is considered as administered in a specific country. The resulting design can be analysed using IRT software that supports the analysis of data collected in an incomplete design. We here refer to items with country-specific parameters as split items.

The method is motivated by the assumption that a substantial part of the items function the same in all countries and a limited number of items have CDIF. In the IRT model, it is assumed that all items pertain to the same latent variable θ. Items without CDIF have the same item parameters in every country. However, items with CDIF have item parameters that differ across countries. These items refer to the same latent variable θ as all the other items, but their location on the scale differs across countries. For instance, the number of cars in the family may be a good indicator of wealth, but the actual number of cars at a certain level of wealth may vary across countries, or even within countries. Having a car in the inner city of Amsterdam is clearly a sign of wealth, but, in the rural eastern part of the Netherlands, an equivalent level of wealth would probably result in the ownership of three cars. The number of items given country-specific item parameters is a matter of choice where two

considerations are relevant. First, there should remain a sufficient number of anchor items in the scale. Second, the model including the split items should fit the data.

3.2.5 Estimation Procedures

The procedures we used for parameter estimation and evaluation of model fit are based on marginal maximum likelihood (MML). Most of the procedures we discuss are documented in more detail elsewhere (see Adams & Wu, 2006; Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988; De Jong et al., 2007; Gibbons & Hedeker, 1992; Glas, 1999; Glas & Jehangir, 2014; Jennrich & Bentler, 2011; Glas & Jehangir, 2014). We used the MIRT/LEXTER public domain software package (Glas, 2010) in the calculations. Additional estimation and testing procedures were used for the bi-factor model, with unidimensional models as special cases, and random item parameters as a generalization.

3.3 Results

3.3.1 Step 1: Scale and Item Characteristic Under the PCM and GPCM

Each component was modelled using both the PCM and the GPCM, with estimated global reliability for each country. Table 3.2 provides the descriptive statistics at country level for the first parental involvement component under the PCM and GPCM, including estimated global reliability. Results for the other components are provided in Appendix A Tables A.1 to A.4. Components 1 (early literacy activities), 2 (help with homework), and 5 (school practices on parental involvement, school perspective) were evaluated using nine, eight, and 15 items, respectively (see also Table 3.1). Their global reliability under the GPCM is generally > 0.70, which is an acceptable level for country inferences. Components 3 (school practices on parental involvement, parental perspective) and 4 (parental involvement from a student perspective), were evaluated using three items and five items, respectively; the global reliability under the GPCM of these estimates was thus correspondingly lower (<.70 for component 4 and even <.67 for component 3).

Table 3.2
Country Characteristics Component 1: Early Literacy Activities Before Beginning Primary School

Country	N	$\bar{X}$	PCM			GPCM		
			$\mu(\theta)$	$\sigma(\theta)$	ρ	$\mu(\theta)$	$\sigma(\theta)$	ρ
Azerbaijan, Republic of	4509	6.56	0.44	1.05	0.74	0.36	0.98	0.74
Australia	3232	4.46	-0.55	1.30	0.77	-0.49	1.19	0.77
Austria	4393	5.90	0.10	1.01	0.73	0.08	0.94	0.74
Belgium (French)	3383	6.46	0.30	1.01	0.74	0.29	0.94	0.74
Bulgaria	5137	6.10	0.12	1.57	0.84	0.12	1.46	0.85
Canada	18848	4.57	-0.49	1.25	0.76	-0.44	1.14	0.76
Chinese Taipei	4242	8.41	0.98	1.11	0.78	0.90	1.03	0.78
Colombia	3798	5.79	0.11	1.19	0.77	0.13	1.10	0.77
Croatia	4539	4.62	-0.38	0.97	0.69	-0.35	0.90	0.69
Czech Republic	4397	5.28	-0.10	0.90	0.68	-0.09	0.84	0.69
Denmark	4322	6.10	0.18	0.96	0.72	0.18	0.90	0.73
Finland	4423	6.23	0.24	0.80	0.65	0.24	0.74	0.65
France	4111	5.94	0.12	1.02	0.74	0.11	0.95	0.74
Georgia	4640	4.46	-0.44	1.11	0.72	-0.44	1.02	0.72
Germany	3197	5.56	-0.01	0.96	0.71	-0.02	0.89	0.71
Hong Kong, SAR	3604	8.45	1.01	0.97	0.73	0.91	0.90	0.74
Hungary	4912	5.27	-0.11	0.92	0.69	-0.12	0.85	0.69
Indonesia	4588	6.90	0.48	1.02	0.74	0.45	0.94	0.75
Iran, Islamic Republic of	5653	7.82	0.82	1.06	0.76	0.75	0.99	0.76
Ireland	4268	4.58	-0.47	1.24	0.76	-0.43	1.14	0.76
Israel	3261	4.81	-0.33	1.11	0.74	-0.30	1.03	0.74
Italy	3873	4.97	-0.23	1.00	0.71	-0.20	0.93	0.71
Lithuania	4406	5.67	0.04	0.96	0.71	0.01	0.90	0.71
Malta	3274	5.24	-0.18	1.14	0.76	-0.17	1.06	0.76
Netherlands	2273	5.53	-0.03	0.96	0.71	-0.02	0.89	0.71
New Zealand	3357	4.37	-0.60	1.33	0.77	-0.54	1.22	0.77
Norway	2909	5.76	0.06	0.97	0.71	0.08	0.90	0.72
Northern Ireland	2107	4.02	-0.74	1.28	0.75	-0.68	1.18	0.75
Poland	4920	5.06	-0.20	0.99	0.71	-0.20	0.92	0.71
Portugal	3887	5.76	0.05	1.09	0.75	0.04	1.01	0.76
Qatar	3650	6.49	0.35	1.08	0.75	0.30	1.00	0.76
Romania	4535	5.59	-0.12	1.57	0.83	-0.12	1.46	0.84
Russian Federation	4412	4.02	-0.70	1.19	0.73	-0.68	1.09	0.73
Saudi Arabia	4369	6.52	0.36	1.04	0.75	0.36	0.97	0.75
Singapore	6194	7.16	0.51	1.24	0.80	0.47	1.15	0.81
Slovak Republic	5481	5.02	-0.24	1.08	0.74	-0.23	1.00	0.74
Slovenia	4313	4.78	-0.33	1.02	0.71	-0.31	0.94	0.71
Spain	7945	5.13	-0.18	1.03	0.72	-0.16	0.95	0.73
Sweden	4013	6.06	0.15	1.03	0.74	0.15	0.96	0.75
Trinidad and Tobago	3497	4.85	-0.33	1.17	0.75	-0.29	1.08	0.76
United Arab Emirates	13305	6.52	0.35	1.03	0.74	0.32	0.96	0.75

Note. N is the sample size, and $\bar{X}$ the observed mean score on the component. $\mu(\theta)$ is the estimated mean, $\sigma(\theta)$ is the standard deviation, and ρ is the estimated global reliability under the partial credit model (PCM) or the generalized partial credit model (GPCM). The origin of the latent scale was identified by setting the sum of the country means to zero. The variance of the scale was identified by fixing it to one.

We also investigated the item characteristics for each component under the PCM and GPCM. Table 3.3 shows the item characteristics for the early literacy activities scale, including estimated local reliability at the mean of the ability distribution. For the other components the results are provided in Tables A.5-A.8. Local reliability was assessed using the "slope" parameter in the GPCM. The relatively high value for PIRLS item

ASBH02A ('read books'), indicates that this item of the scale performed best in this respect. Local reliability is further supported if the item location parameters agree closely with the mean of a latent distribution. In this respect, item ASBH02G ('play word games') performed best, because the latent distributions of the countries were normed to an overall mean of zero. Together the intercept and slope parameters determine the information value of an item. Higher values for the item information at $\theta = 0,$ namely $I(0),$ indicate that the item made a greater contribution to the local reliability of the component.

Table 3.3
Response Frequencies and Item Parameter Estimates Under the Generalized Partial Credit Model for Items in Component 1: Early Literacy Activities

Item	Slope	Intercept	*I(0)*	Relative frequency response categories		
				Cat0	Cat1	Cat2
ASBH02A	1.26	1.84	0.44	0.54	0.41	0.05
ASBH02B	1.24	1.47	0.46	0.48	0.46	0.07
ASBH02C	0.77	0.98	0.23	0.49	0.41	0.11
ASBH02D	1.09	0.85	0.45	0.43	0.44	0.14
ASBH02E	0.95	1.80	0.24	0.62	0.34	0.04
ASBH02F	1.18	0.82	0.45	0.36	0.52	0.12
ASBH02G	1.24	0.57	0.52	0.33	0.51	0.16
ASBH02H	1.06	1.12	0.38	0.47	0.44	0.10
ASBH02I	1.07	0.89	0.43	0.44	0.42	0.13

Note. Slope and intercept are the parameters a_{i0} and the mean of the location parameters b_{i1}, b_{i2},… respectively, under the general partial credit model (GPCM). *I(0)* is the information value of the item at $\theta = 0$. Cat0, Cat1, Cat2 indicate the frequency with which item categories 0, 1 and 2 are endorsed, respectively. The content of the components, items and corresponding category labels are described in Table 3.1.

For component 1 (early literacy activities), the item ASBH02C ('sing songs') has a lower information value than the other items. This should be taken into account when redesigning the instrument for future surveys; in other words, this item may be the first candidate for replacement. Compared to component 1 (early literacy activities), most items in component 2 (helping with homework) were more informative, while items in component 3 (school practices on parental involvement, parent perspective) performed poorly. Components 4 (school practices for parental involvement from a student perspective) and 5 (school practices for parental involvement from a school perspective) provided differing results; in particular, the last two items of component 5 ('parental support for student achievement within school' and 'parental involvement in school activities') performed particularly poorly.

3.3.2 Step 2: GPCM with Random Item Parameters

Comparing the parameter estimates in the GPCM and the GPCM with random item parameters (henceforth the random GPCM) revealed that the agreement between the

slopes and intercepts under the GPCM and the means of the slopes and intercepts under the random GPCM was high (for component 1: Table 3.4, for components 2-5: Tables A.9-A.12 in Appendix A). A higher variance provides an initial indication that the item functions differently in different countries, a topic we address in more detail later. Here, the effects are global over countries and thus only permit global inferences. For instance, for component 1, the last item, ASBH02I ('read aloud signs and labels') has the lowest CDIF because the variance of the intercepts and slopes across the countries is the lowest among the items. A low variance indicates that the item parameters do not vary much across countries. Evaluating the relative CDIF of the other eight items is more difficult, because of the trade-off between the standard deviation for the slope and the intercept. This pattern is repeated for component 2; the items ASBH09F ('helping child practice reading') and ASBH09G ('helping child practice math skills') performed slightly better than the other items. Conversely, component 3 showed a substantial difference between the item parameters estimated with the GPCM and those estimated using the random GPCM, indicating this short scale was quite unstable. The analyses of components 4 and 5 indicated all the items performed comparably with respect to CDIF, although questions surrounding specific item-by-country interaction and the influence of the inferences on country means remain unanswered.

Table 3.4
Item Parameter Estimates Under the Generalized Partial Credit Model (GPCM) and GPCM with Random Item Parameters for Items in Component 1: Early Literacy Activities

Item	GPCM		GPCM random item parameters			
	Slope	Intercept	Slope	*SD* (Slope)	Intercept	*SD* (Intercept)
ASBH02A	1.26	1.84	1.37	0.22	2.06	0.66
ASBH02B	1.24	1.47	1.25	0.15	1.50	0.31
ASBH02C	0.77	0.98	0.80	0.12	1.03	0.34
ASBH02D	1.09	0.85	1.21	0.18	0.86	0.44
ASBH02E	0.95	1.80	1.01	0.19	2.01	0.68
ASBH02F	1.18	0.82	1.33	0.23	0.93	0.42
ASBH02G	1.24	0.57	1.35	0.15	0.60	0.27
ASBH02H	1.06	1.12	1.16	0.16	1.17	0.41
ASBH02I	1.07	0.89	1.09	0.11	0.87	0.22

Note. *SD* (Slope) indicates the standard deviation of the slope. *SD* (Intercept) indicates the standard deviation of the intercept. The content of the items and corresponding category labels are described in Table 3.1.

3.3.3 Step 3: Residual Modelling of CDIF

We compared CDIF as identified by the random GPCM with CDIF as identified using the latent residuals defined by Eq. (3.3) and aggregated over countries. Results

are shown for the first component in Table 3.5. Results for the other components are in Appendix A (Tables A.13-A.16). Overall the agreement between the methods was high. For instance, item ASBH02I performed strongly in all methods, as did item ASBH02G. In general, the residuals with the GPCM are smaller than those with the PCM, because the latter model has fewer parameters. However, we found that differences between the PCM and the GPCM were very small. A striking exception, again, was component 3. Here the fit of the GPCM was worse than the fit of the PCM, which leads to the conclusion that the slopes are very hard to estimate. This is in agreement with the reported low global reliability. Obviously, variance in the θ-distribution is too small to support a proper estimate of the slope parameters.

Table 3.5
Absolute Differential Item Functioning Under the Partial Credit Model and the Generalized Partial Credit Model and Standard Deviations of Random Item Parameters on Items in Component 1: Early Literacy Activities

Item	PCM	GPCM	*SD* (Slope)	*SD* (Intercept)
ASBH02A	0.12	0.11	0.228	0.667
ASBH02B	0.08	0.08	0.158	0.318
ASBH02C	0.09	0.10	0.126	0.349
ASBH02D	0.12	0.12	0.183	0.443
ASBH02E	0.10	0.10	0.192	0.688
ASBH02F	0.09	0.09	0.239	0.421
ASBH02G	0.07	0.07	0.155	0.279
ASBH02H	0.10	0.10	0.161	0.416
ASBH02I	0.07	0.07	0.112	0.229

Note. The columns labelled PCM and GPCM give the mean residuals as estimated under the unidimensional versions of these two models. SD (Slope) indicates the standard deviation of the slope. SD (Intercept) indicates the standard deviation of the intercept. The content of the items and corresponding category labels are described in Table 3.1.

We then addressed the distribution of country-by-item interaction across countries and items, to determine whether the sizes and directions of the residuals were randomly distributed across all countries and items, or whether they exhibited notable patterns of interaction (Table 3.6 and Tables A.17-A.20 in Appendix A). Residuals were defined by Eq. (3.3), estimated under the GPCM, and calculated for every country, with that country as a focus and all other countries as a reference. To simplify, here we shall not consider the specific values of the residuals, but instead concentrate on the outlying values. For example, if we examine results obtained for the Republic of Azerbaijan and Australia for component 1 (early literacy activities), it is clear that, aggregated over the items, the mean absolute residual for the Republic of Azerbaijan is much larger than the mean absolute residual for Australia. The responses

Table 3.6
Residual Analysis for Country-by-Item Interactions for Component 1: Early Literacy Activities

Country	*N*	Item									10% CDIF	20% CDIF	Absolute residual
		1	2	3	4	5	6	7	8	9			
Azerbaijan, Republic of	4509			++		++	--		--	+	4	5	0.146
Australia	3232	-									0	1	0.072
Austria	4393	-			+						0	2	0.096
Belgium (French)	3383										0	0	0.062
Bulgaria	5137	+									0	1	0.058
Canada	18848										0	0	0.059
Chinese Taipei	4242							--	++		2	2	0.106
Colombia	3798	++		-						--	2	3	0.095
Croatia	4539	+									0	1	0.060
Czech Republic	4397								++		1	1	0.104
Denmark	4322	--			++	--					3	3	0.150
Finland	4423	--	+		++	--		++			4	5	0.160
France	4111						++				1	1	0.080
Georgia	4640					++					1	1	0.080
Germany	3197	-			+	-			+		0	4	0.108
Hong Kong, SAR	3604										0	0	0.068
Hungary	4912		--					--	++	++	4	4	0.148
Indonesia	4588		++			+	-	-	--	+	2	6	0.186
Iran, Islamic Republic of	5653										0	0	0.067
Ireland	4268										0	0	0.073
Israel	3261										0	0	0.066
Italy	3873	+									0	1	0.055
Lithuania	4406			++							1	1	0.072
Malta	3274									+	0	1	0.060
Netherlands	2273	--		-	+				+		1	4	0.129
New Zealand	3357										0	0	0.068
Northern Ireland	2909	-		--	++	--		++			4	5	0.168
Norway	2107								+		0	1	0.086
Poland	4920										0	0	0.040
Portugal	3887						+				0	1	0.070
Qatar	3650	+		++	-						1	3	0.111
Romania	4535	++									1	1	0.074
Russian Federation	4412			++							1	1	0.097
Saudi Arabia	4369	++			-				-		1	3	0.129
Singapore	6194				-	+					0	2	0.075
Slovak Republic	5481										0	0	0.071
Slovenia	4313				+						0	1	0.057
Spain	7945	+					++				1	2	0.091
Sweden	4013	-			++	--		+			2	4	0.118
Trinidad and Tobago	3497										0	0	0.074
United Arab Emirates	13305	+			-						0	2	0.092

Note. + indicates that residual belongs to the 20% most positive residuals, ++ indicates that residual even belongs to the 10% most positive residuals. - indicates that residual belongs to the 20% most negative residuals, -- indicates that residual even belongs to the 10% most negative residuals. The 10% cultural differential item functioning (CDIF) and 20% CDIF columns give the number of outliers in the two respective regions. Absolute residual refers to the means over items of the absolute values of the residuals. The content of the items is described in Table 3.1.

were coded 0, 1 and 2, so the residuals, which are the differences between a mean observed and expected response are also on a scale from 0 to 2. Closer inspection at the item level for Republic of Azerbaijan reveals that items 3 and 5 have residuals among the 10% most positive among the countries, while the items 6 and 8 have residuals among the 10% most negative among the countries. Australia, however, has only one negative residual, and this is among the 20% most negative residuals among the countries. Checking the absolute residuals further reveals Poland fits the model best with the lowest CDIF, while Indonesia has the most significant CDIF.

In a similar way, component 2 (helping with homework) functions very differently in the Netherlands than in other countries (Table A.17, Appendix A), probably because giving students homework is not a daily practice in Dutch primary schools. This different item functioning is indicated by both the high mean of the absolute values of the residuals and the large number of outliers of the residuals. Canada fits the model best with the lowest CDIF for this component. For component 3 (school practices on parental involvement, parent's perspective) the highest mean absolute residual was found for Germany. The scale for measuring school practices on parental involvement from the school perspective (component 5) however, showed relatively little evidence for CDIF.

We undertook a marginal count of the outliers for the items aggregated over the countries (Table 3.7). No one item count was prominent, although the first item in component 3 ('my child's school includes me in my child's education') seemed more susceptible to CDIF than other items, since this item revealed the most outliers in residuals over countries: 13 times in the 10% outliers region and 15 times in the 20% outliers region. Items 5 ('volunteering') and 13 ('organize workshops or seminars for parents on learning or pedagogical issues') within component 5 also scored more highly than other items in the component. However, this does not, of course, mean that these items have CDIF; if 10% and 20% extreme values are considered, then 10% and 20% of the residuals must be included, and such information only serves as a tool to further scrutinize the items.

Table 3.7
Distribution of Cultural Differential Item Functioning Across Items on Parental Involvement

% CDIF	Component	Item														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
10	1	6	2	5	4	6	3	4	5	2						
	2	7	2	5	5	1	9	2	2							
	3	13	0	0												
	4	1	2	3	11	5										
	5	7	1	0	1	10	3	0	5	0	2	0	10	13	9	4
20	1	17	3	7	12	9	5	6	9	5						
	2	15	5	8	9	6	12	5	5							
	3	22	2	1												
	4	7	6	8	14	8										
	5	14	1	2	1	18	11	5	11	3	6	0	15	18	13	11

Note. Count of the number of times a residual is in the extreme 10% and extreme 20% region of the distribution of residuals. The content of the components, items and corresponding category labels are described in Table 3.1.

3.3.4 Step 4: Country Differences Under the Bi-Factor GPCM Model

We also calculated country-specific factor loadings for the bi-factor model, where we first transformed country-specific factor loadings to standard normals, and then identified the 2.5% and 5% most extreme outlying values (for component 1: Table 3.8, for components 2-5: Tables A.21-A.24, Appendix A). This distribution of country-specific factor loadings gives an indication of the extent to which items load on a country-specific factor in addition to the general factor of the item, and can, as in our earlier residual analysis, be used to determine whether the sizes and directions of the factor loadings are randomly distributed across all countries and items, or whether they exhibit notable patterns of interaction.

For component 1 most outliers of the country-specific factor loadings and the highest mean absolute factor loading were found for Colombia, suggesting a high level of CDIF. Interestingly, in the residual analysis for this component a total of 15 countries showed a higher mean absolute residual. Regarding help with homework (component 2) Malta showed the highest number of outliers in country-specific factor loadings (Table A.21, Appendix A). The Netherlands, indicating most CDIF earlier (Table A.17, Appendix A) also showed many outliers, but definitely not the most. For component 3 the results of the counts of outliers provided little information, as only three outliers were counted in the 2.5% region. Hungary did show a high mean absolute country-specific factor loading on this component, though the questionable reliability of the scale must be kept in mind. Student perception of parental involvement (component 4) was measured with the least CDIF in Denmark, whereas

the school practices on parental involvement from the school perspective showed the least CDIF for Italy.

Table 3.8
Outliers of Country-Specific Factor Loadings in the Bi-Factor Model for Component 1: Early Literacy Activities

Country	Item									2.5% Outlier	5% Outlier	Mean absolute loading
	1	2	3	4	5	6	7	8	9			
Azerbaijan, Republic of					+	--		--		3	3	0.062
Australia										0	0	0.034
Austria				--				--	--	3	3	0.071
Belgium (French)	+							++	+	2	3	0.042
Bulgaria										0	0	0.029
Canada								+		1	1	0.050
Chinese Taipei										0	0	0.036
Colombia	--		--	-	--		--	--		5	6	0.095
Croatia	--	--	-		--	--				4	5	0.079
Czech Republic	--		-	+	--	--				4	5	0.083
Denmark	--				--	--				3	3	0.080
Finland	--	--								2	2	0.066
France		-			--					1	2	0.048
Georgia				--			--	--	--	4	4	0.077
Germany										0	0	0.029
Hong Kong, SAR				--		--		--	--	4	4	0.082
Hungary		++				-	-	--		1	4	0.066
Indonesia				++						0	1	0.033
Iran, Islamic Republic of								--		1	1	0.053
Ireland	--	--			-	-				2	4	0.062
Israel	--	-			-	--				2	4	0.064
Italy	--	--	--				-			3	4	0.078
Lithuania					-	--				1	2	0.044
Malta	--	-	--		--	-				3	5	0.076
Netherlands	++									0	1	0.028
New Zealand	--		--		--	--				4	4	0.081
Northern Ireland	--	-		--	--				--	4	5	0.088
Norway	--		-	-	--		--			3	5	0.087
Poland			++	-			--	--		2	4	0.056
Portugal										0	0	0.034
Qatar				++						0	1	0.047
Romania				--			-	--		2	3	0.060
Russian Federation					++					0	1	0.032
Saudi Arabia	--	--				-				2	3	0.065
Singapore				--			-	--		2	3	0.057
Slovak Republic	--				--	--				3	3	0.062
Slovenia	--				--	--				3	3	0.065
Spain		-			--	--				2	3	0.049
Sweden				--			-	-		1	3	0.052
Trinidad and Tobago	-	++					-	--		1	4	0.058
United Arab Emirates	--	--			-					2	3	0.065

Note. + indicates factor loading belongs to the 5% most positive loading, ++ indicates factor loading belongs to the 2.5% most positive loading. - indicates factor loading belongs to the 5% most negative loading, -- indicates factor loading belongs to the 2.5% most negative loading. The 2.5% cultural differential item functioning (CDIF) and 5% CDIF columns give the number of outliers in the two respective regions. Mean absolute loading refers to the means over items of the absolute values of country-specific factor loadings. The content of the items is described in Table 3.1.

Aggregating the items over the countries provides a tool for further investigation of items (Table 3.9), with the same caveats as before; if the 2.5% and 5% most extreme values are considered, then similarly 2.5% and 5% of the country-specific factor loadings must fall in this region, but this does not imply that 2.5% and 5% of the items have CDIF. No item count is prominent. Item 5 ('talk about things you had done') in component 1 did seem more susceptible to CDIF than other items, since this item revealed the most outliers in country-specific factor loadings over countries.

Table 3.9
Distribution of Outliers of Country-Specific Factor Loadings in the Bi-Factor Model Across Items on Parental Involvement

Region	Component	Item														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2.5%	1	7	6	4	8	13	11	4	12	5						
	2	13	1	1	3	1	14	15	8							
	3	1	0	2												
	4	3	2	3	6	0										
	5	9	9	3	1	2	0	0	3	0	0	0	0	0	0	0
5%	1	19	8	13	13	18	15	10	14	5						
	2	15	2	1	6	2	20	21	22							
	3	1	1	3												
	4	6	6	9	11	1										
	5	10	10	3	1	2	0	0	3	0	2	0	0	1	0	0

Note. Count of the number of times an outlier of the country-specific factor loading is in the extreme 2.5% and extreme 5% region of the distribution of factor loadings. The content of the components, items and corresponding category labels are described in Table 3.1.

3.3.5 Step 5: Comparison of the Residual Analyses GPCM and the Bi-Factor GPCM

We then addressed whether the residual analyses using the GPCM and the bi-factor GPCM analyses led to the same conclusions (see Table 3.10). A priori, this would be unexpected. The residual analyses target so-called uniform CDIF, namely a shift in the item location (item intercept) parameters over countries. The bi-factor analyses target non-uniform CDIF, namely differences in the slopes and the dimensionality across items. The correlations between residuals under the generalized partial credit model and country-specific factor loadings in the bi-factor for components 2, 4 and 5 were moderate, while for component 1, the correlation was much lower, and for component 3, the correlation completely vanished. The result for component 3 is probably because both the residuals and the country-specific factor loadings are poorly estimated for a scale containing only three items.

Though the correlation between the residuals and the country-specific factor loadings is a reasonable estimate between the two measures, it does not properly indicate to what extent the two measures have the same outliers. To investigate this, we ordered and classified the residuals and country-specific factor loadings in three categories according to their size (a category with negative values, a category with positive values and a middle category). Further, we varied the definition of which values were outliers by varying the size of the middle group (assigning it variously as 33, 40 and 80% of values). The calculation of Kappa (Cohen, 1960) establishes the agreement in categorization between the residual analyses using the GPCM and the bi-factor GPCM. This revealed that agreement was poor throughout for component 3, while, for component 1, the agreement was poor in the 33% category; for other categories, agreement was only fair to moderate. In general, the results indicate that it is not a good policy to rely on one approach for the investigation of CDIF.

Table 3.10
Relation Between Residuals Under the Generalized Partial Credit Model and Country-Specific Factor Loadings in the Bi-Factor GPCM

Component	Correlation	Kappa classification CDIF		
		Size middle group		
		33%[a]	40%[b]	80%[c]
1	0.228	0.15	0.20	0.24
2	0.603	0.21	0.29	0.27
3	-0.044	0.07	0.17	0.10
4	0.651	0.46	0.41	0.41
5	0.519	0.34	0.31	0.25

Note. Correlation between the GPCM residuals and the country-specific factor loadings, over countries and items. The content of the components is described in Table 3.1. Size middle group indicates the classification of the ordered residuals and country-specific factor loadings in three categories according to their size: a category with negative values, a category with positive values and a middle category. Norms for Kappa: poor agreement: 0.00–0.19, fair agreement: 0.20–0.39, and moderate agreement: 0.40–0.59 (Landis & Koch, 1977).
[a] Three equally-sized categories.
[b] A middle category contained 40% of the values, the two extreme categories each contained 30%.
[c] A middle category contained 80% of the values, the two extreme categories each contained 10%.

We investigated the influence of CDIF by calculating the correlation and rank correlation between country means estimated with no, 10%, and 20% CDIF parameters, and with random item parameters (Table 3.11). Estimates of the means using the unidimensional GPCM without country-specific item parameters and using the bi-factor GPCM could not be distinguished, so we exclude them from further discussion. In general, correlations were high, indicating that in the estimation of the country means and the rank order of the country means the CDIF had little impact. Component 3 remained the exception; both correlations and rank correlations were

low. Further, for components 2 and 4, the correlations between the means estimated using the GPCM with random item parameters and the other three models were also low, but this was not the case for the rank correlations. This is because the relationship between the means is not linear.

Table 3.11
Correlation and Rank Correlation Between Country Means Estimated with no, 10%, and 20% Cultural Differential Item Functioning Parameters, and Random Item Parameters

Component	Parameter	Correlation			Rank correlation		
		No CDIF	10% CDIF	20% CDIF	No CDIF	10% CDIF	20% CDIF
1	10% CDIF	0.99			0.98		
	20% CDIF	0.99	0.99		0.98	0.98	
	Random	0.98	0.97	0.97	0.97	0.96	0.97
2	10% CDIF	0.99			0.98		
	20% CDIF	0.98	0.99		0.98	0.99	
	Random	0.66	0.64	0.58	0.95	0.93	0.95
3	10% CDIF	0.83			0.94		
	20% CDIF	0.80	0.82		0.93	1.00	
	Random	0.53	0.38	0.33	0.62	0.64	0.63
4	10% CDIF	0.98			0.97		
	20% CDIF	0.97	0.98		0.95	0.95	
	Random	0.50	0.44	0.37	0.94	0.92	0.89
5	10% CDIF	0.97			0.97		
	20% CDIF	0.97	1.00		0.97	1.00	
	Random	0.97	0.98	0.98	0.97	0.99	0.99

Note. The content of the components is described in Table 3.1.

3.4 Discussion and Conclusions

The main purpose of the research presented in this chapter was to develop a psychometric framework aimed at identifying and modelling cultural differential item functioning (CDIF) in parental involvement, as a first step to assess its relationship with student reading literacy. This framework was to shed light on the extent to which there are any cultural differences (differences between countries) in the components that measure dimensions of parental involvement. We developed tools for the identification and modelling of CDIF that were based on five models: the GPCM, GPCM with random item parameters, the GPCM with 10% and 20% country-specific parameters, and the bi-factor GPCM. Firstly, we found that all models clearly and consistently supported the identification of CDIF. However, we also found the results obtained by the models varied. There was reasonable agreement for components 2 (helping with homework), 4 (student's perception of parental involvement) and 5

(school practices for parental involvement from a school perspective). The methods clearly disagreed for component 1 (early literacy activities) and for component 3 (school practices on parental involvement, parent perspective); the latter was likely because of the poor reliability of this component, probably due to the shortness of the instrument. Disagreement in the other four tests is because different aspects of model fit are assessed by the models. In fact, the method using residuals specifically targets uniform CDIF, while the bi-factor GPCM specifically targets non-uniform CDIF. In conclusion, practitioners should therefore better not rely on one model and one approach to investigate CDIF, but diversify in their methods.

In PIRLS, the literacy test and background questionnaires are translated and adapted for each country. Considerable effort is devoted to guaranteeing the international validity of these instruments. For example, the translations and the layout of the instruments are thoroughly reviewed by independent verifiers, and all necessary adaptations are documented in detail. However, it is not unlikely that there are cultural differences in the way respondents interpreted some of the questionnaire items. The main purpose of this study was to establish whether there were any cultural differences in the measurement of parental involvement in PIRLS and, if so, whether correction for these differences led to different results with regard to its relation with reading literacy. Although some of the PIRLS scales for parental involvement may require improvements to increase their reliability, the overall conclusion is that these scales are internationally valid. In the next chapter the scales, both with and without adaptations for CDIF will be related to student reading literacy in order to shed more light on the effects of CDIF in secondary analyses on cross national analyses.

Chapter 4
Modelling Parental Involvement and Reading Literacy: The Relationship Investigated Across Countries.[3]

Abstract

The inconsistent results in the effects of parental involvement on student achievement may be caused by differences between educational systems and cultural differences, or by the great variation in the methods used to assess student achievement and parental involvement across studies. This chapter describes how data from the Progress in International Reading and Literacy Study 2011 is used to explore the relationship between different dimensions of parental involvement and student reading literacy across a large number of countries. Reading literacy was regressed in latent multilevel models on dimensions of parental involvement, where these dimensions were scaled using item response theory both with and without corrections for country-item interactions as described in the previous chapter. Results showed that early literacy activities and help with homework were both related to student achievement, but the impact of parental involvement on reading literacy was not large. Impact differences across countries did prove to be quite large, especially for helping with homework, where the effect of this dimension was found to be stronger in high-achieving countries. The comparison between analyses with and without corrections for cultural differential item functioning (CDIF) in the parental involvement scales indicated that CDIF did not influence the inferences.

---

[3]Based on Chapters 3, 5, and 6 from: Punter, R. A., Glas, C. A. W., & Meelissen, M. R. M. (2016). *Psychometric Framework for Modeling Parental Involvement and Reading Literacy*. Cham: Springer.

4.1 Introduction

Parental involvement is seen as one of the most malleable factors of the student's home situation, though found effects on student performance are not univocal (Punter, Glas, & Meelissen, 2016a). The literature review on dimensions of parental involvement and their relation to student attainment by Punter et al. indicated that parental involvement is generally positively correlated to or has positive effects on student attainment. However, for the individual dimensions of parental involvement, the results were less definitive. One of the explanations for the mixed results the authors point to, is the complexity of the parental involvement concept which has led to a lack of agreement on definitions and measurement inconsistencies, making it troublesome to compare findings across studies. Another limitation that may have contributed to the inconsistent outcomes, is the strong dependency on single-source reports. For example, how parents perceive their involvement at school might differ from the school's perception of the parents' involvement. The authors therefore encourage the use of parent, as well as student and school information, as this capacitates triangulation of essential perspectives on involvement and thereby allows for a more precise determination of parental involvement and its influence on student outcomes. Finally, a point is made by Punter et al. that it is likely that cultural differences in the perception of parental involvement exist. For international comparative studies in education this could mean that different parental perceptions of what is important for the education of their child can also have consequences for how survey questions about parental involvement are interpreted.

Altogether, empirical research is desired into the measurement of student achievement and indicators of the parental involvement, and how comparisons can be made between educational systems (countries), to find out to what extent and under which conditions, parental involvement influences student achievement. In-depth analyses of large-scale international comparative data, such as that contained in the Progress in International Reading and Literacy Study (PIRLS) may provide valuable additions to the research on parental involvement and student achievement. The benefit of using data from an international large-scale assessment study such as PIRLS is not only the richness in data resulting from achievement tests, as well as student, home, teacher and school questionnaires, but also, obviously, the large number of countries for which these data are available.

The previous chapter described how parental involvement scales were constructed from the PIRLS-2011 data, taking potential cultural differences in interpretations of the items, i.e. the potential risk of cultural differential item functioning (CDIF), into account. The present chapter continues by modelling the relation between parental involvement and reading literacy using these parental involvement scales. The research question is: to what extent are the different dimensions of parental involvement related to student achievement in reading literacy, taking into account student background characteristics and differences between countries? To address this question, a multilevel analysis of the PIRLS-2011 data explored the relationship between parental involvement and student reading literacy.

The parental involvement construct was modelled using the five scales constructed from the PIRLS-2011 data, as described in the previous chapter. The scales were modelled both using the generalized partial credit model (GPCM; Muraki, 1992) and by applying the GPCM with adjustment for CDIF by means of fixed item parameters, random item parameters and a bi-factor structure as described in Chapter 3. Having already compared the outcomes for the means on the latent scales for the different corrections for CDIF, comparing the results of the structural multilevel model reveals the extent to which CDIF may influence the relationship between parental involvement and students' reading literacy, and level of control provided by the measurement models.

4.2 Method

4.2.1 Constructs and Countries in PIRLS-2011

Table 4.1 provides an overview of the components and items from PIRLS-2011 used for the analyses in this study. The five variables for the five parental involvement components were identified using the five item response theory (IRT) models considered in the previous chapter: the GPCM, the GPCM with country-specific item parameters for the 10 and 20% most extreme country-by-items interaction, the GPCM with random item parameters and the bi-factor GPCM (for details see Chapter 3). We obtained expected a posteriori (EAP) estimates for the parental involvement constructs and entered these estimates as independent variables into a multilevel regression model. Socioeconomic status (SES) and gender were included as control variables, with SES measured by three indicators. The dependent variable was achievement on the PIRLS-2011 reading comprehension assessment. To account for

the unreliability of this outcome variable, five plausible values are available in the PIRLS dataset. All analyses were repeated for all five plausible values and then aggregated to overall estimates of fixed and random factors, thus incorporating the differences in standard errors for the different effect sizes (Von Davier, Gonzalez, & Mislevy, 2009). The sampling procedure was accounted for by including a student-class weight at the student level and a school weight at the school level. It is important to concurrently use school and student weights in the analyses, because schools were sampled first and then students were sampled within schools. For this study, we considered the data from 41 countries: all countries participating in PIRLS-2011, with the exclusion of countries for which the average achievement was not reliably measured, or that did not administer the home questionnaire (Mullis, Martin, Foy, & Drucker, 2012).

Table 4.1
Overview of Components and Items for Secondary Analyses of PIRLS-2011

Construct	Scale or item (Item in international datasets)
Parental involvement [a]	Component 1: Early literacy activities before beginning primary school Component 2: Help with homework Component 3: School practices on parental involvement, parent perspective Component 4: Parental involvement, student perspective Component 5: School practices on parental involvement, school perspective
Socioeconomic status	About how many books are there in your home? [b] (ASBH14) What is the highest level of education completed by the child's father? [c] (ASBH17A) What is the highest level of education completed by the child's mother? [c] (ASBH17B)
Gender	Are you a girl or a boy? (ASBG01)
Reading literacy	PIRLS reading comprehension assessment (ASRREA01 – ASRREA05)

Note. The international datasets are described in detail in Foy and Drucker (2013).
[a] Details on the parental involvement components are described in Chapter 3 (Table 3.1).
[b] Category labels are: 0 - 1 to 10, 1 - 11 to 25, 2 - 26 to 100, 3 – 101 to 200, 4 - More than 200.
[c] Category labels are: 0 - Did not go to school, 1 - Some ISCED level 1 or 2, 2 - ISCED level 2, 3 - ISCED level 3, 4 - ISCED level 4, 5 - ISCED level 5B, 6 - ISCED level 5A, 7 - Beyond ISCED level 5A, 8 - Not applicable.

4.2.2 Latent Regression Model

The research question was addressed by using a three-level regression model in which students (level 1) were clustered within schools (level 2) and schools were clustered within countries (level 3). Although it is important to recognize that the countries participating in PIRLS-2011 cannot necessarily be regarded as being representative of the whole world, incorporating a country level in our analyses provides some

indication of the variance component of the influence of parental involvement on reading achievement over countries.

First an empty model (Model 0) was estimated to see how the variance in the outcome variable is distributed over the three levels. Subsequently, control variables (student background characteristics) were added as fixed effects (Model 1). The resulting model can be seen as a baseline to which the models including the parental involvement variables of interest can be compared. The separate parental involvement components were added as fixed effects on either the student level (i.e., components 1–4) or school level (component 5), resulting in models 2A–2E. We also created a model that included all five components simultaneously (Model 3).

By entering the five components as fixed factors, each factor was assumed to have the same effect across all countries. However, in the context of this study, we also wanted to determine the extent of differences in the effects of parental involvement across countries. Therefore, we also considered a model with random slopes at the country level for the parental involvement components (Model 4). A random-slopes model includes a variance component for the slope of one or more predictor variables, while the other models may be considered special versions obtained by fixing parameters. The full model, a random-intercepts-and-slopes model, with one component for parental involvement is given by

$$Y_{ijk} = \beta_{0\,jk} + \beta_1 \text{Gender} + \beta_2 \text{Books} + \beta_3 \text{Education} + \beta_{4\,jk} \text{Construct} + \varepsilon_{ijk}, \quad (4.1)$$

where ε_{ijk} is a normally distributed error term with variance $VAR\left(\varepsilon_{ijk}\right) = \sigma^2$ that is independent over students *i*, schools *j*, and countries *k*. The first term on the right-hand side is a random intercept, decomposed as $\beta_{0\,jk} = \gamma_{000} + u_{0\,jk} + v_k$, where γ_{000} is the grand mean, and the other two terms are independent normally distributed error components with mean zero and variance $VAR(u_{0\,jk}) = \xi_0^2$ and $VAR(v_k) = \tau_0^2$. The regression coefficients β_1, β_2 and β_3 pertain to gender, the number of books in the home, and the highest educational level attained by one of the parents, respectively. The regression coefficient for the parental involvement component is decomposed as $\beta_{4\,jk} = \gamma_{400} + u_{4\,jk} + u_{4k}$, where γ_{400} is the average slope for the component over all countries and schools, and the two error terms have variances $VAR(u_{4\,jk}) = \xi_4^2$ and

$VAR(u_{4k}) = \tau_4^2$, respectively. Finally, random intercepts and random slopes are allowed to covary; at the country level this leads to a parameter $COV(v_k, u_{4k}) = \tau_{04}^2$.

The fixed effects models are obtained by setting $VAR(u_{4k}) = \tau_4^2 = 0$, and the baseline models, Model 0 and Model 1, are obtained by removing the appropriate predictors.

To keep the model interpretable and relevant, only the components showing a meaningful effect in the fixed model were entered as covariates in the random-intercepts-and-slopes model.

The final step of the analyses was to assess the impact of CDIF by also estimating the final model with the EAP estimates from the IRT models correcting for CDIF in the parental involvement scales as presented in the previous chapter, to see if outcomes differ. All analyses were conducted using the Mplus software package, version 7.11 (Muthén & Muthén, 1998-2012).

4.3 Results

We first modelled the relation between the parental involvement scales and student reading literacy using the GPCM without correction for CDIF in the parental involvement scales. Model 0 indicates that most of the variance in student achievement in reading literacy was situated at the student level (44%, Table 4.2). Differences between countries were also considerable: 39% of the variance could be accounted for by between-country differences. This was to be expected based on the large range of average country scores reported in the international report of PIRLS-2011 (Mullis, Martin, Foy, & Drucker, 2012).

As expected, Model 1 indicated that gender and the two SES-indicators are important predictors of reading literacy. On average, girls outperformed boys by almost 13 points on the PIRLS test. The number of books at home and the educational level of the parents are both positively related to reading achievement. The three background variables explain a considerable amount of variance; 40% at school level and 61% at country level. This suggests that a substantial part of the differences in achievement scores between PIRLS countries can be attributed to individual differences in student background characteristics. However, within countries, individual differences play a

minor role, because they only explain 10% of the variance of the students within a country.

Table 4.2
Effects of Student Background Characteristics and Components of Parental Involvement on Reading Literacy Achievement of Grade 4 Students in 41 PIRLS Countries, Random Intercept Model, Without Correction for Cultural Differences

Effects	Model 0 Empty model		Model 1 Student background characteristics	
	Effect	*SE*	Effect	*SE*
Fixed effects				
Intercept γ_{000}	525.49	7.12	541.75	6.30
Male difference β_1			-12.83	.85
Books at home (low-high) β_2			8.55	.49
Parental education (high-low) β_3			-13.50	.77
Random effects				
Variance between students σ^2	4305.31 (44%)	168.27	3878.90 (61%)	165.14
Variance between schools ξ_0^2	1658.19 (17%)	269.24	1001.59 (16%)	177.31
Variance between countries τ_0^2	3887.53 (39%)	.30	1497.43 (23%)	327.48
Explained by predictors				
At student level			10%	
At school level			40%	
At country level			61%	

In models 2A–2E, we explored the fixed effects of the different components of parental involvement, taking into account the effect of the three background variables (Table 4.3). The effect sizes of the background variables did not change noteworthy when the different components of parental involvement were included in the model. Parental report of literacy activities before their child starts in first grade, and helping with homework were both related to student reading literacy, but each in a different way. Students of parents reporting spending more time on early literacy activities with their child showed higher achievement levels than those whose parents spent less time on these activities (the scale runs from high involvement to low involvement, therefore the effect in Table 4.3 appears as a negative score). With regard to helping with homework, there is a negative relationship (this scale also runs from high involvement to low involvement, therefore the effect in Table 4.3 appears positive).

Because of the large number of respondents recorded in the data (over 200,000 students), each relationship with achievement, even when very weak, is significant. Therefore, the relevance of these relationships was assessed in terms of changes in the mediated achievement score as a function of parental involvement. The standard deviation of the EAP estimates of early literacy activities was 0.97. If parents'

perceptions of the time they spent on early literacy activities increased one point (i.e., from average to one standard deviation above the mean), the score of the student on the PIRLS test increased by nine points as can be inferred from the regression coefficient of -9.03 (Table 4.3). On a scale with a mean of 500 and a standard deviation of 100, this could be considered a small effect. This also applies to the negative association between helping with homework and reading achievement. The standard deviation of this component was also 1, and the reduction in student achievement was 9.3 if parents reported that they spent one standard deviation more in time in helping their child.

At first glance, it seems that students whose parents had less positive views about school practices outperformed the classmates whose parents held more positive views. However, as the standard deviation of the EAP estimates on this scale was 1.6 and the regression coefficient was 5.15, the increase in scores per standard deviation was only three points. The same was true for students' perception of parental involvement (an increase of three points) and school perception of parental involvement (decrease of almost three points). These are very small effects.

Models 2A–2E revealed that for each of the five components, the percentages suggested hardly any alteration in the variance, as compared with Model 1. Thus, in Model 3, we entered the fixed effects for all components of parental involvement simultaneously (Table 4.4). The influence of early literacy activities and helping with homework increased slightly when the effects of the other components were held constant; variance increased to 15% at student level, 46% at school level, and 69% at country level.

Table 4.3

Effects of Student Background Characteristics and Components of Parental Involvement on Reading Literacy Achievement of Grade 4 Students in 41 PIRLS Countries, Random Intercept Model, Without Correction for Cultural Differences

Effects	Model 2A Early literacy activities		Model 2B Help with homework		Model 2C Parent's view of school practices		Model 2D Student's perception		Model 2E School's perception	
	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*	Effect	*SE*
Fixed effects										
Intercept γ_{000}	542.69	6.20	541.53	5.88	541.89	5.81	541.62	6.36	541.59	6.34
Male difference β_1	-11.73	.83	-12.73	.84	-12.73	.86	-12.43	.82	-12.78	.85
Books at home (low-high) β_2	7.17	.51	8.72	.48	8.32	.50	8.53	.49	8.52	.50
Parental education (high-low) β_3	-12.78	.72	-13.51	.79	-13.26	.75	-13.53	.75	-13.53	.78
Early literacy activities before primary school (high-low) γ_{400}	-9.03	.57								
Help with homework (high-low) γ_{500}			9.32	1.20						
Parent's view of school practices on parental involvement (high-low) γ_{600}					5.15	1.75				
Student's perception of parental involvement (high-low) γ_{700}							-2.64	.68		
School's perception of parental involvement (high-low) γ_{800}									3.10	.74
Random effects										
Variance between Students σ^2	3825.10	161.13	3813.42	171.89	3849.27	164.10	3861.90	161.85	3886.61	162.96
Variance between schools ξ_0^2	977.68	173.60	993.75	176.31	977.54	172.83	977.60	172.91	978.71	168.07
Variance between countries τ_0^2	1469.37	319.06	1395.61	330.34	1301.80	315.67	1507.00	326.07	1545.41	346.23
Explained variance by predictors										
At student level	11%		11%		11%		10%		10%	
At school level	41%		41%		41%		41%		41%	
At country level	62%		64%		67%		61%		60%	

Table 4.4
Effects of Student Background Characteristics and Parental Involvement on Reading Literacy Achievement of Grade 4 Students in 41 PIRLS Countries, Random Intercept Model, Without Correction for Cultural Differences

Effects	Model 0 Empty model		Model 1 Student background characteristics		Model 3 Parental involvement	
	Effect	*SE*	Effect	*SE*	Effect	*SE*
Fixed effects						
Intercept γ_{000}	525.49	7.12	541.75	6.30	541.82	5.46
Male difference β_1			-12.83	.85	-10.45	.81
Books at home (low-high) β_2			8.55	.49	6.60	.48
Parental education (high-low) β_3			-13.50	.77	-12.15	.70
Early literacy activities before primary school (high-low) γ_{400}					-12.91	.56
Help with homework (high-low) γ_{500}					13.08	1.14
Parent's view of school practices on parental involvement (high-low) γ_{600}					4.12	1.50
Student's perception of parental involvement (high-low) γ_{700}					-4.09	.49
School's perception of parental involvement (high-low) γ_{800}					2.7	.75
Random effects						
Variance between students σ^2	4305.31 (44%)	168.27	3878.90	165.14	3675.41	163.91
Variance between schools ξ_0^2	1658.19 (17%)	269.24	1001.59	177.31	888.24	153.69
Variance between countries τ_0^2	3887.53 (39%)	0.30	1497.43	327.48	1190.78	346.23
Explained variance by predictors						
At student level			10%		15%	
At school level			40%		46%	
At country level			61%		69%	

The next step was to estimate two models with random slopes at country level for early literacy activities and helping with homework, to determine whether the effects of these components of parental involvement differed across countries. While recognizing this is still open for discussion, we considered these two components as showing a small, but meaningful relation with reading achievement. We included the three background variables as fixed effects in the random effect model (Table 4.5).

For early literacy activities there is a very small difference in the average overall effect, from –9.0 in Model 2A to –8.7 in Model 4A. The variance over countries is 14.57; relative to the total variance in the outcome variable this is very small, but relative to the effect of early literacy activities, the effect is clearly larger. Ninety-five percent of the range of the slope over countries lay roughly between –16.3 and -1.0.

Table 4.5
Effects of Student Background Characteristics and Parental Involvement on Reading Literacy Achievement of Grade 4 Students in 41 PIRLS Countries, Random-Intercepts-and-Slopes Model, Without Correction for Cultural Differences

Effects	Model 4A Early literacy activities		Model 4B Help with homework		Model 4C Component 1 + 2	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Fixed effects						
Intercept γ_{000}	542.01	6.42	538.54	5.99	538.67	5.78
Male difference β_1	-11.72	0.84	-12.52	0.84	-10.93	0.83
Books at home (low-high) β_2	7.19	0.51	8.45	0.45	6.58	0.45
Parental education (high-low) β_3	-12.76	0.72	-12.96	0.78	-11.97	0.72
Early literacy activities γ_{400}	-8.67	0.57			-12.66	0.68
Help with homework γ_{500}			11.75	1.38	15.15	1.40
Random effects						
Students σ^2	3789.18	160.03	3741.12	171.88	3618.13	16.048
Variance intercepts schools ξ_0^2	960.26	170.61	979.91	180.05	930.05	171.372
Variance slopes schools ξ_4^2	33.62	6.04			30.86	5.57
Variance slopes schools ξ_5^2			33.47	6.79	22.74	4.86
Variance intercepts countries τ_0^2	1386.39	292.43	1402.06	321.18	1447.48	371.59
Variance slopes countries τ_4^2	14.57	3.80			20.09	5.59
Covariance intercepts and slopes τ_{04}^2	27.25	28.10			41.40	30.56
Variance slopes countries τ_5^2			74.42	14.32	73.96	14.25
Covariance intercepts and slopes τ_{05}^2			109.23	47.79	134.36	48.19

A covariance of 27.25 indicates a relationship between the intercept of a country and the steepness of the slope within a country. A positive covariance means that the relationship between parental involvement and reading achievement is stronger in countries that performed strongly in the PIRLS test; a negative covariance means that the association between the predictor and dependent variable becomes stronger as the country average of reading achievement decreases. The standard error of the covariance between intercept and slope for early literacy activities was larger than the covariance itself, indicating that there was no relation between the intercept and slope. From the variance components and the covariance, we obtained a correlation of 0.19, which must be considered small to moderate.

For helping with homework, the average effect size increased from 9.3 in Model 2B to 11.8 in Model 4B. The variance of the slope over countries was 74.42; 95% of the range of the slope over countries lay between –5.5 and 29.0. Further, a positive covariance of 109.23 led to a correlation of 0.34, which is also substantial. As this scale runs from high involvement to low involvement, although the effect reported was positive, in truth it is a negative effect (more help equals lower achievement). The

positive covariance suggests that this negative association of helping with homework with achievement was stronger in high-performing countries.

To assess the impact of CDIF, we replicated the last analysis (Table 4.5) with the EAP estimates of the latent parental involvement parameters from all five IRT models (Table 4.6). Estimates from all models were very close and never more than one standard error away from the estimates under the GPCM. We conclude that CDIF did not bias the inferences.

Table 4.6
Random-Intercepts-and-Slopes Model for Effects of Student Background Characteristics and Parental Involvement on Reading Literacy Achievement, Without and With Correction for Cultural Differences

Effects	GPCM		GPCM 10% Split	GPCM 20% Split	Random GPCM	Bi-factor GPCM
	Estimate	*SE*	Estimate	Estimate	Estimate	Estimate
Fixed effects						
Intercept γ_{000}	538.67	5.78	539.01	539.31	539.39	538.27
Male difference β_1	-10.93	0.83	-10.91	-10.90	-10.92	-11.11
Books at home (low-high) β_2	6.58	0.45	6.59	6.58	6.58	6.74
Parental education (high-low) β_3	-11.97	0.72	-11.99	-11.98	-11.98	-12.10
Early literacy activities γ_{400}	-12.66	0.68	-12.58	-12.56	-12.52	-12.106
Help with homework γ_{500}	15.15	1.40	15.05	15.51	15.59	13.084
Random effects						
Students σ^2	3618.13	16.048	3619.85	3615.16	3621.08	3628.78
Variance intercepts schools ξ_0^2	930.05	171.372	927.67	928.90	928.49	931.63
Variance slopes schools ξ_4^2	30.86	5.57	30.36	30.91	29.84	29.07
Variance slopes schools ξ_5^2	22.74	4.86	25.34	26.09	24.91	23.49
Variance intercepts countries τ_0^2	1447.48	371.59	1469.70	1418.14	1463.75	1347.119
Variance slopes countries τ_4^2	20.09	5.59	20.47	20.52	21.49	25.418
Covariance intercepts and slopes τ_{04}^2	41.40	30.56	40.32	32.40	40.43	21.73
Variance slopes countries τ_5^2	73.96	14.25	72.36	82.66	85.51	47.521
Covariance intercepts and slopes τ_{05}^2	134.36	48.19	138.30	132.56	145.50	101.94

4.4 Discussion and Conclusions

In the overall goal of providing a framework for assessing the relation between parental involvement and reading literacy, the previous chapter focused on the modelling of dimensions of parental involvement and ways to address the issue of CDIF in these scales. This chapter further extended the framework by relating the previously constructed scales for parental involvement to student reading literacy assessment data. The research question at hand was: to what extent are the different dimensions of parental involvement related to student achievement in reading literacy, taking into account student background characteristics and differences between

countries? Multilevel analyses were conducted to explore the association between parental involvement and student achievement for all countries that participated in PIRLS-2011. A three-level (student, school and country) random intercept model was explored, as well as a random three-level model.

The results of the three-level models with a random intercept showed that, controlled for student's gender and SES, and taking into account between-schools and between-countries variance, there is a rather weak but positive relationship between early literacy activities and student achievement in reading literacy at Grade 4. We may here only confirm a positive association and cannot make any claims about causality, as PIRLS is cross-sectional. The results only indicate that other types of studies (experimental studies) measuring the real effects of early literacy activities on reading achievement are relevant, assuming that there is agreement among scholars in how these activities should be measured.

Both early literacy activities and helping with homework are home-based activities, confirming that what parents do at home with their child is important for student achievement. In this study we found school-based involvement from the perspective of the school (component 5) had negligible effect. As the constructed scale for school-based involvement from the perspective of the parent (component 3) turned out to be unreliable (Chapter 3), we are unable to draw valid conclusions for this component regarding its relationship with student achievement.

Overall, the impact of parental involvement on reading literacy is not large. When all five components were entered into the model, it explained approximately 15% at the student, 46% at the school level, and 69% at the country level. However, the impact differences across countries proved to be quite substantial, especially for helping with homework, where regression coefficients, with a mean value of 11.8, range over countries from -5.5 to 29.0.

From the positive correlation between the country-level intercept and slopes for helping with homework we conclude that in low-achieving PIRLS countries, the effect of helping with homework is smaller than in high-performing countries. This means that, in exploring the achievement effect of helping with homework, the educational context should be taken into account. The sometimes-contradictory results of earlier studies on this subject may also be explained by such differing effects between

countries. Another explanation for the positive correlation between intercepts and slopes on the country level would be that, in low-achieving countries, parents' reading competency will also be low, so parents are themselves less able to read and hence provide effective support. However, it is beyond the possibilities of the present research to draw conclusions in this respect.

Finally, analysing the influence of CDIF on the estimates of country means and on the outcomes of latent regression analyses led to the conclusion that CDIF did not influence the results.

Chapter 5
An IRT Model for the Interaction Between Item Properties and Group Membership: Reading Demand in the TIMSS-2015 Mathematics Test[4]

Abstract

An IRT model is presented that combines multiple angles to detect and model differential item functioning in large-scale assessments. The approach starts with the generalized partial credit model as a baseline. Then two generalizations follow: a bi-factor model and a model where item parameters are regressed on student and item characteristics. Finally, both models are combined into one overall model. An empirical investigation was done on data from TIMSS-2015 mathematics test from four European countries, with reading demand classifications as the item properties of interest for students not speaking the test language at home. Results showed that the full model fitted best, indicating that reading demand is a factor to take into account for students not speaking the test language at home.

---

[4]Based on Punter, R. A., Meelissen, M. R. M., Eggen, T. J. H. M., & Glas, C. A. W. An IRT model for the interaction between item properties and group membership: Reading demand in the TIMSS-2015 mathematics test. Submitted for publication.

5.1 Introduction

In international large-scale assessments (ILSA), differences in item response behaviour for students of equal proficiency can complicate inferences regarding proficiency differences between countries or specific student populations. These differences in response behaviour are often labelled as differential item functioning (DIF) and may be attributable to the items' framing. Many achievement tests in ILSA apply contextualized items to test the targeted construct in a meaningful, real-life setting. Although it increases test validity to have students applying their skills in a variety of contexts, it may also evoke DIF, for example by the reading demand posed by contextualized mathematics items. In this study, we present an item response theory (IRT) based approach to identify and handle DIF and apply this approach to the Trends in International Mathematics and Science Study (TIMSS) 2015 mathematics data to explore the potential DIF due to reading demand in mathematics items for students with different language backgrounds.

5.1.1 IRT Perspectives for Modelling of DIF

The detection and handling of DIF has received a lot of attention in the past decades (Holland & Wainer, 1993). In the present article we focus on parametric IRT-based methods for DIF detection over more general techniques such as the Mantel-Haenszel statistic (Holland & Thayer, 1988). This allows for modelling DIF in addition to identifying it, so (sub)populations can still be compared on their latent proficiency.

5.1.2 Group-Specific Item Parameters

Detection and modelling of DIF in the IRT framework can be approached from two angles. One is viewing DIF as related to item properties and modelling DIF using virtual item parameters that are allowed to vary across groups, or regressing item parameters on item properties, as far as they are available (Glas, 1998; Glas & Jehangir, 2014). Glas and Jehangir present a method to identify the strongest cases of DIF using a Lagrange Multiplier test statistic. For these items, group-specific item parameters are estimated. The method is motivated by the assumption that a substantial part of the items function the same in all groups and a limited number of items have DIF. In the IRT model, it is assumed that all items pertain to the same latent variable. Items without DIF have the same item parameters in every group. However, items with DIF have item parameters that differ across groups. These items refer to the same latent variable as all other items, but their location on the scale differs across groups. This is

a bottom-up approach as no a priori expectations of DIF are taken into account, other than the group specification. Also, it assumes that the structure of the latent trait measured by the test is equal for all groups.

This approach is in line with the explanatory approach presented in De Boeck & Wilson (2004), which seeks to relate item responses to variables pertaining to student or item characteristics in a context of generalized linear and non-linear mixed models. An example of this approach is the linear logistic test model (Fischer, 1973). This model does not specifically address the issue of DIF, but it does focus on the item difficulty parameters by placing linear constraints on the item parameters. As demonstrated by Fischer, the regression of item parameters on item properties enables one to investigate hypothesized psychological functioning of items, such as processes students go through when answering a mathematics item. In the context of DIF, this approach helps to explain the functioning of DIF based on item characteristics.

5.1.3 Multidimensionality

In addition to focusing on DIF as differences in item difficulty parameters between groups, DIF can be viewed as related to differences in the multidimensional proficiency distributions of distinguished groups. Though scores are often modelled by a unidimensional model, the student's score may actually represent a composite of abilities (Ackerman, 1992). When groups of students differ in these latent abilities, but only a single score is reported, DIF can occur (Roussos & Stout, 1996). Shealy and Stout (1993) presented a multidimensional model for DIF where DIF is due to (1) an item being sensitive to both the construct the item intends to measure and a secondary, confounding nuisance dimension, and (2) a difference exists between subpopulations in the secondary construct, given their proficiency on the target construct.

Contrary to this traditional perception of the additional factors as construct-irrelevant and the exploratory fashion in which the analyses are often conducted, Walker and Beretvas (2001) demonstrate the inclusion of more dimensions as intentional, contributing to a more authentic multidimensional representation of the construct of interest. They advocate that since test developers are confronted with items that show DIF for no apparent reason (Angoff, 1993), DIF methodology needs to be considered

that hypothesizes substantive reasons for the occurrence of DIF at forehand, preferably in a multidimensional framework.

5.1.4 Proposed Model

In this article, we propose an IRT model that combines the two described DIF approaches: a model that considers both multidimensionality in the latent trait by applying a bi-factor structure, as well as an item difficulty effect by allowing for an interaction between student background characteristics and specific properties of a test item. The approach starts with using the generalized partial credit model (GPCM; Muraki, 1992) as a base line. The GPCM, which is commonly used in ILSA for free response items requiring partial credit scoring, assumes one latent variable needed to explain response behaviour. Furthermore, it defines a parameter related to an item's discriminating ability and response category threshold parameters providing information on the salience of response alternatives. It is a generalization of the partial credit model (Masters, 1982) which assumes the discrimination parameters for all items are constrained to be equal. Other IRT models suitable for polytomously scored items are the sequential model (Tutz, 1990) and the graded response model (Samejima, 1969). Since the response curves of these models are hard to discern based on empirical data (see for instance Verhelst, Glas, & de Vries, 1997), the choice for the GPCM is not fundamental for the present application.

In the proposed model, the first generalization of the GPCM allows for differences in proficiency distributions between populations. This model is closely related to the bi-factor model (Gibbons & Hedeker, 1992) where each item is an indicator of a general dimension and one of several other dimensions. Here, each item is an indicator of a general dimension the test sets out to measure, and one group-specific secondary dimension. In this way, the model incorporates potential group-specific differences in the dimensionality to which the test pertains, while still being able to measure and compare the proficiency level on the main component. Given its latent structure, this model will be referred to as the "bi-factor GPCM". In the second generalization, our so-called "hybrid model", the item parameters are regressed on an interaction between item properties and student characteristics. This content-by-group interaction provides information on the extent to which item characteristics may contribute to DIF. Finally, both models are combined into a full model.

5.1.5 Reading Demand in TIMSS-2015 Mathematics

The approach is exemplified with data from the TIMSS-2015 mathematics test. The ILSA TIMSS measures the achievement of students in Grade 4 and 8 in science and mathematics, and has its framework organized around content and cognitive dimensions. The mathematics test relies strongly on the use of word problems to cover the broad domain of the math curriculum to be assessed. To ensure that the linguistic features of the items are appropriate for the student population, the contextualized mathematics items are formulated concise and succinctly (Mullis & Martin, 2013). The items nevertheless demand sufficient reading skills to handle the required reading demand. This reading demand in testing mathematics proficiency might lead to DIF in the TIMSS items for students with different language backgrounds. In this application the proposed model is estimated to evaluate the potential confounding role of language in the TIMSS-2015 mathematics test for students not speaking the test language at home.

The required reading demand in a mathematics test may particularly affect student subpopulations with different language backgrounds such as Second Language Learners (SLL), described by Barwell (2012, p.147) as "students from linguistic minority backgrounds whose home languages are not well-represented or recognized in wider society". Many studies have focused on differences in achievement between SLL and non-SLL students, finding consistently lower achievement for SLL students. Andon, Thompson, and Becker (2012) focused on students in large-scale assessment studies and reported underperformance of SLL students in the Programme for International Student Assessment (PISA) and the Progress in Reading Literacy Study (PIRLS). Furthermore, TIMSS-2015 showed that in most of the participating countries, in particular European countries, students reporting to speak the language of the test at home only sometimes, tended to score lower than students always speaking the test language at home (Mullis, Martin, Foy, & Hooper, 2016).

The issue of reading demand in word problems was addressed using TIMSS data when in 2011 the main data collection for TIMSS and PIRLS coincided (Martin & Mullis, 2013). In 34 countries the same students participated in both studies, providing the opportunity to further research the relationship between reading achievement and achievement in science and mathematics. To study the effects of reading demand, the TIMSS test items were globally classified as low, medium or high in reading demand.

The classification was based on four indicators of reading difficulty: (1) number of words, (2) number of different symbols, (3) number of different specialized vocabulary words, and (4) total number of elements in the visual displays. Martin and Mullis showed that performance on the TIMSS test of students with lower reading proficiency was influenced by the reading demands of the test items, but not for students with a high reading proficiency.

Several other studies found DIF in mathematics items for students with different language backgrounds. Banks, Jeddeeni, and Walker (2016) compared probabilities of answering mathematical word problems correctly between SLL students and non-SLL students of equal math proficiency. They applied the nonparametric SIBTEST approach developed by Shealy and Stout (1993) and found a significantly lower average total score for SLL students yet no persistent DIF against SLL students. Abedi and Lord (2010) reduced the linguistic complexity of items from the large-scale assessment NAEP and found that SLL students and low SES students in grade eight benefited most from these linguistic modifications. This suggests an interaction effect between language demand in a content area assessment and student's linguistic and socioeconomic background (Abedi & Lord, 2010). In a fourth-grade state mathematics test Martiniello (2009) found that greater item linguistic complexity corresponded with greater differences in IRT difficulty parameter estimates, favouring non-SLL over SLL students. In addition to that, SLL students' think-aloud interviews in solving the DIF items indicated that linguistic complexity is a source of DIF for SLL students (Martiniello, 2009). Haag, Heppt, Stanat, Kuhl, and Pant (2013) used the same operationalization of DIF as Martiniello and found that math word problems on a third-grade state math test which were more difficult for German students who heard and/or spoke a language other than German at home.

In this study we explore the proposed IRT model to investigate whether differences in mathematics achievement between students with different language backgrounds can be explained by reading demand in the items. This is studied for four west-European countries (Denmark, Flemish Belgium, Germany, and the Netherlands) that showed a large achievement gap on the TIMSS-2015 mathematics scale between students that always and students that only sometimes speak the test language at home (Mullis, Martin, Foy, et al., 2016). We model it in two ways; by applying a group-specific secondary component that is unrelated to the main factor (the bi-factor GPCM) and

by including an interaction effect between student language background and item reading demand on the item difficulty (the hybrid model). The contribution of this analysis does not lie in a mere statistical description of item parameters, but rather in providing insight into the role of reading demand in the mathematics test. The following research questions are addressed:

1. How are the TIMSS-2015 mathematics items distributed across levels of reading demand?
2. To what extent are there student differences on the latent dimension in the GPCM with respect to language spoken at home in four west-European countries?
3. To what extent are there student differences on the main dimension, when a bi-factor structure is in place?
4. Are the factor loadings on the secondary component suggestive of the noise component pertaining to reading demand?
5. Can differences in performance between students that do and do not speak the testing language at home be attributed to the reading demand of the items?
6. Which model fits the TIMSS-2015 mathematics test better in terms of the DIC: a model that incorporates a second factor reflecting noise in the proficiency of students due to the linguistic demands, a model reflecting changes in item difficulty due to their linguistic demands, or a model reflecting both effects?

5.2 The Model

Here we present the full model from which the other models can be obtained by invoking certain restrictions. The measurement model pertains to a polytomously-scored response Y_{ni} of a student *n* to an item *i*. The item scores range from *0* to M_i and the score of student *n* on item *i* is denoted by the variables $x_{nij}(j=0,...,M_i)$ where $x_{nij}=1$ if the response is in category *1* and *0* otherwise. Note that M_i has an index *i*, which indicates that the maximum score *M* of an item can differ among items. In the full model (Model 4) the probability of scoring in category j $(j=0,...,M_i)$ is given by

$$P(Y_{ni}=m\mid\theta_n)=\frac{\exp\left(m\left(\alpha_{i0}\theta_{n0}+\alpha_{ig(n)}\theta_{ng(n)}-\sum_{p,q}\gamma_{pq}z_{np}x_{iq}\right)-\sum_{j=1}^{m}\beta_{ij}\right)}{1+\sum_{h=1}^{M_i}\exp\left(h\left(\alpha_{i0}\theta_{n0}+\alpha_{ig(n)}\theta_{ng(n)}-\sum_{p,q}\gamma_{pq}z_{np}x_{iq}\right)-\sum_{j=1}^{h}\beta_{ij}\right)} \quad (5.1)$$

with $z_{np}=\begin{cases}1 \text{ if student } n \text{ belongs to a particular group } p\\ 0 \text{ otherwise,}\end{cases}$

and $x_{iq}=\begin{cases}1 \text{ if item } i \text{ has item characteristic } q\\ 0 \text{ otherwise,}\end{cases}$

where θ_{n0} is the score of a student n on the latent scale pertaining to the main latent construct (the proficiency that the test intents to measure), $\theta_{ng(n)}$ is the score on a second latent dimension, and the index $g(n)$ indicates the group to which student n belongs. Further, α_{i0} and $\alpha_{ig(n)}$ are the factor loadings of item i on these two dimensions, and β_{ih} $(h=1,...,m_i)$ are the item response category threshold parameters. The variables x_{iq} are dummy codes for item characteristics, z_{np} are dummy codes for a particular student background characteristic. The associated regression coefficients for the interaction between z_{np} and x_{iq} are γ_{pq}.

It is assumed that within each group the dimensions θ_0 and θ_g have a bi-variate normal distribution $N(\theta_{n0},\theta_{ng};\boldsymbol{\mu}_{\mathbf{g}},\Sigma_g)$. For the two-dimensional group mean $\boldsymbol{\mu}_{\mathbf{g}}=(\mu_0,\mu_g)$, it holds that the mean on the second dimension is fixed at zero, that is $\mu_g=0$. The covariance matrix is given by $\Sigma_g=\begin{bmatrix}\sigma_g^2 & 0\\ 0 & 1\end{bmatrix}$. To further identify the model, the latent proficiency scale has a standard normal proficiency distribution for one of the groups.

From this overall model, a unidimensional GPCM (Model 1; Muraki, 1992), assuming only one latent variable, is obtained by dropping the group-specific dimensions $\theta_{ng(n)}$ and the associated factor loadings $\alpha_{ig(n)}$ presented in Eq. (5.1), as well as the

regression coefficients γ_{pq}. The latent student parameters θ_0 have a univariate normal distribution with a mean μ_g and a variance σ_g^2.

A bi-factor GPCM (Model 2) is defined by the inclusion of a group-specific second latent component, which in general can be regarded as a noise component. The factor loadings $\alpha_{ig(n)}$ on this second latent dimension index the extent to which items load on this extra dimension for a specific student group. The model results from Eq. (5.1) by restricting $\sum_{p,q} \gamma_{pq} z_{np} x_{iq}$ to zero.

Model 3, the hybrid model, includes specific item characteristics. In this model, the coefficients γ_{pq} indicate the effect on the item location parameter of a specific item characteristic for students in a particular subpopulation. It is therefore an interaction between the item characteristic and the student characteristic. The model is defined by Eq. (5.1) by dropping the group-specific dimensions $\theta_{ng(n)}$ and the associated factor loadings $\alpha_{ig(n)}$.

The models were estimated in a Bayesian framework. This framework treats parameters as random variables and allows for the concurrent estimation of the item parameters and estimates regarding students and subpopulations. Estimates of the parameters come from the multivariate conditional distribution of the parameters (i.e., the posterior distribution), which naturally incorporates uncertainty from one subset of random variables or parameters into inferences for another subset (Platz & Junker, 1999). As Platz and Junker point out, this extra uncertainty can be especially important when more complex IRT models are applied, resulting in a relatively small ratio of examinees to parameters.

The models were estimated using the OpenBUGS software (Lunn, Spiegelhalter, Thomas, & Best, 2009) which is one among several open source packages that provide a general-purpose Markov chain Monte Carlo (MCMC) sampler to estimate complex IRT models in a Bayesian framework. Models were compared on their fit with the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van der Linde, 2002) that is provided by the software.

The script for estimation of the overall model is provided in Appendix B. The MCMC sampling procedure in the application were run with over 50000 iterations for each of two chains. Each chain had a burn-in of 20000 iterations. These large numbers were chosen to ensure convergence and sufficiently small estimation errors. The sampling procedures were checked for convergence by visual inspection of the trace plots and by comparing the estimates of the two chains. Also, we inspected the MCMC errors to establish that the error associated with the point estimates were acceptable small (<5% of estimated standard deviation; Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012).

5.3 Application to the TIMSS-2015 Mathematics Test

5.3.1 Dataset

Response data from the TIMSS-2015 mathematics test from Denmark, Germany, the Netherlands and Belgium (Flanders) were used, together with data from the student question "*How often do you speak <<language of the test>> at home?*". Based on this latter question, two groups of students were created: students in Language Group 1 (LG1) indicated to speak the test language at home "always" or "almost always" (the non-SLL group), whereas students in Language Group 2 (LG2) indicated to speak the test language "sometimes" or "never" at home (the SLL group).

5.3.2 Item Coding

The TIMSS-2015 mathematics items were classified as either (1) low, (2) medium, and (3) high in reading demand according to the guidelines for classification of the TIMSS-2011 items (Martin & Mullis, 2013). The items were also scored on four indicators of reading demand, previously identified to influence reading demand of a test item by Martin and Mullis (2013). Discrepancies in scoring and classification of the new items between scorers were resolved among the scorers. A discriminant analysis on the reading demand indicators was conducted to validate the holistic classification of the items.

5.3.3 Item Classification on Reading Demand

Table 5.1 shows the number of items coded as low, medium, or high in reading demand for the trend, new and total set of TIMSS-2015 mathematics items. The proportion of items with high reading demand in the newly added mathematics items was somewhat higher than in the trend items.

Table 5.1
Classification of TIMSS-2015 Mathematics Items According to Reading Demand, Number of Items per Level

Level of reading demand	Trend items (%)	New items (%)	All items (%)
Low	36 (35)	18 (27)	54 (32)
Medium	37 (36)	25 (37)	62 (37)
High	29 (28)	24 (36)	53 (31)
Total	102 (100)	67 (100)	169 (100)

A discriminant analysis on the reading demand indicators was conducted to validate the global classification of the items. As for the 2011 items (Martin & Mullis, 2013), the first discriminant function was sufficient to discriminate between the item groups (2011: Wilks' Lambda = .260, p<.001, 2015: Wilks' Lambda = .268, p<.001). The indicator loading most heavily on the first discriminant function was, as in 2011, the number of words.

Table 5.2 shows how the items with different levels of reading demand were distributed across the cognitive and content domains known in TIMSS. There appears to be a relation between the cognitive domain and the level of reading demand as items belonging to "knowing" were more often classified as low in reading demand, whereas items in the domains "applying" and "reasoning" were mostly categorized as medium and high in reading demand. Items in the content domain "numbers" usually demanded the least reading. Items in the "data display" domain were most textual: none of these items were labelled as low in reading demand.

Table 5.2
Distribution of Low, Medium and High Reading Demand Items per Cognitive and Content Domain

	Reading demand		
Cognitive domain	Low	Medium	High
Knowing	36	18	10
Applying	15	31	26
Reasoning	3	13	17
Content domain			
Number	41	32	16
Data Display	0	2	22
Geographic Shapes and Measures	13	28	15

5.3.4 Item Response Modelling

Model 1.

First, we fitted the unidimensional GPCM (Model 1) to the response data from students in both LG1 and LG2 (Table 5.3). From the comparison of both groups it becomes clear that in each of the four countries students in LG2 scored lower on the

mathematics test than student in LG1. This difference was the largest in Belgium, the smallest in the Netherlands.

Table 5.3
Sample Sizes, Means and Standard Deviations Under Model 1 According to Language Group, and Group Differences in Means

Country	LG1			LG2			Difference in means
	N	$\mu(\theta)$	$\sigma(\theta)$	N	$\mu(\theta)$	$\sigma(\theta)$	$\mu(\theta)_{LG1} - \mu(\theta)_{LG2}$
Denmark	3166	0.082	1.097	424	-0.348	1.200	0.430
Germany	2668	-0.216	1.012	659	-0.652	1.011	0.436
Netherlands	3560	-0.212	0.847	787	-0.539	0.933	0.327
Belgium*	4111	0.000	1.000	1204	-0.517	0.934	0.517

Note. LG1 refers to students that indicated to speak the test language at home "always" or "almost always". LG2 refers to students that indicated to speak the test language at home "sometimes" or "never". *The mean $\mu(\theta)$ and standard deviation $\sigma(\theta)$ of the latent scale are fixed for students in LG1 in Belgium.

Model 2.

To further investigate where this difference may originate from, Model 2, a bi-factor version of the GPCM, was fitted to the response data with separate groups specified for every language group within every country. Table 5.4 shows the means and variances of the main component in this model as well as the differences in means between both groups. Results showed that the achievement gap between both student groups (in Model 1) was reduced by incorporating a noise component (Model 2). Comparing the variance of the means for students in Language Group 2 on the main component in Model 1 and Model 2, shows that the variances were also reduced by inclusion of a secondary component. That is, the secondary component explained part of the total variance in proficiency in Model 1.

Table 5.4
Country Means and Standard Deviations of the Main Component in a Bi-Factor GPCM (Model 2) According to Language Group, and Group Differences in Means

Country	LG1			LG2			Difference in means
	N	$\mu(\theta_0)$	$\sigma(\theta_0)$	N	$\mu(\theta_0)$	$\sigma(\theta_0)$	$\mu(\theta_0)_{LG1} - \mu(\theta_0)_{LG2}$
Denmark	3166	-0.441	0.742	424	-0.598	0.759	0.157
Germany	2668	-0.826	0.715	659	-1.045	0.714	0.219
Netherlands	3560	-0.647	0.701	787	-0.863	0.818	0.216
Belgium*	4111	0.000	1.000	1204	-0.456	0.733	0.456

Note. LG1 refers to students that indicated to speak the test language at home "always" or "almost always". LG2 refers to students that indicated to speak the test language at home "sometimes" or "never". *The mean $\mu(\theta_0)$ and standard deviation $\sigma(\theta_0)$ of the latent scale are fixed for students in LG1 in Belgium.

The factor loadings on the secondary component ($\alpha_{ig(n)}$) indicate how strong items measure other constructs in addition to the main component (mathematical

proficiency). Low factor loadings for items with low reading demand and high loadings for items with high reading demand provide support for interpreting this second dimension as a language component. Table 5.5 further summarizes this distribution of the mean factor loadings on the second factor according to the item classification for reading demand. Results suggest that the mean loading increases with the level of reading demand, so that items high in reading demand do load heavier on the secondary component, though these differences are small. From the ten items with the highest loadings (averaged over the eight student groups), six were classified as high, two as medium, and two as low in reading demand. From the ten items with the smallest *a*-values, only three were coded as low in reading demand, five as medium and two as high in reading demand.

Table 5.5
Mean Factor Loadings on the Second Dimension of the Bi-Factor Model According to Item Reading Demand Classification

Level of reading demand	*N*	$\bar{\alpha}_{ig}$	*SD*	Minimum	Maximum
Low	54	0.943	0.266	0.376	1.763
Medium	62	1.005	0.327	0.444	2.159
High	53	1.142	0.434	0.508	2.838

Model 3.

Two more models were fitted to see if the modelling of student responses for students in LG2 can be further improved by taking the reading demand of the items into account. In these models (Model 3 and Model 4), the effects of medium and high reading demand were indicated by γ_1 and γ_2, respectively. Table 5.6 shows the γ estimates for the model fitted to the four countries. For both parameters the 95% credibility intervals were above zero, indicating a significant effect of an item being coded as medium or as high in reading demand. Also, the effect of high reading demand items (i.e. γ_2) was slightly larger than the effect for medium reading demand items (i.e. γ_1), indicating that items high in reading demand result in more difficult items for LG2 students than do items with medium reading demand.

Table 5.6
Parameter Estimates and Credibility Intervals for the Regression Coefficients of Item Reading Demand Classification in the Hybrid Model (Model 3)

Parameter	Mean	*SD*	Lower bound 95% CI	Upper bound 95% CI
γ_1	0.195	0.019	0.158	0.234
γ_2	0.235	0.020	0.194	0.274

Model 4.

Finally, the combined full model (Model 4) was fitted to the data with countries only specified as groups for the bi-factor structure. Table 5.7 shows the means of the latent proficiency scores on the main component. The ordering of countries according to the main component, including modelling of potential DIF, was identical to the ordering in the international study report of TIMSS-2015 (Mullis, Martin, Foy, et al., 2016). The resulting γ estimates, shown in Table 5.8, were comparable to the estimates for Model 3. The factor loadings on the group-specific component averaged over countries showed the smallest mean loading for items classified as low in reading demand and the highest for item high in reading demand. This result corresponds to results for Model 2. Again, the differences were only minimal with the means ranging from 0.882 for items with low to 1.020 for items with high reading demand.

Table 5.7
Country Means and Standard Deviations on the Main Component in a Bi-Factor GPCM with Reading Demand as a Covariate for Students in LG2 (Model 4) and Achievement on the International TIMSS-Scale*

Country	$\mu(\theta_0)$	$\sigma(\theta_0)$	Average Scale Score (*SD*)
Denmark	-0.059	0.774	539 (2.7)
Germany	-0.888	0.692	522 (2.0)
Netherlands	-0.694	0.798	530 (1.7)
Belgium**	0.000	1.000	546 (2.1)

Note. LG2 refers to students that indicated to speak the test language at home "sometimes" or "never". *Mullis, Martin, Foy, et al. (2016). **The mean $\mu(\theta_0)$ and standard deviation $\sigma(\theta_0)$ of the latent scale are fixed for students in Belgium.

Table 5.8
Parameter Estimates and Credibility Intervals for the Regression Coefficients of Item Reading Demand Classification in Model 4

Parameter	Mean	*SD*	Lower bound 95% CI	Upper bound 95% CI
γ_1	0.190	0.021	0.147	0.232
γ_2	0.213	0.022	0.169	0.256

Model Comparison.

A comparison of model fit for the four models was done based on the DIC. Table 5.9 shows the DIC value for each model together with the number of structural item parameters as an indication of model complexity. Model 4 showed the lowest DIC, which suggests that the full model showed the best fit to the TIMSS-2015 mathematics test including information on item reading demand and student language background.

Table 5.9
Model Comparisons Based on Deviance Information Criterion

Model	DIC	Number of structural item parameters
I – GPCM	420200	361
II – bi-factor GPCM	402100	1714
III – Hybrid model	420300	355
IV – Full model	401900	1031

Note. The number of structural parameters includes the item parameters and estimates of the population distributions. Estimates for student proficiency are excluded from the count. Model 1 and 2 both specified separate proficiency distributions for each language group per country, whereas Model 3 and 4 were fitted with proficiency distributions only specified per country, resulting in a lower number of structural parameters.

5.4 Discussion

In this paper we presented an IRT model to model latent proficiency, taking into account potential group differences in the item functioning due to item characteristics and a component additional to the principal component of interest, say the noise or nuisance component. The model can serve to investigate and model DIF in a large-scale assessment. The full model and more restricted, simpler models were fitted to data from the TIMSS-2015 mathematics test in four west-European countries for which an achievement gap was reported between students with different language background (Mullis, Martin, Foy, et al., 2016). This achievement gap was confirmed by the GPCM (Model 1), the baseline model, as student not speaking the test language at home had a lower score than students that do speak the test language at home in each of the four countries.

The full model we presented is characterized by two main features. The first is the bi-factor structure which allows for an item to load on a group-specific nuisance component. A specific hypothesis can be postulated on how this secondary component might be interpreted. In the presented application it was hypothesized that it was related to a language component which items would pertain to in addition to the main component, that is, student's math proficiency. Results showed that modelling

the response data according to the bi-factor model resulted in a smaller gap on the main component between the language groups in all four countries, suggesting the initially found differences in scores may partly be due to a nuisance factor.

If items are classified according to a certain characteristic that is hypothesized to loading on a secondary component, as in the application items were classified according to reading demand, the factor loadings can help evaluate the a priori hypotheses. However, as this secondary component can be regarded as a nuisance component, no interpretation for the component is required. The model is therefore also suitable to investigate and model DIF in a more exploratory fashion. In the application the results were indeed suggestive of items heavier in reading demand loading stronger on the noise component. However, there were also item loadings that did not fit the hypothesized pattern.

This may be because TIMSS items are already formulated very concisely and the difference between items low and high in reading demand are limited. Great differences in effects for the classification are therefore not expected. In addition to that, a general deficiency of indices for reading complexity in mathematics is that they do not measure text cohesion. As Walker, Zang, & Surber (2008, p. 177-178) explain: "If a mathematics test item consists of two short consecutive sentences that are comprised of simple words, then it will be graded as relatively easy to read, even if the two sentences are completely unrelated to each other. Unfortunately, it is precisely this lack of cohesion that may lead to difficulty comprehending the text presented in the problem". Another possibility is that certain items may require low reading demand in the item stem, but do require students to come up with a highly verbal answer – and vice versa. Taking the response format also into account could therefore provide valuable additional information.

The second feature of the model is the interaction effect on the item difficulty between student and item characteristics. This offers the opportunity to evaluate hypotheses regarding specific item characteristics that might lead to DIF as a result of increased item difficulty for specific subgroups. In the TIMSS-2015 mathematics data for the countries studied, the level of item reading demand showed to increase item difficulty for students not speaking the test language at home. This confirms the need for great care in test construction of contextualized items to reduce the reading

demand. In addition to that, it advocates taking reading demand into consideration when modelling mathematical proficiency based on contextualized items for students with different language backgrounds.

Both features of the model can be applied separately, as was exemplified here by the specification of Model 2 and Model 3, but they can also be combined to an overall model (Model 4). In the application, this full model showed the best fit for the TIMSS-2015 mathematics data in the four west-European countries. Modelling the response data according to this full model did not result in a different country ranking among the four countries than their respective ranking on the international TIMSS-2015 scale. Though only four countries were studied here, the results do not suggest that the TIMSS results on a national level are affected by DIF due to item reading demand.

Though the results of this study show support for the differentiating role of reading demand in mathematics, some findings in this study do ask for further research. Items showed a clear relation between cognitive and content domains and classification according reading demand. Covering certain areas in the math curriculum (e.g. data display) will require more contextualized items than others (e.g. numbers). The same goes for the cognitive domains. For example, items in the domain "reasoning" encompass unfamiliar situations, complex contexts, and multi-step problems (Mullis, & Martin, 2013), which explains their higher level of reading demand. Ideally, the effects of reading demand would therefore be assessed within a content or cognitive domain.

Also, a valuable next step would be to validate the model with the combined data from TIMSS-2011 and PIRLS-2011 by comparing not only the students with different language backgrounds, but also based on their PIRLS reading achievement: a score for which language spoken at home can only serve as an approximation. This future research could shed more light on whether the achievement gap comes from a disadvantage due to language factors for SLL students or whether other factors such as SES might play a bigger role.

Chapter 6
The Role of Reading Proficiency in Testing Mathematics Achievement of Second Language Learners

Abstract

The role of language proficiency in testing proficiency in mathematics is widely acknowledged, particularly for contextualized items. Concerns are raised that this dependency may lead to biased findings in math achievement, especially for second language learner (SLL) students. This study sets out to evaluate to what extent this potential bias can be attributed to limited reading proficiency. Using the data from the combined TIMSS-2011 and PIRLS-2011 administration in four European countries, this research focuses on modelling mathematical proficiency, taking into account reading proficiency. In addition, the effect of student language background is incorporated to investigate if being an SLL student affects achievement in mathematics beyond effects of reading proficiency. The final modelling step incorporates effects of item reading demand classifications for SLL students. The response and structural parts of the models were estimated concurrently, using Bayesian estimation procedures.

Results in this study confirmed the often-reported relationship between student achievement in reading and mathematics. In addition, it gave way to test the extent to which the often-observed differential item functioning in the mathematics test for SLL students can be attributed to their more limited reading achievement. Results showed a small but negative effect of being an SLL student, even after controlling for reading proficiency. This shows that the achievement gap between SLL and non-SLL students cannot be fully explained by their level of reading literacy. Also, a small interaction effect was found between item reading demand and student language background, while controlling for reading proficiency. Implications of these findings for the development of future international large-scale assessments relate to taking potential negative effects of item wording into account. This should be done at the phase of test construction, as well as in analyses, once all the data is collected.

6.1 Introduction

An important goal of an international large-scale assessment (ILSA) is to measure student proficiency in a valid and unbiased way. That is, to have differences in test scores reflect true differences in the targeted proficiency. In assessments of subjects other than language, language skills are nevertheless needed to understand and answer the items, which may lead to biased results. In this study, the focus is on how reading literacy interacts with assessing mathematics and whether the often-found achievement gap between students with different language backgrounds can be attributed to differences in reading literacy.

6.1.1 Validity in Large-Scale Assessments

For a test to be valid, it must be free of construct-irrelevant variance (Messick, 1989). Construct-irrelevant variance emerges if a test requires certain skills that are not targeted by the test, for example due to language demand in mathematics testing. In mathematics assessments, scores are expected to indicate students' proficiency in this area and validity of the test scores depends on the extent to which performance on assessments are indeed accurate indicators of students' competencies (Kane, 2013). Many achievement tests in ILSA apply contextualized items to test the targeted proficiency in a meaningful, real-life setting. Although having students applying their skills in a variety of contexts can limits the risk of construct-underrepresentation, it may also evoke construct-irrelevant variance by, for example, the reading demand posed by contextualized items. This can lead to test bias as students with low reading proficiency may score lower on the mathematics test than equally proficient students in mathematics but with higher reading proficiency.

The differences in item response behaviour for students of equal proficiency in the subject targeted by the test are often labelled as differential item functioning (DIF) and these differences can complicate inferences regarding proficiency differences between countries or specific subpopulations. DIF can be viewed as related to differences in the multidimensional proficiency distributions of distinguished groups. Though scores are often modelled by a unidimensional model, the student's score may actually represent a composite of abilities (Ackerman, 1992). When groups of students differ in these latent abilities, but only a single score is reported, DIF can occur (Roussos & Stout, 1996).

6.1.2 Reading and Mathematics

The ILSA Trends in International Mathematics and Science Study (TIMSS) measures the achievement of fourth- and eighth-grade students in science and mathematics, and has its framework organized around content and cognitive dimensions. The mathematics test relies strongly on the use of contextualized items to cover the broad domain of the curriculum to be assessed. To ensure that the formulations of the items are suited for the student population, the word problems are worded concisely and succinctly (Mullis & Martin, 2013). The items nevertheless require sufficient reading skills to handle the reading demand.

The issue of reading demand in word problems was addressed using TIMSS data when in 2011 the main data collection for TIMSS and the Progress in Reading Literacy Study (PIRLS) coincided (Martin & Mullis, 2013). PIRLS collects data to provide information on trends in reading literacy achievement of fourth-grade students. In 34 countries the same students participated in both assessments, providing the opportunity to further research the relationship between reading achievement and achievement in science and mathematics. To study the effects of item reading demand in particular, the TIMSS test items were classified as low, medium or high in reading demand. Martin and Mullis (2013) reported an effect of item reading demand on the performance on the TIMSS test for students with lower reading proficiency, but not for students with a high reading proficiency.

6.1.3 Second Language Learners

The requirement of a certain level of reading proficiency to do well on a mathematics test may particularly affect students with different language backgrounds such as Second Language Learner (SLL) students. SLL students are described by Barwell (2012, p.147) as "students from linguistic minority backgrounds whose home languages are not well-represented or recognized in wider society". Several studies focused on differences in achievement between SLL and non-SLL students and found consistently lower achievement for SLL students as opposed to non-SLL students. A study by Andon, Thompson, and Becker (2012) focused on students in large-scale assessment studies and reported underperformance of SLL students in both the Programme for International Student Assessment (PISA) and PIRLS. Furthermore, TIMSS-2015 showed that in most of the participating countries, in particular European countries, students reporting to speak the language of the test at home only

sometimes, tended to score lower than students always speaking the test language at home (Mullis, Martin, Foy, & Hooper, 2016).

Other studies focused specifically on DIF in mathematics items for students with different language backgrounds. Banks, Jeddeeni, and Walker (2016) compared the probabilities of answering mathematical word problems correctly between SLL students and non-SLL students of equal math proficiency. They found a significantly lower average total score for SLL students, yet no DIF against SLL students. Abedi and Lord (2010) reduced the linguistic complexity of items from a large-scale national assessment and found that eighth-grade SLL students and low SES students benefited most from these linguistic modifications. The results suggested an interaction effect between language demand and student's linguistic and socioeconomic background (Abedi & Lord, 2010). In a fourth-grade US state assessment in mathematics it was found that test items with greater item linguistic complexity where more difficult for SLL students than for non-SLL students (Martiniello, 2009). In addition to that, SLL students' think-aloud interviews while solving the DIF items identified linguistic complexity as a source of the found DIF for SLL students (Martiniello, 2009). Ercikan et al. (2015) examined the comparability of PISA-2009 mathematics and science scores for students with an English language background and non-English language background in four English speaking countries. They found indications of differential score meaning and comparability. More generally, they reported that up to 43% of the variance in mathematics, and up to 79% in science, was accounted for by reading proficiency.

In Chapter 5 of this thesis, the classification scheme for item reading demand from Martin and Mullis (2013) was applied to the TIMSS-2015 mathematics items. These classifications were incorporated in an item response theory (IRT) model to test for potential differential effects of item reading demand for students not speaking the test language at home. Results showed that this effect was indeed present. The study also showed that the mathematics test addressed a secondary component, which might relate to reading proficiency. Unfortunately, no test data on students' reading proficiency was available to further explore the role of reading proficiency in testing mathematics.

6.1.4 Research objectives

This study sets out to address the role of student reading literacy and item reading demand in the mathematics test in four European countries. In addition to that, the study aims to investigate whether this relationship differs across countries and whether being an SLL student affects mathematical achievement beyond effects of reading literacy. The following research questions are formulated:

1. To what extent is the estimated reading literacy score related to the mathematics score and does this effect differ across countries?
2. What is the effect of language background on the mathematics score, when controlling for reading achievement and does this effect differ across countries?
3. Does item reading demand interact with language background on item difficulty, when controlling for student reading literacy and test language spoken at home?

To answer these questions, the response models on the TIMSS-2011 mathematics and PIRLS-2011 reading literacy test are estimated, along with several latent regression structures on proficiency in mathematics, using Bayesian estimation.

6.2 Methods

6.2.1 Data

Response data was used from the TIMSS-2011 mathematics test (Mullis, Martin, Foy, & Arora, 2012), which consists of 175 items. For the reading literacy, data was used from the PIRLS-2011 reading literacy test (Mullis, Martin, Foy, & Drucker, 2012), consisting of 10 reading passages with a total of 135 questions related to these passages. For both TIMSS and PIRLS a booklet rotation system was in place, with test items distributed across 14 and 13 booklets, respectively. For each test, a sampled student was assigned one of the booklets.

Data was used from Finland, Germany, Ireland and Sweden. These four west-European countries participated in the combined administration of TIMSS and PIRLS in 2011, resulting in data on both tests for the same participating students. Data on both tests was available from 16786 students in total.

In addition to the response data on the mathematics and reading literacy test items, data from the question "*How often do you speak <<language of the test>> at home?*" in the student questionnaire was used. Based on this question, two student subpopulations

were defined: students that indicated to speak the test language at home "always" or "almost always" (the non-SLL group) and students in that indicated to speak the test language "sometimes" or "never" at home (the SLL group). Table 6.1 shows the total number of students per country and the number of the SLL and non-SLL students within each country.

At the item level, the classification according to the level of reading demand available from Martin and Mullis (2013) was used. This classification states for each item in the mathematics whether it is low, medium or high in reading demand.

Table 6.1
Sample Sizes

Country	N_{Total}	$N_{SLL\ student}$	$N_{Non\text{-}SLL\ student}$
Finland	4525	4028	497
Germany	3544	2825	719
Ireland	4303	3554	749
Sweden	4414	3353	1061

Note. SLL refers to students that indicated to speak the test language at home "sometimes" or "never".

6.2.2 Statistical Analyses

Eight nested models were defined to answer the research questions.

Model 1 and 2.

The response model on the mathematic test items (Model 1) is based on the generalized partial credit model (GPCM; Muraki, 1992). The GPCM, which is commonly used in ILSA for items requiring partial credit scoring, assumes one latent variable is needed to explain response behaviour. Furthermore, an item discrimination parameter a_i and response category parameters b_{im}, provide information on the salience of the item scores. The GPCM is a generalization of the partial credit model (Masters, 1982) which assumes the discrimination parameters for all items are constrained to be equal.

In the response model for the TIMSS mathematics items for student n from country g, with latent math proficiency $\theta^{(T)}_{ng(n)}$, the probability of attaining a score Y_{ni} in response category $m = 0, 1, \ldots, M_i$ on item i is given by

$$\mathrm{P}(Y_{ni} = m \mid \theta^{(T)}_{ng(n)}) = \frac{\exp\left(m a_i \theta^{(T)}_{ng(n)} - b_{im}\right)}{1 + \sum_{h=1}^{M_i} \exp\left(h a_i \theta^{(T)}_{ng(n)} - b_{ih}\right)}, \quad (6.1)$$

with $\theta^{(T)}_{ng(n)} \sim N(\mu_g, \sigma^2_g)$, where the upper index (T) indicates that this latent proficiency relates to the TIMSS test. To identify the latent scale, the latent distribution for one of the countries is set to standard normal (i.e. $\mu_1 = 0,$ and $\sigma^2_1 = 1$). Note that the maximum item score *M* can differ among items, hence the subscript *i*.

The response model on the PIRLS reading literacy items (Model 2) is also based on the GPCM, but differs from Eq. (6.1) as it also incorporates a testlet effect to model the dependency of responses to items nested in a reading passage (Wainer, Bradlow, & Wang, 2007). The PIRLS reading achievement test consist of several passages, with multiple items relating to the same passage (Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009). These items form a group which is called a testlet. To take the correlation among these items into account, the response model on the PIRLS-2011 items includes a testlet effect γ for each respondent. This γ parameter has index *d(i)* indicating to which testlet item *i* belongs. The probability of attaining a score U_{ni} in response category *m* is then given by

$$\mathrm{P}(U_{ni} = m \mid \theta^{(P)}_{ng(n)}) = \frac{\exp\left(mc_i\theta^{(P)}_{ng(n)} + \gamma_{nd(i)} - d_{im}\right)}{1 + \sum_{h=1}^{M_i} \exp\left(hc_i\theta^{(P)}_{ng(n)} + \gamma_{nd(i)} - d_{ih}\right)}, \qquad (6.2)$$

with $\theta^{(P)}_{ng(n)} \sim N(\mu_g, \sigma^2_g)$, where the population distributions for one of the countries is again set to standard normal and $\gamma_{nd(i)} \sim N(0, \sigma^2_{\gamma d})$. The upper index (P) indicates that the latent proficiency relates to the PIRLS test. The parameters c_i and d_{im} refer to the item's discrimination parameter and the response category parameters, respectively.

Models 3 to 8.

To answer the research questions, six latent regression model are estimated. Each latent regression model is estimated concurrently with the response models as described in Eq. (6.1) and Eq. (6.2). We will discuss each model in turn.

The first latent regression models incorporate the estimated reading proficiency as a predictor for math achievement to answer research question one. In Model 3 the effect is estimated across all four countries. In Model 4 the country-specific effects are estimated. The regression structure for Model 3 can be written as

$$\theta^{(T)}_{ng(n)} = \beta_{0g(n)} + \beta_1 \theta^{(P)}_{ng(n)} + \varepsilon_{ng(n)}, \text{ with } \varepsilon_{ng(n)} \sim N(0, \sigma^2_\varepsilon), \tag{6.3}$$

with $\beta_{0g(n)}$ indicating a country-specific intercept, which is fixed for one of the countries at zero to further identify the model, and β_1 the effect of reading literacy achievement across countries. Model 4 is defined analogous to the model in Eq. (6.3) only with a country-specific β_{1g} indicating the country-specific effect of reading proficiency.

Models 5 and 6 address research question two. Here, reading literacy and student language group are included as predictors for math achievement, with effects of both predictors (β_1 and β_2) assumed either equal (Model 5) or unique (Model 6) for each country. Whether a student is an SLL is indicated by

$$z_n = \begin{cases} 1 & \text{if student } n \text{ is a SLL student} \\ 0 & \text{otherwise.} \end{cases}$$

For Model 5 the regression structure is defined as

$$\theta^{(T)}_{ng(n)} = \beta_{0g(n)} + \beta_1 \theta^{(P)}_{ng(n)} + \beta_2 z_n + \varepsilon_{ng(n)}, \text{ with } \varepsilon_{ng(n)} \sim N(0, \sigma^2_\varepsilon). \tag{6.4}$$

This is also the definition for Model 6, only then the effects are given country-specific parameters β_{1g} and β_{2g} for each country g.

In the final models, the previous models are extended to include an interaction between reading demand in the test and the language background of the student to answer the third research question. Effects of item reading demand are incorporated in the response model of the TIMSS data as an effect on the latent mathematics dimension $\theta^{(T)}_{ng(n)}$. To study the potential differential effects of item reading demand for SLL and non-SLL students, an interaction term is added. The response model on TIMSS then becomes

$$\mathrm{P}(Y_{ni} = m \mid \theta^{(T)}_n) = \frac{\exp\left(m\left(a_i \theta^{(T)}_n - \sum_q \delta_q z_n x_{iq} \right) - b_{im} \right)}{1 + \sum_{h=1}^{M_i} \exp\left(h\left(a_i \theta^{(T)}_n - \sum_q \delta_q z_n x_{iq} \right) - b_{ih} \right)},$$

with $x_{iq} = \begin{cases} 1 & \text{if item } i \text{ has Language Demand } q \\ 0 & \text{otherwise,} \end{cases}$

where q = 1 for medium, and q = 2 for high reading demand. In Model 7 this response model is combined with a latent regression structure that includes a cross-country effect for reading literacy as shown in Eq. (6.3). In Model 8 the response model is combined with the latent regression structure including both reading proficiency and language background, as shown in Eq. (6.4).

6.2.3 Estimation

All models were estimated in a Bayesian framework using the OpenBUGS software (Lunn, Spiegelhalter, Thomas, & Best, 2009), which is an open-source package that provides a general-purpose Markov chain Monte Carlo (MCMC) sampler to estimate complex IRT models in a Bayesian framework. Models were compared on their fit with the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van der Linde, 2002) that is provided by the software.

The script, including prior specifications, for estimation of the most extensive model (Model 8) is provided in Appendix C. The other models can be inferred from this script by removing code related to parameters not presented in the nested models. From this script it becomes clear that the booklet structure of both TIMSS and PIRLS is taken into account. The MCMC sampling procedure in the application were run with over 11000 iterations for each chain. Each chain had a burn-in of 4000 iterations. The sampling procedures were checked for convergence by visual inspection of the trace plots. As an additional check for convergence and the influence of the priors, the response models as shown in Eq. (6.1) and Eq. (6.2) were also estimated in the framework of marginal maximum likelihood (MML; see for example, Bock & Aitkin, 1981) using the software-package LEXTER (Glas, 2017).

6.3 Results

6.3.1 Model 1 and 2

To obtain an impression of the available data, the first analysis step was to estimate math achievement (Model 1) and reading literacy (Model 2) separately, both with the MML and the fully Bayesian approach. The correlations between the parameter estimates obtained by the two methods was always above 0.99. This shows that the

priors for the parameters had little influence and the convergence of the MCMC procedure was appropriate for the response models.

Table 6.2 gives the Bayesian estimates of the parameters of the proficiency distributions. In this table the reported country scores on the international TIMSS and PIRLS scales are also shown. As would be expected, the country ranking is the same for the mathematics test across both scales. The same ranking would also be expected for country performance on reading literacy, but here we see a slight deviation: the ranks are switched for Germany and Sweden. However, these countries did not show a significant difference on the international scale (Mullis, Martin, Foy & Drucker, 2012).

Table 6.2
Population Means on the TIMSS Mathematics and PIRLS Reading Literacy Assessment, International Scale Means, Latent Scale Means and Credibility Regions

		Int. score (SD)	μ_g			σ_g^2		
			Mean	Lower bound 95% CI	Upper bound 95% CI	Mean	Lower bound 95% CI	Upper bound 95% CI
TIMSS								
	Finland	545 (2.3)	0.000*			1.000*		
	Germany	528 (2.2)	-0.089	-0.131	-0.038	0.863	0.792	0.942
	Ireland	527 (2.6)	-0.266	-0.310	-0.214	1.248	1.155	1.351
	Sweden	504 (2.0)	-0.535	-0.531	-0.486	0.821	0.758	0.890
PIRLS								
	Finland	568 (1.9)	0.000*			1.000*		
	Germany	541 (2.2)	-0.273	-0.343	-0.234	1.041	0.967	1.134
	Ireland	552 (2.3)	-0.177	-0.248	-0.135	1.532	1.435	1.655
	Sweden	542 (2.1)	-0.376	-0.443	-0.340	1.089	1.019	1.177

Note. *Means and variances for Finland are fixed to anchor the scale. Posterior means and 95% credibility regions for the population means and variances. Internationally reported scores on the TIMSS mathematics (Mullis, Martin, Foy, & Arora, 2012) and the PIRLS reading literacy assessment (Mullis, Martin, Foy, & Drucker, 2012)

Then, to get an impression of the correlation between both estimated proficiencies, the IRT models for the TIMSS and PIRLS data were estimated concurrently with the assumption that the proficiency parameters within every country had a bivariate normal distribution. The correlations are given in Table 6.3. Results show high correlations: all are above .70.

Table 6.3
Estimates of Correlation Between both Latent Dimensions for each Country

	Mean	*SD*	Lower bound 95% CI	Upper bound 95% CI
Finland	0.736	0.008	0.720	0.795
Germany	0.812	0.004	0.811	0.823
Ireland	0.816	0.005	0.802	0.813
Sweden	0.787	0.005	0.766	0.789

6.3.2 Models 3 to 8

The first latent regression model (Model 3) includes reading literacy as a predictor for math achievement, assuming the effect is equal for the four countries. The intercept β_0 represents the baseline scores for countries on the mathematics test and the regression coefficient β_1 represents the effect of reading literacy on math score. Model 4 uses the same input variables, but here the effect of reading literacy is allowed to differ across countries. Table 6.4 shows the results for both models which together answer the first research question. The table shows that the credibility interval for the overall effect of reading literacy on math achievement lies above zero, indicating a positive effect of reading achievement. Results for Model 4 show that the found effect of reading proficiency at country level is the smallest in Sweden. However, the means hardly differ between countries.

Table 6.4
Parameter Estimates and Credibility Intervals for the Regression Coefficients and Variance Components in Model 3 and Model 4

		Model 3				Model 4			
Parameter	Country	Mean	SD	Lower bound 95% CI	Upper bound 95% CI	Mean	SD	Lower bound 95% CI	Upper bound 95% CI
Intercept β_0	Finland	0.000*				0.000*			
	Germany	0.100	0.015	0.072	0.129	0.094	0.014	0.068	0.124
	Ireland	-0.073	0.013	-0.098	-0.048	-0.066	0.013	-0.091	-0.042
	Sweden	-0.159	0.013	-0.184	-0.134	-0.169	0.014	-0.196	-0.142
Reading Literacy β_1		0.507	0.016	0.480	0.506				
	Finland					0.492	0.022	0.455	0.544
	Germany					0.485	0.023	0.444	0.535
	Ireland					0.502	0.022	0.465	0.552
	Sweden					0.453	0.022	0.416	0.501
σ_e^2		0.169	0.011	0.145	0.189	0.157	0.013	0.138	0.191

Note. CI indicates the credibility interval. *The latent ability distribution for one of the countries is fixed to identify the latent scale.

To answer the second research question, Model 5 and 6 include language background as a predictor variable for math achievement, in addition to reading proficiency. In

Model 5, these effects are estimated as fixed across countries, whereas in Model 6 the effects are allowed to vary over countries. Table 6.5 shows the estimates for both models, where β_1 pertains to the effect of reading literacy and β_2 to the effect of student language group. The overall effect of language background is negative. This means that for SLL students, after controlling for their demonstrated reading proficiency, there still is a negative effect on their estimated math proficiency. However, it should be noted that this effect is very small. Looking at the variance component at the level of the student, the variance reduces from 0.169 in Model 3 to 0.158, so the language background explains an additional 7%.

Table 6.5
Parameter Estimates and Credibility Intervals for the Regression Coefficients and Variance Components in Model 5 and Model 6

Parameter	Country	Model 5				Model 6			
		Mean	SD	Lower bound 95% CI	Upper bound 95% CI	Mean	SD	Lower bound 95% CI	Upper bound 95% CI
Intercept β_0	Finland	0.000*				0.000*			
	Germany	0.104	0.014	0.074	0.131	0.112	0.016	0.082	0.143
	Ireland	-0.070	0.015	-0.101	-0.042	-0.070	0.014	-0.095	-0.040
	Sweden	-0.151	0.017	-0.187	-0.120	-0.146	0.013	-0.171	-0.118
Reading Literacy β_1		0.487	0.020	0.444	0.521				
	Finland					0.464	0.015	0.439	0.495
	Germany					0.445	0.016	0.413	0.476
	Ireland					0.478	0.014	0.452	0.508
	Sweden					0.422	0.015	0.394	0.451
Language Background β_2		-0.056	0.012	-0.080	-0.032				
	Finland					-0.054	0.025	-0.101	-0.002
	Germany					-0.140	0.023	-0.186	-0.095
	Ireland					0.045	0.022	0.000	0.089
	Sweden					-0.071	0.018	-0.107	-0.037
σ_e^2		0.158	0.012	0.132	0.180	0.137	0.006	0.125	0.149

Note. Language Background refers to whether or not student is an SLL student. CI indicates the credibility interval.
*The latent ability distribution for one of the countries is fixed to identify the latent scale.

From the estimates for Model 6 it can be seen that the effect of being an SLL student is largest in Germany. The credibility interval for β_2 in Ireland does not give reason to assume the effect differs from zero, which means there does not appear an effect of langue group after controlling for reading literacy. The inclusion of effects at country

level as opposed to overall effects, results in a reduction to 14% error variance at the student level.

In the last models, Model 7 and 8, reading demand classification at the item level is also incorporated in the response model on the TIMSS items. Estimates for these models (see Table 6.6) help answer the third research question, pertaining to the effect of item reading demand on the mathematics score for SLL students. The relevant parameters in these models are δ_1 and δ_2, which refer to the effects of medium and high reading demand, respectively.

Table 6.6
Parameter Estimates and Credibility Intervals for Effects of Item Reading Demand and Reading Literacy and Language Group on Math Achievement

Parameter	Country	Model 7				Model 8			
		Mean	SD	Lower bound 95% CI	Upper bound 95% CI	Mean	SD	Lower bound 95% CI	Upper bound 95% CI
Intercept β_0	Finland	0.000*				0.000*			
	Germany	0.104	0.015	0.075	0.135	0.105	0.014	0.078	0.131
	Ireland	-0.071	0.013	-0.096	-0.046	-0.063	0.013	-0.091	-0.038
	Sweden	-0.156	0.014	-0.185	-0.130	-0.144	0.014	-0.174	-0.118
Reading Literacy β_1		0.508	0.025	0.469	0.574	0.486	0.016	0.455	0.515
Language Background β_2						-0.042	0.015	-0.074	-0.013
Reading Demand δ_{medium}		0.066	0.019	0.029	0.107	0.033	0.023	-0.013	0.077
Reading Demand δ_{high}		0.070	0.021	0.028	0.113	0.031	0.024	-0.019	0.076
σ_e^2		0.158	0.012	0.108	0.160	0.158	0.010	0.140	0.176

Note. Language Background refers to whether or not student is as SLL student. Reading Demand refers to the interaction effect between item reading demand classification and student language background. CI indicates the credibility interval. *The latent ability distribution for one of the countries is fixed to identify the latent scale.

In Model 7, small effects were found for item reading demand on item difficulty for SLL students. The effect for items high in reading demand was, as to be expected, slightly larger than the effect for items with medium reading demand. The parameter estimates for δ_1 and δ_2 are smaller than the effects found for the TIMSS-2015 mathematics data (respectively 0.195, 0.235; Chapter 5). When the language group of the student was also included as a predictor in the latent regression model (Model 8),

the interaction effects between item reading demand and student language background proved to be around zero. This indicates equal functioning of TIMSS mathematics items with different levels of reading demand across both SLL and non-SLL students, when controlled for both student's language background and reading proficiency.

6.3.3 Model comparison

The models were compared based on the DIC (Table 6.7). The smallest DIC was found for Model 5, which incorporates both student reading literacy as well as student language background as predictors for math achievement. However, model comparisons based on the DIC did not show substantive differences between the models. This can be explained by the large number of item parameters in each of the models, required for the response models. The fact that the less restrictive models, where the effects are allowed to vary over countries, show a model fit similar to their more restrictive counterparts supports the notion that there are limited differences in the effects between countries.

Table 6.7
Model Comparisons Based on Deviance Information Criterion

Model	DIC
3	932400.0
4	931900.0
5	930900.0
6	931100.0
7	932400.0
8	931000.0

6.4 Discussion

The main purpose of this study was to explore the relationship between student reading and mathematical achievement using the TIMSS-2011 and PIRLS-2011 combined data from four European countries. Another point of interest was to see to what extent reading literacy can explain the commonly observed achievement gap between SLL and non-SLL students. Three research questions guided the study, each of which will be addressed in turn.

1. *To what extent is the estimated reading literacy score related to the mathematics score and does this effect differ across countries?*

The correlation between both latent constructs appeared very high and positive across all four countries, which corresponds to the often-reported relationship between both skills. In the regression models, this result echoed in finding a substantial effect of the

reading literacy achievement across countries. Effects did not differ much across the countries studied.

In interpreting these effects, we have to take into account that the PIRLS literacy test does not only invoke a student's reading skills. In absence of other information on student's educational attainment, the test also serves as a measure of a student's general educational proficiency. This underlying general proficiency contributes to both achievement in mathematics and reading. These findings are in line with a study by Lynn and Mikk (2009) on data from PISA-2006, which showed high correlations (>.84) between both national scores in reading comprehension and mathematical ability with national IQ measures. Also, the TIMSS and PIRLS test administrations are very similar in nature: they both appeal to working well on a paper-and-pencil test during a standardized classroom test administration. The likeness of the testing situation is likely to result in higher correlations as opposed to correlations between measures of different assessment modes.

2. *What is the effect of language background on the mathematics score, when controlling for reading achievement and does this effect differ across countries?*
Underachievement of SLL students is often hypothesized as due to their limited proficiency in the test language. Results of this study showed that there is a negative effect for SLL students, even when their reading proficiency is taken into account. This raises questions and motivates further research into which factors can explain this additional negative effect. Perhaps SLL students are hindered by cultural notations in wording problems or less familiar contextualization of items. It may also be that students are not necessarily disadvantages by the test, but they are hindered in the learning process. Cultural differences may lead to less effective instruction and learning in mathematics, resulting in lower mathematical proficiency.

In the four countries studied here, the largest effect of language background was found for Germany. Ireland, on the other hand, showed no effect. So, whether being an SLL student hinders achievement in mathematics, when controlling for reading proficiency, differs among countries. Future research could focus on explanatory factors for these country differences. Perhaps the SLL student population differs among countries in characteristics such as social economic status. More insight into

these characteristics, may also provide more insight into challenges faced by these students in their educational trajectories.

3. *Does item reading demand interact with language background on item difficulty, when controlling for student reading literacy and test language spoken at home?*
After establishing main effects of both reading literacy and language background, we tested a model that included an interaction term on an item's reading demand and student background. As in Chapter 5 where this interaction was tested on TIMSS-2015 data, we found a small effect for the interaction. The effect was smaller than in Chapter 5, which was to be expected as in this study we also corrected for student reading literacy. The implications of a small, yet present, effect of the interaction between language background and item reading demand, indicates that for SLL students the items more demanding in reading skills proved to be more challenging than for non-SLL students – even while correcting for reading literacy. Future research could look into explaining factors. For example, whether SLL students are more intimidated by a wordily math question or whether these high reading demand items address less familiar contexts for SLL students.

Regarding the estimation procedures used in this study we found that the Bayesian framework offered ways to test more complex models. The MCMC sampling was implemented by the open-source package OpenBUGS (Lunn et al., 2009). Though the great advantage of not having to worry about the sampling process other than ensuring convergence, it did require considerable computational time. Using dedicated samplers, set up for specific models, will most likely speed up the estimation time. This does however, require more statistical knowledge and programming skills of the researcher.

The current model assumes reading literacy as compensatory in nature to the mathematics achievement. Future research could also consider applying a conjunctive model (also called a non-compensatory model), which is characterized by an interaction term between both latent dimensions. As a result of this interaction, the chance of responding to an item correctly is limited by the lowest latent dimension score. In the case of the mathematics test, the achievement of a poor reader will be limited by his low reading proficiency, regardless of the student's potential excellence

in core mathematics. Also, a limited math proficiency will result in lower achievement, regardless of potentially great reading skills.

Implications from this study for future cycles of TIMSS relate to addressing the potential negative effects of item wording for students with limited reading proficiency. This should be done at the phase of test construction, as well as at the phase of data analysis. As generally no ILSA data is available on student reading proficiency, it will be difficult to fully correct for this in analyses. Chapter 5 offers ways to address this by means of a bi-factor structure where language ability and other unidentified noise variables are estimated as a country-specific secondary component next to the estimation of math ability component. In the framing of math test items, the context should also be kept as familiar as possible to both non-SLL and SLL students to avoid further bias for the latter student group.

Chapter 7
Epilogue

This thesis presented several studies based on data from international large-scale assessment (ILSA) studies. Though the objective of each study differed, they all set out to contribute to substantive knowledge in the field of educational studies and to explore methodologies on identifying and handling potential differential item functioning (DIF) in the framework of item response theory (IRT). In this chapter, reflections are provided on their contributions. Also, perspectives are given on the implications of this thesis for future ILSAs and resulting secondary analyses.

7.1 Contributions to Educational Research

Studies in this thesis intended to contribute to research in the field of education by both deploying ILSA data in research areas where the availability of standardized data from multiple countries offered new research opportunities, and improved modelling of the large datasets that characterise ILSAs.

This improved modelling is reflected in Chapter 2 by multidimensional modelling of the International Computer and Information Literacy Study (ICILS) data. In the international report, a unidimensional scale was reported, showing that girls outperformed boys. The proposed three-dimensional modelling in Chapter 2 offered more insight into the construct of computer- and information literacy (CIL). The study showed that CIL, as measured in ICILS-2013, entails information-oriented dimensions and a dimension oriented towards computer literacy. From studying achievement on the different dimensions, it became clear that the internationally reported gender difference in ICILS originated primarily from girls outperforming boys in the information-oriented dimensions, whereas no gender differences were shown on the computer literacy dimension. This study showed how, as pointed out by Reckase (2009, p. 54), "careful development of multidimensional IRT models should lead to better descriptions of the interactions between persons and test items that can be provided by unidimensional IRT models".

In Chapters 3 and 4, the availability of data from the same instruments across many countries provided a unique contribution to research on the effects of parental involvement and reading achievement. But also in Chapter 3, great attention was directed at more elaborate modelling of the parental involvement construct itself. The study illustrated how the Progress in International Reading and Literacy Study (PIRLS) measures a variety of the different components and perspectives of parental involvement and how they can be estimated within the framework of IRT. In addition, close attention was directed at correcting for potential country-specific effects, to gain more precise estimates of the parental involvement constructs. Chapter 4 showed that early literacy activities and help with homework were both related to reading literacy. The impact of parental involvement on reading literacy was not large, but did show great variation across countries. The different modelling techniques to correct for country-specific item functioning did not lead to different conclusions.

In Chapters 5 and 6, the attention was directed back from modelling the questionnaire data (as in Chapter 3) to modelling test data (as in Chapter 2). Both chapters focus on the role of reading in the Trends in International Mathematics and Science Study (TIMSS) mathematics test, due to the contextualization of test items. In Chapter 5, the results on the TIMSS-2015 data suggested the items address a secondary component, likely related to reading demand, in addition to the main component (mathematics). The findings in Chapter 6 were in line with these results, as the regression models showed a strong relation between the PIRLS-2011 reading literacy test and the TIMSS-2011 mathematics test. The findings contribute to the scientific field of mathematics and language by testing the relation between both proficiencies, measured using standardized tests across many students.

The role of reading demand in mathematics tests has implications for students with limited reading proficiency. Because of high migration across the globe, educational systems are faced with educating children who come from multiple language and cultural backgrounds. This makes the learning process of second language learner (SLL) students highly relevant. By studying the test functioning for these students, more knowledge is gained on whether differences in math test achievement are due to limited math proficiency or that the reading demands of the test hinders students. In Chapters 5 and 6, the availability of background information on the students (i.e., language spoken at home) was used to study the test functioning for this specific

subpopulation. Results from both chapters were consistent, hinting that the item functioning indeed differs for students with different language backgrounds (Chapter 5) and that there was a negative effect on math achievement for SLL students even beyond the effect of reading literacy (Chapter 6).

7.2 Identifying and Handling Differential Item Functioning

As pointed out in Chapter 1, for the primary purpose of ILSA, making valid comparisons between countries is of fundamental importance. But also for secondary analyses, gathering evidence to support the adequacy and appropriateness of inferences is essential (Messick, 1989). Therefore, in addition to contributing to the field of educational research, this thesis set out to focus on measurement invariance across groups of interest. In this, it follows the recommendation already formulated by Lin, Bumgarner, Chatterji (2014, p. 31) to "provide ongoing psychometric and research resources so as to continually address or mitigate various sources of cultural, linguistic or other biases in regional and national ILSA reports".

To address the issue of item bias in comparisons across (sub)populations, this study presents several ways to identify and handle differential item functioning (DIF):

- Comparisons of item parameters (Chapter 2)
- Residual approach (Chapters 2 and 3)
- Random item parameters (Chapter 3)
- Bi-factor approach (Chapters 3 and 5)
- Hybrid model on item and person characteristics (Chapters 5 and 6)

The methodologies demonstrated for handling and identifying DIF were not intended to lead to the ultimate choice for a method. They should rather be regarded as a showcase of suitable methods for ILSA data. That is, they are part of a toolbox to investigate DIF from different perspectives. Each research question requires reflection on how to build a case for validity and which modelling techniques can best serve this goal. Some reflections on appropriate use will be provided next.

The first approach mentioned, comparisons of item parameters, is very intuitive in its interpretation: high correlations between parameter estimates across groups provide great support for measurement invariance. Also, the estimation of the correlations is very straightforward. However, care must be taken that models are specified identically

across groups with the same identification of the latent scale. A difficulty with this approach is that it is hard to say when a correlation is strong enough. In addition, the approach does not identify items showing most DIF.

To more thoroughly test differences in difficulty parameters, a residual approach can be adopted. This approach calculates residuals as the differences between observed mean item scores in subgroups and their expected values under the response model. In principle, the larger this residual, the stronger the indication of DIF. The Lagrange multiplier (LM) test statistic (Rao, 1947) can be applied to formally test this. This statistic has a sound framework for application in IRT modelling (e.g., Glas, 1998). However, given the large sample sizes in ILSAs, even small differences will show a significant effect in the LM test. Fortunately, as illustrated in Chapter 3, one can instead focus on the size of the test statistics to identify items showing the largest DIF.

After identifying DIF, it is then up to the researcher to deal with items showing DIF. When analyses are done after pilot testing, the test items themselves can be amended. For researchers working with already available data from ILSA, other ways must be explored. For example, as is done in Chapter 3, by "splitting" the items, i.e. estimating group-specific item parameters for the worst cases of DIF, while making sure sufficient items retain the same parameters across the groups to ensure measurements can still be compared. Unfortunately, no clear guidelines are available to what extent items can be split.

Alternatively, one can argue that the item parameters should be allowed to vary across (sub)populations by default, as identical item functioning across all populations is an illusion. In Chapter 3 this is demonstrated by allowing for random item parameters (De Jong, Steenkamp, & Fox, 2007). The distribution of the parameters across (sub)populations is informative of the extent to which items functioning differs over groups. Though informative, it does not tell us which item-group interaction is the largest. It is therefore primarily informative in the first stage of scale analysis to assess general functioning of a scale across groups of respondents. When inferences are made based on a scale including random item parameters, this may lead to questions of linking: can scores still be compared validly?

Then in Chapters 3 and 5, a bi-factor structure is proposed. This approach relates strongly to the multidimensional perspective of DIF (e.g., Shealy & Stout, 1993) as the

interaction is considered to come from differences in a residual component addressed by the test. The bi-factor structure is characterized by a main component across all groups, on which consequential comparisons can be made, and a group-specific, secondary component. At item level, the item factor loadings (i.e., discrimination parameters for the dimensions) provide insight into the extent to which an item pertains to the common factor. In this approach, deletion of items showing large DIF is not necessary from a validity perspective.

When there are item characteristics in a test or questionnaire that are expected to result in an item-group interaction, the hybrid model as presented in Chapter 5 can be highly informative. If the coefficient for the interaction effect differs from zero, the effects of the item characteristics at hand should be taken into account; either by including it in the model, or at the most basic level, by studying and reporting effects at the different levels of the item characteristics.

7.3 Estimation Methods

Models in this thesis are estimated in both a frequentist (Chapters 2 to 4) and a Bayesian framework (Chapters 5 and 6). The choice for either framework was guided primarily by practical considerations. The LM test was available in the publically available MIRT/LEXTER (Glas, 2010) software which is based on marginal maximum likelihood (MML) estimation. The estimation procedure required little data handling and estimations had relatively short computational times.

For the more complex, highly dimensional or innovative models, the freely available OpenBUGS software (Lunn, Spiegelhalter, Thomas, & Best, 2009) was used to estimate the parameters of interest. The Markov chain Monte Carlo sampler provided great freedom for modelling, as little thought had to be given to the posterior distribution. Results in Chapter 6 showed high correlations (>.99) between estimates for the response models on TIMSS-2011 and PIRLS-2011 from the MML and Bayesian estimation procedure.

From this perspective, the Bayesian framework provides researchers with great flexibility in building and estimating models and this makes it recommendable. However, the estimation using the general-purpose software did take quite long. Also, it required a load of data handling from the researchers to get the data (including its grouping structure, use of booklets, and possibly testlets) correctly specified into the

program. Using dedicated samplers set up for specific models will most likely speed up the estimation time. This does, however, require a high level of statistical knowledge and programming skills from the researcher.

7.4 Conclusions

The synergy between the methodological focus on validity and on more substantive research questions throughout this thesis shows how DIF analyses can be insightful. More than a mere check, a task to tick off before the "real" questions are studied, the analyses can lead to insights into effects underlying test results. As ILSA data provide a lot of background information at the student level, a great variety of subpopulations can be studied in secondary analyses. A standard DIF check for common groups (e.g. according to country or gender) does not suffice for building a case for valid cross-group comparisons between other groups (e.g. according to language background). Throughout the studies in this thesis, it is therefore shown how, in studies with a substantive interest in group comparisons, the study of validity on both test and questionnaire items should be integrated in the methodology. Though no clear-cut one-method-fits-all strategy is presented here, the thesis shows that there are many ways to approach the issue.

7.5 Future Directions

The findings in this thesis provide several directions for educational research in general and more specific for ILSA projects and secondary analyses on ILSA data. For example, regarding research on CIL, more insight into the gender difference might be gained by assessing the identified computer literacy dimension with more challenging tasks. From the chapters on parental involvement, interest grows for an exploration of the educational context for helping with homework. This could shed light on why larger effects for parents' help with homework on student achievement were found for countries performing higher in PIRLS. The research on the role of reading in mathematics test demands for further investigation into factors that explain the negative effects on the mathematics test of being an SLL-student, even after correcting for reading proficiency.

As described in Chapter 1, ILSAs already make great efforts to ensure valid cross-national measurements of achievement. For future ILSA projects and following secondary analyses, this thesis provides an additional toolbox of ways to check for this. It also shows that this evaluation should not only be directed towards country comparisons, but to comparisons of other subpopulations as well, depending on the research goal.

Where ILSAs are moving from paper-and-pencil tests to more interactive digital assessments, this brings along motivation to check for measurement invariance across assessment modes. For example, the interaction model presented in Chapter 5 may be very useful to test for potential interactions between specific features of the digital assessment and specific student populations.

This attention to measurement invariance should not only be directed at the achievement tests, but also at questionnaire data. Since the student questionnaire requires reading demand, the issues as discussed in Chapters 5 and 6 may be relevant here as well. Also, as Chapter 3 showed, there are country differences in the functioning of questionnaire items.

In doing secondary analyses, correct handling of the data is important to come to valid findings (Rutkowski, Gonzalez, Joncas, & Von Davier, 2010). To stimulate the correct handling of data, efforts are made by the International Association for the Evaluation of Educational Achievement (IEA) to educate and support educational researchers in their use of ILSA data. Examples of these efforts are the courses offered by the IEA-ETS Research Institute (IERI, n.d.) and particularly the development of IEA International Database Analyzer software (IEA, 2017), which help researchers to generate advanced syntax for SPSS (IBM, 2013) via intuitive drag-and-drop menus. Modelling steps proposed in this thesis may provide directions for future developments in these initiatives so that barriers to study DIF are taken down.

Finally, in all studies in this thesis, country comparisons are made. Given the cross-sectional nature of the ILSA data, no causal claims can be made for reasons behind the country differences. For interpretations on where the country differences originate from, this thesis can serve as a valuable starting point.

Appendix A
Additional Tables on the Modelling of Parental Involvement

Table A.1
Country Characteristics Component 2: Help with Homework

Country	N	$\bar{X}$	PCM			GPCM		
			$\mu(\theta)$	$\sigma(\theta)$	ρ	$\mu(\theta)$	$\sigma(\theta)$	ρ
Azerbaijan, Republic of	4541	2.99	-0.95	2.02	0.76	-0.63	1.30	0.76
Australia	3234	5.27	0.53	1.23	0.79	0.33	0.80	0.80
Austria	4430	6.26	0.83	1.22	0.81	0.57	0.81	0.82
Belgium (French)	3356	3.58	-0.44	1.74	0.78	-0.30	1.16	0.79
Bulgaria	5126	4.82	-0.22	2.28	0.83	-0.13	1.50	0.84
Canada	18844	3.99	-0.04	1.41	0.77	-0.02	0.92	0.78
Chinese Taipei	4244	5.73	0.53	1.52	0.83	0.33	1.00	0.84
Colombia	3824	3.03	-0.72	1.74	0.75	-0.46	1.12	0.76
Croatia	4532	5.08	0.44	1.28	0.79	0.32	0.88	0.82
Czech Republic	4418	4.42	0.30	1.10	0.73	0.22	0.74	0.76
Denmark	4303	5.32	0.54	1.23	0.79	0.36	0.82	0.80
Finland	4410	8.31	1.45	0.92	0.78	0.96	0.64	0.80
France	4115	3.63	-0.23	1.48	0.76	-0.15	0.99	0.78
Georgia	4622	3.05	-0.83	1.90	0.76	-0.53	1.23	0.77
Germany	3195	6.05	0.72	1.33	0.82	0.49	0.90	0.84
Hong Kong, SAR	3609	5.94	0.49	1.70	0.85	0.28	1.13	0.85
Hungary	4903	3.91	-0.26	1.71	0.79	-0.15	1.13	0.80
Indonesia	4577	3.99	-0.23	1.70	0.79	-0.21	1.10	0.79
Iran, Islamic Republic of	5650	4.68	0.16	1.53	0.80	0.07	1.01	0.81
Ireland	4268	2.99	-0.69	1.68	0.75	-0.46	1.11	0.76
Israel	3271	5.84	0.63	1.38	0.82	0.43	0.92	0.83
Italy	3867	3.78	-0.22	1.57	0.78	-0.12	1.05	0.80
Lithuania	4395	5.49	0.53	1.35	0.81	0.35	0.92	0.83
Malta	3285	4.23	0.06	1.41	0.78	0.04	0.93	0.79
Netherlands	2280	9.36	1.63	1.10	0.83	1.09	0.76	0.85
New Zealand	3351	5.28	0.43	1.44	0.81	0.26	0.94	0.82
Norway	2105	2.40	-0.82	1.41	0.69	-0.54	0.93	0.70
Northern Ireland	2908	3.56	-0.20	1.39	0.75	-0.15	0.91	0.76
Poland	4923	3.82	-0.10	1.40	0.76	-0.05	0.94	0.78
Portugal	3889	3.82	-0.32	1.72	0.79	-0.23	1.14	0.80
Qatar	3653	3.20	-0.54	1.63	0.76	-0.35	1.04	0.76
Romania	4533	3.71	-0.81	2.32	0.80	-0.50	1.52	0.81
Russian Federation	4417	3.39	-0.40	1.56	0.76	-0.28	1.01	0.77
Saudi Arabia	4256	3.79	-0.33	1.74	0.79	-0.23	1.12	0.79
Singapore	6190	5.83	0.56	1.51	0.83	0.33	0.99	0.84
Slovak Republic	5489	4.99	0.31	1.47	0.81	0.22	0.99	0.83
Slovenia	4340	4.78	0.25	1.42	0.80	0.18	0.96	0.82
Spain	7945	3.15	-0.67	1.76	0.76	-0.43	1.17	0.77
Sweden	3985	4.78	0.31	1.34	0.79	0.19	0.88	0.80

(continued)

Table A.1 (continued)

Country	*N*	$\bar{X}$	PCM			GPCM		
			$\mu(\theta)$	$\sigma(\theta)$	ρ	$\mu(\theta)$	$\sigma(\theta)$	ρ
Trinidad and Tobago	3499	2.41	-1.09	1.76	0.72	-0.68	1.12	0.73
United Arab Emirates	13287	3.12	-0.61	1.67	0.76	-0.40	1.07	0.76

Note. N indicates the sample size, $\bar{X}$ the observed mean score on the component. $\mu(\theta)$ is the estimated mean, $\sigma(\theta)$ is the standard deviation, and ρ is the estimated global reliability under the partial credit model (PCM) or the generalized partial credit model (GPCM). The origin of the latent scale was identified by setting the sum of the country means to zero. The variance of the scale was identified by fixing it to one.

Table A.2
Country Characteristics Component 3: School Practices on Parental Involvement, Parent Perspective

Country	*N*	$\bar{X}$	PCM			GPCM		
			$\mu(\theta)$	$\sigma(\theta)$	ρ	$\mu(\theta)$	$\sigma(\theta)$	ρ
Azerbaijan, Republic of	4401	0.79	-2.02	1.40	0.47	-15.97	11.11	0.51
Australia	3185	3.51	0.39	0.13	0.04	3.12	1.04	0.36
Austria	4349	4.11	0.63	0.19	0.10	4.96	1.49	0.49
Belgium (French)	3269	4.14	0.63	0.13	0.04	0.63	1.00	0.35
Bulgaria	5029	2.70	-0.01	0.58	0.42	-0.07	4.62	0.67
Canada	18567	3.66	0.45	0.16	0.07	3.58	1.29	0.43
Chinese Taipei	4189	1.85	-0.35	0.23	0.10	-2.77	1.84	0.54
Colombia	3738	1.31	-1.31	1.31	0.55	-10.37	10.33	0.62
Croatia	4478	3.05	0.18	0.46	0.35	1.43	3.65	0.65
Czech Republic	4316	3.68	0.46	0.15	0.06	3.65	1.16	0.40
Denmark	4243	4.03	0.58	0.13	0.04	4.60	1.01	0.37
Finland	4348	4.53	0.73	0.10	0.02	5.77	0.79	0.28
France	3961	3.86	0.52	0.10	0.02	4.14	0.81	0.27
Georgia	4483	1.63	-0.80	0.96	0.51	-6.31	7.58	0.64
Germany	3097	3.88	0.54	0.13	0.04	4.25	0.99	0.36
Hong Kong, SAR	3593	1.51	-0.60	0.33	0.17	-4.75	2.64	0.59
Hungary	4793	3.38	0.36	0.22	0.13	2.83	1.77	0.52
Indonesia	4549	0.86	-1.65	1.09	0.42	-13.02	8.61	0.57
Iran, Islamic Republic of	5608	1.34	-1.00	0.89	0.45	-7.88	7.07	0.64
Ireland	4187	3.44	0.37	0.36	0.27	2.89	2.87	0.61
Israel	3188	2.47	-0.11	0.55	0.39	-0.87	4.34	0.64
Italy	3755	3.61	0.43	0.11	0.03	3.44	0.86	0.28
Lithuania	4347	3.45	0.38	0.22	0.13	2.99	1.77	0.51
Malta	3188	2.35	-0.25	0.80	0.51	-1.99	6.34	0.66
Netherlands	2265	4.39	0.70	0.13	0.04	5.56	1.01	0.39
New Zealand	3362	3.56	0.42	0.23	0.13	3.29	1.78	0.51
Norway	2091	3.92	0.55	0.19	0.10	4.37	1.51	0.48
Northern Ireland	2884	3.49	0.38	0.11	0.03	3.04	0.89	0.29
Poland	4790	3.25	0.32	0.15	0.05	2.50	1.15	0.38
Portugal	3745	3.60	0.43	0.11	0.03	3.44	0.86	0.28
Qatar	3610	1.87	-0.39	0.40	0.25	-3.06	3.18	0.61
Romania	4477	2.00	-0.52	0.90	0.53	-4.10	7.13	0.66
Russian Federation	4331	3.25	0.31	0.19	0.09	2.44	1.48	0.47
Saudi Arabia	4306	1.39	-1.05	1.03	0.50	-8.29	8.19	0.64
Singapore	4401	0.79	-2.02	1.40	0.47	-15.97	11.11	0.51
Slovak Republic	3185	3.51	0.39	0.13	0.04	3.12	1.04	0.36
Slovenia	4349	4.11	0.63	0.19	0.10	4.96	1.49	0.49
Spain	3269	4.14	0.63	0.13	0.04	0.63	1.00	0.35
Sweden	5029	2.70	-0.01	0.58	0.42	-0.07	4.62	0.67
Trinidad and Tobago	18567	3.66	0.45	0.16	0.07	3.58	1.29	0.43
United Arab Emirates	4189	1.85	-0.35	0.23	0.10	-2.77	1.84	0.54

Note. N indicates the sample size, $\bar{X}$ the observed mean score on the component. $\mu(\theta)$ is the estimated mean, $\sigma(\theta)$ is the standard deviation, and ρ is the estimated global reliability under the partial credit model (PCM) or the generalized partial credit model (GPCM). The origin of the latent scale was identified by setting the sum of the country means to zero. The variance of the scale was identified by fixing it to one.

Table A.3
Country Characteristics Component 4: Student Perception of Parental Involvement

Country	N	$\bar{X}$	PCM			GPCM		
			$\mu(\theta)$	$\sigma(\theta)$	ρ	$\mu(\theta)$	$\sigma(\theta)$	ρ
Azerbaijan, Republic of	4330	1.50	-0.48	0.99	0.51	-0.54	1.18	0.51
Australia	5997	3.31	0.44	0.67	0.57	0.55	0.81	0.58
Austria	4571	1.90	-0.24	0.90	0.54	-0.29	1.09	0.55
Belgium (French)	3680	2.11	-0.12	0.87	0.55	-0.17	1.09	0.57
Bulgaria	5191	2.36	-0.24	1.18	0.64	-0.29	1.44	0.64
Canada	22750	2.46	0.08	0.79	0.56	0.13	0.96	0.57
Chinese Taipei	4276	4.36	0.65	0.84	0.69	0.79	1.04	0.70
Colombia	3793	1.42	-0.74	1.17	0.53	-0.88	1.40	0.53
Croatia	4564	2.08	-0.04	0.74	0.50	-0.06	0.88	0.50
Czech Republic	4483	1.38	-0.52	0.87	0.45	-0.62	1.06	0.47
Denmark	4543	2.58	0.21	0.66	0.51	0.25	0.80	0.52
England	3912	3.30	0.47	0.61	0.53	0.54	0.73	0.54
Finland	4599	3.57	0.55	0.58	0.53	0.67	0.70	0.55
France	4403	2.31	0.01	0.80	0.55	-0.02	1.01	0.57
Georgia	4581	1.56	-0.53	1.05	0.53	-0.62	1.25	0.53
Germany	3600	1.90	-0.16	0.80	0.50	-0.13	0.97	0.53
Hong Kong, SAR	3826	5.34	0.93	0.70	0.67	1.10	0.88	0.68
Hungary	5105	1.95	-0.23	0.91	0.54	-0.31	1.09	0.54
Indonesia	4662	2.32	-0.04	0.90	0.58	-0.01	1.10	0.60
Iran, Islamic Republic of	5727	2.14	-0.07	0.83	0.54	-0.08	0.98	0.54
Ireland	4415	2.27	0.00	0.80	0.54	0.03	0.97	0.56
Israel	4117	2.46	0.06	0.83	0.57	0.08	1.00	0.58
Italy	4100	2.17	-0.01	0.76	0.52	-0.07	0.96	0.54
Lithuania	4591	2.17	-0.02	0.77	0.52	-0.04	0.93	0.53
Malta	3519	2.53	0.09	0.81	0.57	0.14	0.98	0.59
Netherlands	3955	3.56	0.48	0.72	0.61	0.52	0.89	0.61
New Zealand	5549	3.03	0.36	0.67	0.55	0.44	0.80	0.56
Northern Ireland	3523	2.37	0.16	0.61	0.46	0.17	0.75	0.47
Norway	3112	2.50	0.19	0.65	0.49	0.29	0.80	0.53
Poland	4953	2.20	-0.05	0.82	0.54	-0.10	1.01	0.55
Portugal	4037	1.91	-0.19	0.83	0.52	-0.26	1.02	0.53
Qatar	3947	2.82	0.17	0.89	0.62	0.19	1.08	0.62
Romania	4592	1.69	-0.57	1.16	0.57	-0.71	1.39	0.56
Russian Federation	4444	1.82	-0.25	0.86	0.51	-0.31	1.05	0.53
Saudi Arabia	4425	2.55	0.06	0.90	0.60	0.07	1.08	0.61
Singapore	6275	4.25	0.66	0.74	0.65	0.77	0.92	0.66
Slovak Republic	5586	1.76	-0.45	1.06	0.56	-0.57	1.29	0.56
Slovenia	4456	2.13	-0.02	0.75	0.51	-0.02	0.91	0.52
Spain	8501	2.07	-0.15	0.88	0.55	-0.18	1.07	0.56
Sweden	4533	2.45	0.18	0.64	0.49	0.23	0.79	0.51
Trinidad and Tobago	3875	1.52	-0.65	1.12	0.54	-0.75	1.34	0.55
United Arab Emirates	14209	2.23	-0.11	0.94	0.58	-0.14	1.14	0.59
United States	12501	2.72	0.15	0.84	0.60	0.20	1.02	0.61

Note. N indicates the sample size, $\bar{X}$ the observed mean score on the component. $\mu(\theta)$ is the estimated mean, $\sigma(\theta)$ is the standard deviation, and ρ is the estimated global reliability under the partial credit model (PCM) or the generalized partial credit model (GPCM). The origin of the latent scale was identified by setting the sum of the country means to zero. The variance of the scale was identified by fixing it to one.

Table A.4
Country Characteristics Component 5: School Practices on Parental Involvement, School Perspective

Country	N	$\bar{X}$	PCM			GPCM		
			$\mu(\theta)$	$\sigma(\theta)$	ρ	$\mu(\theta)$	$\sigma(\theta)$	ρ
Azerbaijan, Republic of	169	32.89	0.32	0.55	0.70	0.59	0.90	0.73
Australia	269	35.41	0.64	0.64	0.74	1.04	1.06	0.76
Austria	158	31.00	0.05	0.47	0.64	-0.14	0.81	0.71
Belgium (French)	118	23.37	-0.80	0.59	0.73	-1.13	0.92	0.74
Bulgaria	147	30.14	-0.04	0.70	0.79	0.10	1.10	0.81
Canada	1084	33.37	0.38	0.67	0.76	0.56	1.09	0.79
Chinese Taipei	150	34.35	0.54	0.88	0.83	0.77	1.48	0.85
Colombia	149	32.73	0.35	0.77	0.81	0.87	1.29	0.81
Croatia	152	30.50	-0.02	0.46	0.63	0.01	0.74	0.68
Czech Republic	174	28.28	-0.29	0.46	0.64	-0.30	0.81	0.72
Denmark	231	25.93	-0.54	0.42	0.60	-1.03	0.76	0.67
England	120	32.83	0.29	0.57	0.71	0.30	0.89	0.74
Finland	139	25.60	-0.59	0.50	0.68	-1.01	0.85	0.72
France	167	27.52	-0.35	0.57	0.73	-0.60	0.99	0.78
Georgia	171	30.85	0.07	0.69	0.79	0.24	1.17	0.82
Germany	187	30.16	-0.05	0.50	0.68	-0.19	0.82	0.72
Hong Kong, SAR	125	30.18	-0.04	0.63	0.76	-0.31	1.05	0.80
Hungary	143	29.06	-0.18	0.52	0.69	-0.14	0.85	0.73
Indonesia	155	27.53	-0.37	0.75	0.82	-0.60	1.21	0.84
Iran, Islamic Republic of	244	32.60	0.30	0.88	0.85	0.65	1.46	0.85
Ireland	145	27.75	-0.32	0.62	0.76	-0.75	1.07	0.81
Israel	132	32.24	0.25	0.71	0.79	0.31	1.12	0.81
Italy	200	27.80	-0.26	0.65	0.77	-0.38	1.08	0.80
Lithuania	151	30.42	-0.02	0.49	0.67	-0.07	0.83	0.72
Malta	93	30.99	0.09	0.59	0.73	0.02	0.89	0.75
Netherlands	117	26.97	-0.42	0.43	0.61	-0.80	0.77	0.69
New Zealand	175	34.13	0.47	0.57	0.70	0.56	0.88	0.73
Northern Ireland	117	29.23	-0.17	0.55	0.71	-0.53	0.91	0.76
Norway	115	26.03	-0.54	0.39	0.56	-0.93	0.66	0.62
Poland	148	31.57	0.15	0.53	0.69	0.22	0.83	0.71
Portugal	147	29.62	-0.10	0.62	0.76	-0.16	0.98	0.78
Qatar	166	33.86	0.53	0.96	0.85	1.02	1.59	0.85
Romania	147	32.91	0.34	0.71	0.78	0.76	1.21	0.81
Russian Federation	202	34.42	0.46	0.46	0.62	0.69	0.82	0.70
Saudi Arabia	169	26.57	-0.48	0.83	0.85	-0.55	1.43	0.88
Singapore	176	32.40	0.25	0.60	0.73	0.31	1.05	0.79
Slovak Republic	194	28.70	-0.22	0.56	0.72	-0.20	0.97	0.77
Slovenia	191	29.25	-0.14	0.47	0.65	-0.21	0.81	0.71
Spain	302	28.96	-0.18	0.65	0.78	-0.30	1.08	0.81
Sweden	132	27.33	-0.38	0.55	0.71	-0.57	0.89	0.75
Trinidad and Tobago	147	30.84	0.07	0.78	0.82	0.43	1.34	0.85
United Arab Emirates	419	32.42	0.29	0.80	0.82	0.45	1.25	0.83
United States	331	35.36	0.66	0.73	0.78	1.01	1.29	0.81

Note. N indicates the sample size, $\bar{X}$ the observed mean score on the component. $\mu(\theta)$ is the estimated mean, $\sigma(\theta)$ is the standard deviation, and ρ is the estimated global reliability under the partial credit model (PCM) or the generalized partial credit model (GPCM). The origin of the latent scale was identified by setting the sum of the country means to zero. The variance of the scale was identified by fixing it to one.

Table A.5
Response Frequencies and Item Parameter Estimates Under the Generalized Partial Credit Model for Items in Component 2: Help with Homework

Item	Slope	Intercept	*I(0)*	Relative frequency response categories			
				Cat0	Cat1	Cat2	Cat3
ASBH09A	1.17	2.44	0.21	0.78	0.18	0.03	0.02
ASBH09B	1.63	2.01	0.78	0.56	0.32	0.07	0.05
ASBH09C	1.15	2.15	0.18	0.81	0.14	0.03	0.03
ASBH09D	1.10	2.27	0.24	0.73	0.22	0.03	0.02
ASBH09E	1.56	2.51	0.31	0.77	0.17	0.03	0.04
ASBH09F	1.69	1.09	1.25	0.43	0.33	0.10	0.14
ASBH09G	2.26	1.87	1.62	0.43	0.37	0.12	0.08
ASBH09H	1.45	1.66	0.77	0.47	0.38	0.11	0.04

Note. Slope and intercept are the parameters a_{i0} and the mean of the location parameters b_{i1}, b_{i2},… respectively, under the general partial credit model (GPCM). *I*(0) is the information value of the item at $\theta = 0$. *Cat0, Cat1, Cat2,* and *Cat3* indicate the frequency with which item categories 0, 1, 2 and 3 are endorsed, respectively. The content of the components, items and corresponding category labels are described in Table 3.1.

Table A.6
Response Frequencies and Item Parameter Estimates Under the Generalized Partial Credit Model for Items in Component 3: School Practices on Parental Involvement, Parent Perspective

Item	Slope	Intercept	*I(0)*	Relative frequency response categories			
				Cat0	Cat1	Cat2	Cat3
ASBH10A	0.61	1.41	0.15	0.54	0.37	0.07	0.02
ASBH10B	0.61	0.36	0.34	0.30	0.31	0.23	0.16
ASBH10E	0.58	0.52	0.30	0.38	0.29	0.19	0.14

Note. Slope and intercept are the parameters a_{i0} and the mean of the location parameters b_{i1}, b_{i2},… respectively, under the general partial credit model (GPCM). *I*(0) is the information value of the item at $\theta = 0$. *Cat0, Cat1, Cat2,* and *Cat3* indicate the frequency with which item categories 0, 1, 2 and 3 are endorsed, respectively. The content of the components, items and corresponding category labels are described in Table 3.1.

Table A.7
Response Frequencies and Item Parameter Estimates Under the Generalized Partial Credit Model for Items in Component 4: Student Perception of Parental Involvement

Item	Slope	Intercept	*I(0)*	Relative frequency response categories			
				Cat0	Cat1	Cat2	Cat3
ASBG07A	1.01	1.47	0.32	0.67	0.21	0.05	0.07
ASBG07B	0.96	1.15	0.43	0.56	0.27	0.08	0.09
ASBG07C	0.85	1.35	0.22	0.75	0.14	0.04	0.08
ASBG07D	0.77	1.21	0.22	0.73	0.14	0.04	0.09
ASBR09C	0.55	1.55	0.09	0.76	0.18	0.04	0.02

Note. Slope and intercept are the parameters a_{i0} and the mean of the location parameters b_{i1}, b_{i2},… respectively, under the general partial credit model (GPCM). *I*(0) is the information value of the item at $\theta = 0$. *Cat0, Cat1, Cat2,* and *Cat3* indicate the frequency with which item categories 0, 1, 2 and 3 are endorsed, respectively. The content of the components, items and corresponding category labels are described in Table 3.1.

Table A.8
Response Frequencies and Item Parameter Estimates Under the Generalized Partial Credit Model for Items in Component 5: School Practices on Parental Involvement, School Perspective

Item	Slope	Intercept	*I(0)*	Relative frequency response categories				
				Cat0	Cat1	Cat2	Cat3	Cat4
ACBG11AA	0.75	-2.88	0.14	0.00	0.01	0.37	0.62	-
ACBG11AB	0.91	-2.95	0.20	0.00	0.02	0.33	0.65	-
ACBG11AC	0.87	-2.34	0.23	0.00	0.05	0.39	0.56	-
ACBG11AD	0.57	-1.34	0.14	0.03	0.07	0.29	0.62	-
ACBG11BA	0.47	-0.66	0.16	0.07	0.16	0.37	0.40	-
ACBG11BB	0.51	-0.64	0.20	0.06	0.30	0.32	0.32	-
ACBG11CA	0.70	-0.55	0.29	0.05	0.33	0.38	0.23	-
ACBG11CB	0.84	-1.29	0.38	0.03	0.18	0.35	0.44	-
ACBG11CC	1.27	-1.27	0.72	0.01	0.38	0.37	0.24	-
ACBG11CD	1.13	-1.42	0.66	0.01	0.45	0.29	0.25	-
ACBG11CE	1.09	-1.10	0.60	0.03	0.30	0.39	0.29	-
ACBG11CF	0.41	0.02	0.18	0.23	0.25	0.27	0.25	-
ACBG11CG	0.52	0.26	0.23	0.24	0.31	0.30	0.15	-
ACBG12E	0.25	-0.35	0.05	0.02	0.13	0.46	0.31	0.09
ACBG12F	0.20	-0.18	0.03	0.04	0.17	0.46	0.26	0.08

Note. Slope and intercept are the parameters a_{i0} and the mean of the location parameters b_{i1}, b_{i2},… respectively, under the general partial credit model (GPCM). $I(0)$ is the information value of the item at $\theta = 0$. *Cat0, Cat1, Cat2,* and *Cat3* indicate the frequency with which item categories 0, 1, 2 and 3 are endorsed, respectively. The content of the components, items and corresponding category labels are described in Table 3.1.

Table A.9
Item Parameter Estimates Under the Generalized Partial Credit Model (GPCM) and GPCM with Random Item Parameters for Items in Component 2: Help with Homework

Item	GPCM		GPCM random item parameters			
	Slope	Intercept	Slope	*SD* (Slope)	Intercept	*SD* (Intercept)
ASBH09A	1.17	2.44	1.331	0.619	3.686	1.547
ASBH09B	1.63	2.01	1.313	0.534	2.947	1.880
ASBH09C	1.15	2.15	1.396	0.554	2.199	1.203
ASBH09D	1.10	2.27	1.227	0.314	3.736	1.610
ASBH09E	1.56	2.51	1.437	0.634	3.446	1.208
ASBH09F	1.69	1.09	1.477	0.503	0.707	1.251
ASBH09G	2.26	1.87	1.308	0.434	0.796	1.154
ASBH09H	1.45	1.66	1.559	0.224	1.518	1.210

Note. SD (Slope) indicates the standard deviation of the slope. *SD* (Intercept) indicates the standard deviation of the intercept. The content of the items and corresponding category labels are described in Table 3.1.

Table A.10
Item Parameter Estimates Under the Generalized Partial Credit Model (GPCM) and GPCM with Random Item Parameters for Items in Component 3: School Practices on Parental Involvement, Parent Perspective

Item	GPCM		GPCM random item parameters			
	Slope	Intercept	Slope	*SD* (Slope)	Intercept	*SD* (Intercept)
ASBH10A	0.61	1.41	1.218	1.388	4.477	4.172
ASBH10B	0.61	0.36	4.144	1.601	2.751	4.923
ASBH10E	0.58	0.52	3.843	1.791	3.469	5.232

Note. *SD* (Slope) indicates the standard deviation of the slope. *SD* (Intercept) indicates the standard deviation of the intercept. The content of the items and corresponding category labels are described in Table 3.1.

Table A.11
Item Parameter Estimates Under the Generalized Partial Credit Model (GPCM) and GPCM with Random Item Parameters for Items in Component 4: Student Perception of Parental Involvement

Item	GPCM		GPCM random item parameters			
	Slope	Intercept	Slope	*SD* (Slope)	Intercept	*SD* (Intercept)
ASBG07A	1.01	1.47	0.924	0.161	1.473	1.102
ASBG07B	0.96	1.15	0.994	0.357	1.155	0.943
ASBG07C	0.85	1.35	0.989	0.316	1.937	2.614
ASBG07D	0.77	1.21	0.990	0.240	1.917	3.017
ASBR09C	0.55	1.55	0.553	0.050	2.100	2.782

Note. *SD* (Slope) indicates the standard deviation of the slope. *SD* (Intercept) indicates the standard deviation of the intercept. The content of the items and corresponding category labels are described in Table 3.1.

Table A.12
Item Parameter Estimates Under the Generalized Partial Credit Model (GPCM) and GPCM with Random Item Parameters for Items in Component 5: School Practices on Parental Involvement, School Perspective

Item	GPCM		GPCM random item parameters			
	Slope	Intercept	Slope	*SD* (Slope)	Intercept	*SD* (Intercept)
ACBG11AA	0.75	-2.88	0.689	0.664	-1.667	1.396
ACBG11AB	0.91	-2.95	1.029	0.377	-2.122	0.797
ACBG11AC	0.87	-2.34	0.998	0.506	-2.110	0.778
ACBG11AD	0.57	-1.34	0.466	1.042	-1.480	0.461
ACBG11BA	0.47	-0.66	0.645	0.876	-0.581	1.033
ACBG11BB	0.51	-0.64	0.627	0.807	-0.583	0.462
ACBG11CA	0.70	-0.55	0.887	0.491	-0.576	0.434
ACBG11CB	0.84	-1.29	0.890	0.621	-1.120	0.614
ACBG11CC	1.27	-1.27	1.236	0.620	-0.995	0.682
ACBG11CD	1.13	-1.42	1.194	0.515	-1.122	0.625
ACBG11CE	1.09	-1.10	1.132	0.229	-1.023	0.168
ACBG11CF	0.41	0.02	0.548	0.738	0.029	0.342
ACBG11CG	0.52	0.26	0.737	0.514	0.071	0.781
ACBG12E	0.25	-0.35	0.123	1.453	0.551	1.954
ACBG12F	0.20	-0.18	0.279	1.431	-0.030	1.789

Note. *SD* (Slope) indicates the standard deviation of the slope. *SD* (Intercept) indicates the standard deviation of the intercept. The content of the items and corresponding category labels are described in Table 3.1.

Table A.13
Absolute Differential Item Functioning (DIF) Under the Partial Credit Model (PCM) and the Generalized Partial Credit Model (GPCM) and Standard Deviations of Random Item Parameters on Items in Component 2: Help with Homework

Item	PCM	GPCM	*SD* (Slope)	*SD* (Intercept)
ASBH09A	0.11	0.12	0.619	1.547
ASBH09B	0.07	0.07	0.534	1.880
ASBH09C	0.10	0.10	0.554	1.203
ASBH09D	0.10	0.10	0.314	1.610
ASBH09E	0.08	0.08	0.634	1.208
ASBH09F	0.14	0.12	0.503	1.251
ASBH09G	0.08	0.06	0.434	1.154
ASBH09H	0.07	0.07	0.224	1.210

Note. The columns labeled PCM and GPCM give the mean residuals as estimated under the unidimensional versions of these two models. *SD* (Slope) indicates the standard deviation of the slope. *SD* (Intercept) indicates the standard deviation of the intercept. The content of the items and corresponding category labels are described in Table 3.1.

Table A.14
Absolute Differential Item Functioning (DIF) Under the Partial Credit Model (PCM) and the Generalized Partial Credit Model (GPCM) and Standard Deviations of Random Item Parameters on Items in Component 3: School Practices on Parental Involvement, Parent Perspective

Item	PCM	GPCM	*SD* (Slope)	*SD* (Intercept)
ASBH10A	0.13	0.47	1.388	4.172
ASBH10B	0.07	0.36	1.601	4.923
ASBH10E	0.09	0.38	1.791	5.232

Note. The columns labeled PCM and GPCM give the mean residuals as estimated under the unidimensional versions of these two models. *SD* (Slope) indicates the standard deviation of the slope. *SD* (Intercept) indicates the standard deviation of the intercept. The content of the items and corresponding category labels are described in Table 3.1.

Table A.15
Absolute Differential Item Functioning (DIF) Under the Partial Credit Model (PCM) and the Generalized Partial Credit Model (GPCM) and Standard Deviations of Random Item Parameters on Items in Component 4: Student Perception of Parental Involvement

Item	PCM	GPCM	*SD* (Slope)	*SD* (Intercept)
ASBG07A	0.08	0.07	0.161	1.102
ASBG07B	0.09	0.08	0.357	0.943
ASBG07C	0.07	0.08	0.316	2.614
ASBG07D	0.12	0.12	0.240	3.017
ASBR09C	0.07	0.08	0.050	2.782

Note. The columns labeled PCM and GPCM give the mean residuals as estimated under the unidimensional versions of these two models. *SD* (Slope) indicates the standard deviation of the slope. *SD* (Intercept) indicates the standard deviation of the intercept. The content of the items and corresponding category labels are described in Table 3.1.

Table A.16
Absolute Differential Item Functioning (DIF) Under the Partial Credit model (PCM) and the Generalized Partial Credit Model (GPCM) and Standard Deviations of Random Item Parameters on Items in Component 5: School Practices on Parental Involvement, School Perspective

Item	PCM	GPCM	*SD* (Slope)	*SD* (Intercept)
ACBG11AA	0.23	0.21	0.664	1.396
ACBG11AB	0.19	0.17	0.377	0.797
ACBG11AC	0.17	0.16	0.506	0.778
ACBG11AD	0.16	0.16	1.042	0.461
ACBG11BA	0.32	0.35	0.876	1.033
ACBG11BB	0.24	0.24	0.807	0.462
ACBG11CA	0.20	0.18	0.491	0.434
ACBG11CB	0.22	0.23	0.621	0.614
ACBG11CC	0.15	0.13	0.620	0.682
ACBG11CD	0.21	0.17	0.515	0.625
ACBG11CE	0.11	0.11	0.229	0.168
ACBG11CF	0.29	0.32	0.738	0.342
ACBG11CG	0.32	0.34	0.514	0.781
ACBG12E	0.26	0.27	1.453	1.954
ACBG12F	0.25	0.24	1.431	1.789

Note. The columns labeled PCM and GPCM give the mean residuals as estimated under the unidimensional versions of these two models. *SD* (Slope) indicates the standard deviation of the slope. *SD* (Intercept) indicates the standard deviation of the intercept. The content of the items and corresponding category labels are described in Table 3.1.

Table A.17
Residual Analysis for Country-by-Item Interactions for Component 2: Help with Homework

Country	*N*	Item								10% CDIF	20% CDIF	Absolute residual
		1	2	3	4	5	6	7	8			
Azerbaijan, Republic of	4541	+								0	1	0.084
Australia	3234			+			--			1	2	0.097
Austria	4430				-					0	1	0.105
Belgium (French)	3356									0	0	0.056
Bulgaria	5126									0	0	0.056
Canada	18844									0	0	0.027
Chinese Taipei	4244	++			+					1	2	0.100
Colombia	3824									0	0	0.032
Croatia	4532	--		-	--		++			3	4	0.182
Czech Republic	4418	-				-	++		++	2	4	0.148
Denmark	4303	-				+				0	2	0.057
Finland	4410	-		--	++		++			3	4	0.131
France	4115		-						+	0	2	0.097
Georgia	4622		++							1	1	0.108
Germany	3195									0	0	0.065
Hong Kong, SAR	3609	+			++		-	-		1	4	0.132
Hungary	4903				-	+				0	2	0.083
Indonesia	4577	++	+	++			--	--		4	5	0.188
Iran, Islamic Republic of	5650			++					-	1	2	0.125
Ireland	4268									0	0	0.059
Israel	3271					+				0	1	0.076
Italy	3867	-						+		0	2	0.098
Lithuania	4395	-	-			+	++			1	4	0.134
Malta	3285		++							1	1	0.083
Netherlands	2280	++		++	--	++	-	++	--	6	7	0.249
New Zealand	3351			++						1	1	0.097
Northern Ireland	2105						--			1	1	0.074
Norway	2908									0	0	0.055
Poland	4923	-					+			0	2	0.080
Portugal	3889	++								1	1	0.069
Qatar	3653									0	0	0.045
Romania	4533									0	0	0.044
Russian Federation	4417				++					1	1	0.068
Saudi Arabia	4256						--			1	1	0.089
Singapore	6190	++			+			-	+	1	4	0.129
Slovak Republic	5489	--					++			2	2	0.115
Slovenia	4340			-						0	1	0.080
Spain	7945									0	0	0.064
Sweden	3985									0	0	0.057
Trinidad and Tobago	3499									0	0	0.038
United Arab Emirates	13287									0	0	0.037

Note. + indicates that residual belongs to the 20% most positive residuals, ++ indicates that residual even belongs to the 10% most positive residuals. - indicates that residual belongs to the 20% most negative residuals, -- indicates that residual belongs to the 10% most negative residuals. The 10% cultural differential item functioning (CDIF) and 20% CDIF columns give the number of outliers in the two respective regions. Absolute residual refers to the means over items of the absolute values of the residuals. The content of the items is described in Table 3.1.

Table A.18
Residual Analysis for Country-by-Item Interactions for Component 3: School Practices on Parental Involvement, Parent Perspective

Country	*N*	Item			10%	20%	Absolute
		1	2	3	CDIF	CDIF	residual
Azerbaijan, Republic of	4401			+	0	1	0.084
Australia	3185				0	0	0.032
Austria	4349	++			1	1	0.102
Belgium (French)	3269	+			0	1	0.088
Bulgaria	5029	+			0	1	0.110
Canada	18567				0	0	0.058
Chinese Taipei	4189				0	0	0.057
Colombia	3738	--			1	1	0.112
Croatia	4478	--			1	1	0.090
Czech Republic	4316	++			1	1	0.085
Denmark	4243	--			1	1	0.071
Finland	4348	++			1	1	0.096
France	3961	+			0	1	0.081
Georgia	4483	--			1	1	0.088
Germany	3097	++			1	1	0.164
Hong Kong, SAR	3593	+			0	1	0.054
Hungary	4793				0	0	0.026
Indonesia	4549	-	+		0	2	0.142
Iran, Islamic Republic of	5608				0	0	0.034
Ireland	4187				0	0	0.073
Israel	3188				0	0	0.042
Italy	3755	+			0	1	0.106
Lithuania	4347				0	0	0.029
Malta	3188	--			1	1	0.082
Netherlands	2265				0	0	0.039
New Zealand	3362				0	0	0.037
Northern Ireland	2091				0	0	0.030
Norway	2884	--			0	0	0.104
Poland	4790				0	0	0.050
Portugal	3745				0	0	0.037
Qatar	3610	+			0	1	0.075
Romania	4477	--			1	1	0.127
Russian Federation	4331		+		0	1	0.088
Saudi Arabia	4306	-			0	1	0.048
Singapore	6145	++			1	1	0.083
Slovak Republic	5344				0	0	0.049
Slovenia	4246	++			1	1	0.072
Spain	7699				0	0	0.018
Sweden	3974				0	0	0.016
Trinidad and Tobago	3328	-			0	1	0.109
United Arab Emirates	13061				0	0	0.044

Note. + indicates that residual belongs to the 20% most positive residuals, ++ indicates that residual even belongs to the 10% most positive residuals. - indicates that residual belongs to the 20% most negative residuals, -- indicates that residual even belongs to the 10% most negative residuals. The 10% cultural differential item functioning (CDIF) and 20% CDIF columns give the number of outliers in the two respective regions. Absolute residual refers to the means over items of the absolute values of the residuals. The content of the items is described in Table 3.1.

Table A.19
Residual Analysis for Country-by-Item Interactions for Component 4: Student Perception of Parental Involvement

Country	*N*	Item					10% CDIF	20% CDIF	Absolute residual
		1	2	3	4	5			
Azerbaijan, Republic of	4330						0	0	0.040
Australia	5997						0	0	0.060
Austria	4571						0	0	0.037
Belgium (French)	3680				-		0	1	0.076
Bulgaria	5191				+		0	1	0.075
Canada	22750						0	0	0.068
Chinese Taipei	4276	++			--		2	2	0.117
Colombia	3793						0	0	0.034
Croatia	4564	-			++		1	2	0.094
Czech Republic	4483						0	0	0.051
Denmark	4543						0	0	0.056
England	3912	+		-			0	2	0.088
Finland	4599					++	1	1	0.103
France	4403						0	0	0.068
Georgia	4581			+			0	1	0.075
Germany	3600	+			--		1	2	0.146
Hong Kong, SAR	3826		-				0	1	0.087
Hungary	5105	-			++		1	2	0.110
Indonesia	4662					--	1	1	0.080
Iran, Islamic Republic of	5727			+			0	1	0.071
Ireland	4415		+		--		1	2	0.112
Israel	4117				++		1	1	0.120
Italy	4100					+	0	1	0.066
Lithuania	4591						0	0	0.061
Malta	3519					-	0	1	0.078
Netherlands	3955		--	++	++	++	4	4	0.233
New Zealand	5549						0	0	0.027
Northern Ireland	3523	+	++		--	-	2	4	0.197
Norway	3112			+			0	1	0.087
Poland	4953		+		--	++	2	3	0.158
Portugal	4037				+		0	1	0.074
Qatar	3947			+			0	1	0.089
Romania	4592				++		1	1	0.082
Russian Federation	4444						0	0	0.073
Saudi Arabia	4425		-	++			1	2	0.133
Singapore	6275			--		++	2	2	0.104
Slovak Republic	5586						0	0	0.070
Slovenia	4456						0	0	0.059
Spain	8501						0	0	0.056
Sweden	4533	+			--		1	2	0.090
Trinidad and Tobago	3875						0	0	0.049
United Arab Emirates	14209						0	0	0.060
United States	12501						0	0	0.084

Note. + indicates that residual belongs to the 20% most positive residuals, ++ indicates that residual even belongs to the 10% most positive residuals. - indicates that residual belongs to the 20% most negative residuals, -- indicates that residual even belongs to the 10% most negative residuals. The 10% cultural differential item functioning (CDIF) and 20% CDIF columns give the number of outliers in the two respective regions. Absolute residual refers to the means over items of the absolute values of the residuals. The content of the items is described in Table 3.1.

Table A.20
Residual Analysis for Country-by-Item Interactions for Component 5: School Practices on Parental Involvement, School Perspective

Country	*N*	Item															10% CDIF	20% CDIF	Absolute residual
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			
Azerbaijan	169					--											1	1	0.164
Australia	269								+								0	1	0.152
Austria	158					++		+			-		+	-			1	5	0.294
Belgium (French)	118	+		+		-								--			1	4	0.289
Bulgaria	147	+														-	0	2	0.227
Canada	1084					+											0	1	0.170
Chinese Taipei	150							-						++	++		2	3	0.186
Colombia	149	+											--		--		2	3	0.268
Croatia	152	++				--			-					++			3	4	0.352
Czech Republic	174	+					-		+					--		--	2	5	0.286
Denmark	231						-		+				++			++	2	4	0.235
England	120	+						-						-			0	3	0.209
Finland	139		--						++				++	++	++	+	5	6	0.294
France	167					++	++			+			-	--			3	5	0.271
Georgia	171	+											--		-		1	3	0.195
Germany	187					++	++										2	2	0.173
Hong Kong	125					+								++			1	2	0.250
Hungary	143	++				+	-						--		--		3	5	0.237
Indonesia	155					--	+	+						--	++	+	3	6	0.302
Iran	244					--	+										1	2	0.203
Ireland	145							-	++				++		++	+	3	5	0.302
Israel	132																0	0	0.121
Italy	200	++							-								1	2	0.200
Lithuania	151												--				1	1	0.162
Malta	93													+			0	1	0.124
Netherlands	117				--	++							+			+	2	4	0.213
New Zealand	175					++			++						+	+	2	4	0.247
Northern Ireland	117												+	-			0	2	0.156
Norway	115	++				+	-				--				-		2	5	0.289
Poland	148					+			++					++			2	3	0.213
Portugal	147	++		+							-						1	3	0.226
Qatar	166					-	--										1	2	0.153

(continued)

Table A.20 (continued)

Country	*N*	Item														10% CDIF	20% CDIF	Absolute residual
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15			
Romania	147						+								--	--	2	3	0.223
Russian Fed.	202	++					+			-	--		--	++	-		4	7	0.367
Saudi Arabia	169	++				--			--		+		-			--	4	6	0.315
Singapore	176					+								++			1	2	0.223
Slovak Republic	194														--	-	1	2	0.200
Slovenia	191													++			1	1	0.176
Spain	302	+							-								0	2	0.176
Sweden	132									+			++	-			1	3	0.226
Trinidad and Tobago	147										+		--	--	--		3	4	0.232
United Arab Emirates	419																0	0	0.128
United States	331																0	0	0.115

Note. + indicates that residual belongs to the 20% most positive residuals, ++ indicates that residual even belongs to the 10% most positive residuals. - indicates that residual belongs to the 20% most negative residuals, -- indicates that residual even belongs to the 10% most negative residuals. The 10% cultural differential item functioning (CDIF) and 20% CDIF columns give the number of outliers in the two respective regions. Absolute residual refers to the means over items of the absolute values of the residuals. The content of the items is described in Table 3.1.

Table A.21
Outliers of Country-Specific Factor Loadings in the Bi-Factor Model for Component 2: Help with Homework

Country	Item								2.5%	5%	Mean absolute
	1	2	3	4	5	6	7	8	Outlier	Outlier	loading
Azerbaijan, Republic of	+					-			1	2	0.056
Australia						--	--	--	3	3	0.060
Austria								-	0	1	0.037
Belgium (French)	++								0	1	0.043
Bulgaria						--	--	--	3	3	0.051
Canada	+								1	1	0.044
Chinese Taipei						-	--		1	2	0.037
Colombia	--			--		+		-	3	4	0.067
Croatia	--			-	--		-	-	2	5	0.068
Czech Republic							-	-	0	2	0.040
Denmark									0	0	0.029
Finland									0	0	0.047
France	+	+				-	-	-	2	5	0.069
Georgia								-	0	1	0.032
Germany	+					-	-		1	3	0.055
Hong Kong, SAR				-				-	0	2	0.044
Hungary	+					--	--	-	3	4	0.076
Indonesia						--	--	-	2	3	0.057
Iran, Islamic Republic of									0	0	0.048
Ireland									0	0	0.029
Israel	--			--				--	3	3	0.056
Italy									0	0	0.041
Lithuania		++				--	--	--	3	4	0.075
Malta	+		+	--	++		++		3	5	0.072
Netherlands						--	--	--	3	3	0.058
New Zealand						--	--	--	3	3	0.059
Northern Ireland	--							--	2	2	0.053
Norway	+					-	--	--	3	4	0.066
Poland						--	--	-	2	3	0.050
Portugal	++								0	1	0.037
Qatar						--	--		2	2	0.041
Romania						--	--		2	2	0.051
Russian Federation	+						--		2	2	0.043
Saudi Arabia				-		+			1	2	0.048
Singapore									0	0	0.036
Slovak Republic								-	0	1	0.043
Slovenia						-	-	-	0	3	0.051
Spain	+					--			2	2	0.038
Sweden						--	--	-	2	3	0.055
Trinidad and Tobago							--		1	1	0.049
United Arab Emirates								-	0	1	0.036

Note. + indicates factor loading belongs to the 5% most positive loading, ++ indicates factor loading belongs to the 2.5% most positive loading. - indicates factor loading belongs to the 5% most negative loading, -- indicates factor loading belongs to the 2.5% most negative loading. The 2.5% cultural differential item functioning (CDIF) and 5% CDIF columns give the number of outliers in the two respective regions. Mean absolute loading refers to the means over items of the absolute values of country-specific factor loadings. The content of the items is described in Table 3.1.

Table A.22
Outliers of Country-Specific Factor Loadings in the Bi-Factor Model for Component 3: School Practices on Parental Involvement, Parent Perspective

Country	Item			2.5%	5%	Mean absolute
	1	2	3	Outlier	Outlier	loading
Azerbaijan, Republic of	+	++		1	2	1.097
Australia				0	0	0.293
Austria				0	0	0.203
Belgium (French)				0	0	0.223
Bulgaria				0	0	0.262
Canada				0	0	0.423
Chinese Taipei			+	1	1	0.640
Colombia				0	0	0.159
Croatia				0	0	0.194
Czech Republic				0	0	0.393
Denmark				0	0	0.284
Finland				0	0	0.293
France				0	0	0.293
Georgia				0	0	0.240
Germany				0	0	0.409
Hong Kong, SAR				0	0	0.168
Hungary			+	1	1	1.521
Indonesia			++	0	1	0.500
Iran, Islamic Republic of				0	0	0.362
Ireland				0	0	0.279
Israel				0	0	0.216
Italy				0	0	0.131
Lithuania				0	0	0.174
Malta				0	0	0.418
Netherlands				0	0	0.331
New Zealand				0	0	0.260
Northern Ireland				0	0	0.321
Norway				0	0	0.228
Poland				0	0	0.213
Portugal				0	0	0.205
Qatar				0	0	0.297
Romania				0	0	0.430
Russian Federation				0	0	0.153
Saudi Arabia				0	0	0.184
Singapore				0	0	0.150
Slovak Republic				0	0	0.180
Slovenia				0	0	0.228
Spain				0	0	0.347
Sweden				0	0	0.315
Trinidad and Tobago				0	0	0.517
United Arab Emirates				0	0	0.175

Note. + indicates factor loading belongs to the 5% most positive loading, ++ indicates factor loading belongs to the 2.5% most positive loading. - indicates factor loading belongs to the 5% most negative loading, -- indicates factor loading belongs to the 2.5% most negative loading. The 2.5% cultural differential item functioning (CDIF) and 5% CDIF columns give the number of outliers in the two respective regions. Mean absolute loading refers to the means over items of the absolute values of country-specific factor loadings. The content of the items is described in Table 3.1.

Table A.23
Outliers of Country-Specific Factor Loadings in the Bi-Factor Model for Component 4: Student Perception of Parental Involvement

Country	Item					2.5%	5%	Mean absolute
	1	2	3	4	5	Outlier	Outlier	loading
Azerbaijan, Republic of						0	0	0.024
Australia						0	0	0.012
Austria						0	0	0.021
Belgium (French)				+		1	1	0.026
Bulgaria						0	0	0.016
Canada	+		++			1	2	0.048
Chinese Taipei						0	0	0.012
Colombia	--	-				1	2	0.044
Croatia	-	--	--	+		3	4	0.084
Czech Republic						0	0	0.018
Denmark						0	0	0.010
England						0	0	0.027
Finland		-				0	1	0.038
France				-	-	0	2	0.035
Georgia	++	++	+	++		1	4	0.057
Germany	--					1	1	0.034
Hong Kong, SAR						0	0	0.011
Hungary						0	0	0.026
Indonesia				-		0	1	0.029
Iran, Islamic Republic of				+		1	1	0.034
Ireland			-			0	1	0.022
Israel						0	0	0.024
Italy						0	0	0.023
Lithuania		--	++	+		2	3	0.052
Malta						0	0	0.008
Netherlands			-	--		1	2	0.048
New Zealand						0	0	0.031
Northern Ireland		-				0	1	0.038
Norway						0	0	0.021
Poland				++		0	1	0.025
Portugal				-		0	1	0.030
Qatar			++			0	1	0.034
Romania	++					0	1	0.037
Russian Federation						0	0	0.027
Saudi Arabia			-			0	1	0.028
Singapore						0	0	0.031
Slovak Republic						0	0	0.016
Slovenia						0	0	0.018
Spain						0	0	0.020
Sweden						0	0	0.022
Trinidad and Tobago						0	0	0.020
United Arab Emirates			--	--		2	2	0.047
United States						0	0	0.019

Note. + indicates factor loading belongs to the 5% most positive loading, ++ indicates factor loading belongs to the 2.5% most positive loading. - indicates factor loading belongs to the 5% most negative loading, -- indicates factor loading belongs to the 2.5% most negative loading. The 2.5% cultural differential item functioning (CDIF) and 5% CDIF columns give the counts of the outliers in the two respective regions. Mean absolute loading refers to the means over items of the absolute values of country-specific factor loadings. The content of the items is described in Table 3.1.

Table A.24
Outliers of Country-Specific Factor Loadings in the Bi-Factor Model for Component 5: School Practices on Parental Involvement, School Perspective

Country	Item															2.5%	5%	Mean absolute
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Outlier	Outlier	loading
Azerbaijan																0	0	0.043
Australia								+								1	1	0.078
Austria																0	0	0.048
Belgium (French)	+	+														2	2	0.075
Bulgaria																0	0	0.053
Canada																0	0	0.057
Chinese Taipei	+	+														2	2	0.121
Colombia					--											1	1	0.066
Croatia																0	0	0.043
Czech Republic																0	0	0.047
Denmark																0	0	0.041
England																0	0	0.041
Finland		+														1	1	0.066
France																0	0	0.028
Georgia													++			0	1	0.047
Germany	+															1	1	0.060
Hong Kong																0	0	0.041
Hungary																0	0	0.070
Indonesia																0	0	0.059
Iran																0	0	0.035
Ireland																0	0	0.050
Israel	+	+														2	2	0.072
Italy																0	0	0.030
Lithuania																0	0	0.035
Malta					+			+								2	2	0.088
Netherlands																0	0	0.041
New Zealand	+	+														2	2	0.116
Northern Ireland	+	+	+													3	3	0.105
Norway										++						0	1	0.083
Poland	+	+	+													3	3	0.113
Portugal	+	+	+	+												4	4	0.179
Qatar																0	0	0.041
Romania																0	0	0.042

(continued)

Table A.24 (continued)

Country	Item															2.5%	5%	Mean absolute
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Outlier	Outlier	loading
Russian Fed.	+	+														2	2	0.067
Saudi Arabia																0	0	0.040
Singapore	++	++														0	2	0.055
Slovak Republic																0	0	0.043
Slovenia										++						0	1	0.063
Spain																0	0	0.044
Sweden																0	0	0.071
Trinidad and Tobago								+								1	1	0.069
United Arab Emirates																0	0	0.054
United States																0	0	0.053

Note. + indicates factor loading belongs to the 5% most positive loading, ++ indicates factor loading belongs to the 2.5% most positive loading. - indicates factor loading belongs to the 5% most negative loading, -- indicates factor loading belongs to the 2.5% most negative loading. The 2.5% cultural differential item functioning (CDIF) and 5% CDIF columns give the number of outliers in the two respective regions. Mean absolute loading refers to the means over items of the absolute values of country-specific factor loadings. The content of the items is described in Table 3.1.

Appendix B
OpenBUGS Script for the Overall Model in Chapter 5, Including Prior Specifications

```
#C is number of groups in the dataset
#N is number of students per group
#N_booklet is a vector indicating the number of items per booklet
#booklet is a matrix specifying which booklet a student made
#TIMSS15 is the data matrix with response data on the mathematics items (coded 1, 2,..)
#K is the number of scoring categories per item
#Itm.Bklt is a matrix specifying for each booklet which items it contains
#LG refers to language group student belongs to (0=LG1, 1=LG2)
#LanguageDemand is a matrix indicating the classification of reading demand in the item using dummy coding.
#P is the number of items

model{
    for (c in 1:C){
        for (i in 1:N[c]){
            for (j in 1:N_booklet[booklet[c,i]]){
                TIMSS15[c,i,j] ~ dcat( prob[c, i, j, 1:K[ Itm.Bklt[ booklet[c,i], j]]])
            }
        }
    }
    for (c in 1:C) {
        for (i in 1:N[c]) {
            for (j in 1:N_booklet[booklet[c,i]]){
                for (k in 1:K[Itm.Bklt[booklet[c,i],j]]){
                    change[c,i,j,k] <- (k-1)*( gamma[1] * LG[c,i] *
LanguageDemand[Itm.Bklt[booklet[c,i],j],1] +
        gamma[2] * LG[c,i] *
LanguageDemand[Itm.Bklt[booklet[c,i],j],2] )
    difficulty[c,i,j,k] <- beta[ Itm.Bklt[booklet[c,i],j]  ,k]
proficiency[c,i,j,k] <- (k-1)*alpha[Itm.Bklt[booklet[c,i],j]] * theta[c,i] +
(k-1)* alphac[Itm.Bklt[booklet[c,i],j],c] *
```

```
                                thetac[c,i]
                        eta[c,i,j,k] <- exp( proficiency[c,i,j,k] - difficulty[c,i,j,k] -
change[c,i,j,k])
                }
                psum[c,i,j] <- sum(eta[c,i,j,1:K[Itm.Bklt[booklet[c,i],j]]])
                for (k in 1:K[Itm.Bklt[booklet[c,i],j]]){ prob[c,i,j,k] <- eta[c,i,j,k] /
psum[c,i,j]
                }
            }
                theta[c,i] ~ dnorm( mu_theta[c], tau_theta[c] )
                thetac[c,i] ~ dnorm(0, 1)
        }
    }
    mu_theta[4] <- 0
    tau_theta[4] <- 1

    for (c in 1:3) {
            mu_theta[c] ~ dnorm(0,1)
            tau_theta[c] ~ dgamma(1,1)
            var_theta[c] <- 1/tau_theta[c]
    }

    for (c in 1:C) {
        for (j in 1:P){
            alphac[j,c] ~ dnorm(1,1)I(0,)
        }
    }
    for (j in 1:P){
        alpha[j]~dnorm(1,1)I(0,)
        beta[j,1] <- 0.0
        for(k in 2:K[j]){
            beta[j,k]~dnorm(0,1)
        }
    }
   gamma[1] ~ dnorm(0,1)
   gamma[2] ~ dnorm(0,1)
}
```

Appendix C
OpenBUGS Script for Model 8 in Chapter 6, Including Prior Specifications

```
#C is number of countries
#N is number of students per country
#N_bookletT is vector indicating number of items per booklet in TIMSS
#bookletT is a matrix specifying which TIMSS booklet student made
#TIMSS2011 is the data matrix with response data on the mathematics items (coded 1,2,…)
#K is number of response categories for the TIMSS items
#Itm.BkltT is matrix specifying for each TIMSS booklet which items it contains
#ReadingDemand is a matrix indicating the classification of reading demand in the items using dummy coding.
#LG refers to language group student belongs to (0=LG1, 1=LG2)
#N_bookletP is vector indicating number of items per booklet in PIRLS
#bookletP is a matrix specifying which PIRLS booklet student made
#H is number of response categories for the PIRLS items
#Itm.BkltP is matrix specifying for each PIRLS booklet which items it contains
#n.t is the total number of testlets in the PIRLS test
#T is number of items in TIMSS11 mathematics test
#P is number of items in PIRLS11 reading literacy test

model{
#Response model on TIMSS-2011
        for(c in 1:C){
                for (i in 1:N[c]){
                        for(j in 1:N_bookletT[bookletT[c,i]]){
                                TIMSS2011[c,i,j] ~
                                dcat(probT[c,i,j,1:K[Itm.BkltT[bookletT[c,i],j]]])
                        }
                }
        }

        for (c in 1:C){
                for (i in 1:N[c]){
```

```
                    for (j in 1:N_bookletT[bookletT[c,i]]){
                        for (k in 1:K[Itm.BkltT[bookletT[c,i],j]]){
                        change[c,i,j,k] <- (k-1)*( delta[1] *
                        ReadingDemand[Itm.BkltT[bookletT[c,i],j],1]*LG[c,i] +
                        delta[2] *
                        ReadingDemand[Itm.BkltT[bookletT[c,i],j],2]*LG[c,i] )
                        difficulty[c,i,j,k] <- betaT[ Itm.BkltT[bookletT[c,i],j], k]
            ability[c,i,j,k] <- (k-1)*alphaT[Itm.BkltT[bookletT[c,i],j]] * theta[c,i,1]
            etaT[c,i,j,k] <- exp( ability[c,i,j,k] - difficulty[c,i,j,k] - change[c,i,j,k])
                        }
                        psumT[c,i,j] <-
sum(etaT[c,i,j,1:K[Itm.BkltT[bookletT[c,i],j]]])
                        for (k in 1:K[Itm.BkltT[bookletT[c,i],j]]){
                            probT[c,i,j,k] <- etaT[c,i,j,k] / psumT[c,i,j]
                        }
                    }
                }
        }

#Response model on PIRLS-2011
        for(c in 1:C){
            for (i in 1:N[c]){
                for(j in 1:N_bookletP[bookletP[c,i]]){
                    PIRLS2011[c,i,j] ~
dcat(probP[c,i,j,1:H[Itm.BkltP[bookletP[c,i],j]]])
                }
            }
        }

        for(c in 1:C){
            for (i in 1:N[c]){
                for(j in 1:N_bookletP[bookletP[c,i]]){
                    for(k in 1:H[Itm.BkltP[bookletP[c,i],j]]){
                        etaP[c,i,j,k] <- exp((k-1) *
                        alphaP[Itm.BkltP[bookletP[c,i],j]] * (theta[c,i,2]
                          + gamma[c,i,d[Itm.BkltP[bookletP[c,i],j]]] -
                        betaP[Itm.BkltP[bookletP[c,i],j],k] ))
                    }
```

```
                        psumP[c,i,j] <-
                        sum(etaP[c,i,j,1:H[Itm.BkltP[bookletP[c,i],j]]])
                        for(k in 1:H[Itm.BkltP[bookletP[c,i],j]]){

                            probP[c,i,j,k] <- etaP[c,i,j,k] / psumP[c,i,j]
                        }
                    }
                    for(t in 1:n.t){gamma[c,i,t] ~ dnorm(0.0, pr.gamma[t])
                    }
                }
        }

#Specification of latent ability structure
for (i in 1:N[1]){
        theta[1,i,2] ~ dnorm(0.0,1.0) }
for(c in 2:C){
        for (i in 1:N[c]){
            theta[c,i,2] ~ dnorm(mu[c],tau_pc[c]) }}
for(c in 1:C){
        for (i in 1:N[c]){
           exp_theta[c,i,1] <- u_0[c] + gamma_01*theta[c,i,2] + gamma_02*LG[c,i]
            theta[c,i,1] ~ dnorm(exp_theta[c,i,1],tau)
        }
}

#Priors for population model
gamma_01 ~ dnorm(0,1)
gamma_02 ~ dnorm(0,1)
tau ~ dgamma(25,5)
var_e <- 1/tau
for (c in 2 : C) { u_0[c] ~ dnorm(0,1)
                                mu[c] ~ dnorm(0,1)
                                tau_pc[c] ~ dgamma(1,1)
                                var_pc_[c] <- 1/tau_pc[c]
                                }
   u_0[1] <- 0.0
        mu[1] <- 0.0

#Priors on item parameters in TIMSS items
```

```
        for (j in 1:T){
                alphaT[j] ~ dnorm(1,1)I(0,)
                betaT[j,1] <- 0.0
                for(k in 2:K[j]){
                        betaT[j,k] ~ dnorm(0,1)
                }
        }


#Priors on item parameters in PIRLS items
        for(j in 1:P){
        alphaP[j] ~ dnorm(1,1)I(0,)
        betaP[j,1] <- 0
                for(k in 2:H[j]){
                        betaP[j,k] ~ dnorm(0,1)
                }
        }

        for(t in 1:n.t){
                pr.gamma[t] ~ dgamma(40,16)
                sigsq.gamma[t] <- 1.0/pr.gamma[t]
        }

delta[1] ~ dnorm(0,1)
delta[2] ~ dnorm(0,1)


}
```

References

Abedi, J., & Lord, C. (2010). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*(1), 67-91.

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Adams, R. J., & Wu, M. (2006). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57–75). New York: Springer.

Aesaert, K., & Van Braak J. (2015). Gender and socioeconomic related differences in performance based ICT competences. *Computers and Education, 84,* 8-25.

Aitchison, J., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29*, 813–828.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *17*, 251– 269.

Andon, A., Thompson, C. G., & Becker, B. J. (2012). A quantitative synthesis of the immigrant achievement gap across OECD countries. *Large-Scale Assessments in Education, 2*(7), 1–20.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Erlbaum

Banks, K., Jeddeeni, A., & Walker, C. M. (2016). Assessing the effect of language demand in bundles of math word problems. *International Journal of Testing, 16*(4), 269-287.

Barwell, R. (2012). Discursive demands and equity in second language mathematics classrooms. In B. Herbel-Eisenmann, J. Choppin, D. Wagner, & D. Pimm (Eds.), *Equity in discourse for mathematics education: Theories, practices, and policies* (pp. 147–164). Dordrecht: Springer.

BECTA, British Educational Communications and Technology Agency. (2008, August). *How do boys and girls differ in their use of ICT?* Retrieved from http://dera.ioe.ac.uk/8318/1/gender_ict_briefing.pdf

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*(4), 541-561.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika, 46*, 443–459.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12*(3), 261-280.

Charlton, J. P. (1999). Biological sex, sex-role identity, and the spectrum of computing orientations: a re-appraisal at the end of the 90s. *Journal of Educational Computing Research, 21*(4), 393-412.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cooper, J. (2006). The digital divide: the special case of gender. *Journal of Computer Assisted Learning, 22*(5), 320-334.

De Boeck P., & Wilson M. (eds). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

De Jong, M. G., Steenkamp, J. B. E. M., & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34*(2), 260–278.

Drent, M., Meelissen, M. R. M., & Van der Kleij, F. M. (2013). The contribution of TIMSS to the link between school and classroom factors and student achievement. *Journal of Curriculum Studies, 45*(2), 198-224.

Ercikan, K., Chen, M. Y., Lyons-Thomas, J., Goodrich, S., Sandilands, D., Roth, W. M., & Simon, M. (2015). Reading proficiency and comparability of mathematics and science scores for students from English and non-English backgrounds: An international perspective. *International Journal of Testing, 15*(2), 153-175.

Ertl, B., & Helling, K. (2011). Promoting gender equality in digital literacy. *Journal of Educational Computing Research, 45*(4), 477-503.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359-374.

Fox, J-P. (2010). *Bayesian item response modeling.* New York: Springer.

Fox, J-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271-288.

Foy, P., & Drucker, K. T. (Eds.). (2013). *PIRLS 2011 user guide for the international database*. Chestnut Hill, MA, USA: TIMSS & PIRLS International Study Center, Boston College.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age: The IEA international computer and information literacy study international report.* Cham: Springer.

Fraillon, J., Schulz, W., & Ainley, J. (2013). *International computer and information literacy study: assessment framework.* Amsterdam: IEA.

Fraillon, J., Schulz, W., Friedman, T., Ainley, J., & Gebhardt, E. (2015). *ICILS 2013 technical report.* Amsterdam: IEA.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423–436.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8*(3), 647-667.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64*(3), 273–294.

Glas, C. A. W. (2010). *Multidimensional item response theory (MIRT), manual and computer program.* Retrieved from https://www.utwente.nl/bms/omd/medewerkers/temp_test/mirt_package.zip and: https://www.utwente.nl/bms/omd/medewerkers/temp_test/mirt-manual.pdf

Glas, C. A. W. (2017). LEXTER: manual and computer program. Retrieved from: https://www.utwente.nl/nl/bms/omd/Medewerkers/temp_test/Lexter-package.zip and https://www.utwente.nl/nl/bms/omd/Medewerkers/temp_test/lexter-manual.pdf.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika, 72*(2), 159–180.

Glas, C. A. W., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*(2), 87–106.

Glas, C. A. W., & Jehangir, K. (2014). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 97-115). London: Chapman & Hall/CRC Press.

Gui, M. & Argentin, G. (2011). Digital skills of internet natives: different forms of digital literacy in a random sample of northern Italian high school students. *New Media and Society, 13*(6), 963-980.

Haag, N., Heppt, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners' performance in mathematics: disentangling the effects of academic languages. *Learning and Instruction, 28*, 24-34.

Hargittai, E., & Shafer, S. (2006). Differences in actual and perceived online skills: the role of gender. *Social Science Quarterly, 87*(2), 432-448.

Hatlevik, O. E., & Christophersen, K. A. (2013). Digital competence at the beginning of upper secondary school: identifying factors explaining digital inclusion. *Computers and Education, 63*, 240-247.

Hill, N.E., Castellino, D.R., Lansford, J.E., Nowlin, P., Dodge, K.A., Bates, J.E., & Pettit, G.S. (2004). Parent academic involvement as related to school behavior, achievement, and aspirations: Demographic variations across adolescence. *Child Development, 75*, 1491-1509.

Hohlfeld, T. N., Ritzhaupt, A. D., & Barron, A. E. (2013). Are gender differences in perceived and demonstrated technology literacy significant? It depends on the model. *Educational Technology Research and Development, 61*(4), 639-663.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* New York, NJ: Routledge.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J. & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research, 62*(3), 333-353.

Husén, T. (Ed.). (1967). *International study of achievement in mathematics: A comparison of twelve countries (volume II).* Stockholm: Almqvist & Wiksell.

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

IBM Corporation. (2013). *IBM SPSS Statistics* (version 22.0). Somers, NY: IBM Corporation.

IERI (n.d.). Training area. Retrieved from http://www.ierinstitute.org/training-area.html.

Ilomäki, L., Paavola, S., Lakkala, M., & Kantosala, A. (2016). Digital competence – an emergent boundary concept for policy and educational research. *Education and Information Technologies, 21*(3), 655-679.

International Association for the Evaluation of Educational Achievement. (2017). *IDB analyzer* (version 4.0).Hamburg, Germany: IEA Hamburg. Available at http://www.iea.nl/data.html.

Janssen Reinen, I. A. M. J., & Plomp, T. (1993). Gender and computers: another area of inequity in education? In Pelgrum, W. J., Janssen Reinen, I. A. M. J., & Plomp, T. (Eds.), *Schools, teachers, students and computers: a cross-national perspective* (pp. 91-116). The Hague: IEA.

Jehangir, K. (2015). *Methodological issues in large-scale educational surveys*. Enschede: Universiteit Twente.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76*(4), 537–549.

Johansson, S. (2016). International large-scale assessments: What uses, what consequences? *Educational Research, 58*(2), 139-148.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 702-714.

Kreiner, S. & Christensen, K.B. (2014). Analyses of model fit and robustness: A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210–231.

Kuhlemeier, H., & Hemker, B. (2007). The impact of computer use at home on students' Internet skills. *Computers and Education 49*(2), 460-480.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Lau, W. W. F., & Yuen, A. H. K. (2015). Factorial invariance across gender of a perceived ICT literacy scale. *Learning and Individual Differences, 41*, 79-85.

Lenkeit, J., Chan, J., Hopfenbeck, T. N., Baird, J.-A. (2015). A review of the representation of PIRLS related research in scientific journals. *Educational Research Review, 16*, 102-115.

Lin, M., Bumgarner, E., Chatterji, M. (2014). Understanding validity issues in international large scale assessments. *Quality Assurance in Education, 22*(1), 31-41.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine, 28*, 3049-3067.

Lynn, R., & Mikk, J. (2009). *National IQs predict educational attainment in math, reading and science across 56 nations. Intelligence*, *37*(3), 305-310.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA, USA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2013). *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade – implications for early learning.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14*, 160-179.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Meelissen, M. (2008). Computer attitudes and competencies among primary and secondary school students. In J. Voogt & G. Knezek (Eds.) *International handbook of information technology in primary and secondary education* (pp. 381-395). Boston, MA: Springer US.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *TIMSS 2015 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2019/frameworks/.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *ePIRLS 2016 international results in online informational reading*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/pirls2016/international-results/.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

OECD (2010). *PISA 2009 results: What students know and can do – student performance in reading, mathematics and science (Volume I)*. Paris: OECD Publishing.

OECD (2011). *PISA 2009 results: Students on line – digital technologies and performance (Volume VI)*. Paris: OECD Publishing.

OECD (2017). *PISA 2015 technical report*. Retrieved from OECD website: http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf.

Platz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

Punter R. A., Glas C. A. W., Meelissen M. R. M. (2016a). Literature review. In *Psychometric framework for modeling parental involvement and reading literacy (*pp. 5-23). Cham: Springer.

Punter R. A., Glas C. A. W., Meelissen M. R. M. (2016b). Appendix A: Technical details on the implementation of the bi-factor model. In *Psychometric framework for modeling parental involvement and reading literacy (*pp. 95-97). Cham: Springer.

Rao, C. R. (1947). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44*, 50 –57.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371.

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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 34*(4, Pt. 2), 1–100.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, 64(*4), 583-639.

Spring, J. (2008). Research on globalization and education. *Review of Educational Research, 78*(2), 330-363.

Tømte, C. (2011). Challenging our views on ICT, gender and education. *Nordic Journal of Digital Literacy, 6*, 309-324.

Tsai, M. J., & Tsai, C. C. (2010). Junior high school students' Internet usage and self-efficacy: a re-examination of the gender gap. *Computers and Education, 54*(4), 1182-1192.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43*, 39 –55.

Van der Linden, W. J. (Ed.). (2016). *Handbook of item response theory*. Boca Raton, FL: Chapman and Hall/CRC

Van Deursen, A. J. A. M., & Van Diepen, S. (2013). Information and strategic Internet skills of secondary students: a performance test. *Computers & Education, 63*, 218-226.

Verhelst, N. D., Glas, C. A. W., & De Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory*, 123-138. New York, NJ: Springer.

Volman M. (1994). *Computerfreak of computervrees: Sekseverschillen en egalitair informatiekunde-onderwijs.* [Computer freak or computer fright: Gender differences and equality in information and computer literacy education]. Amsterdam: University of Amsterdam.

Volman M. (1997). Gender-related effects of computer and information literacy education. *Journal of Curriculum Studies, 29*(3), 315-328.

Volman, M., & Van Eck, E. (2001). Gender equity and information technology in education: the second decade. *Review of Educational Research, 71*(4), 613-634.

Von Davier, M, Gonzalez, E, & Mislevy, R. J. (2009). What are plausible values and why are they useful? In M. Von Davier & D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large-scale assessments* (pp. 9–36). Hamburg: IER Institute.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: a cognitive explanation for DIF. *Journal of Educational Measurement, 38*(2), 147-163.

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education, 21*(2), 162-181.

Wendt, H., Bos, W., & Goy, M. (2011). On applications of Rasch models in international comparative large-scale assessments: A historic review. *Educational Research and Evaluation, 17*(6), 419-446.

Wong, K. C. K., & Cheung, W. K. W. (2012). A study of gender differences in ICT competency. In S. Fong, K. D. Kwack, & F. Co (Eds.), *International Conference on Information Science and Digital Content Technology* (ICIDT), Jeju Island, South Korea, 26–28 June 2012, pp.12–14.

Yamamoto, K., & Kulick, E. (2000). Scaling methodology and procedures for the TIMSS mathematics and science scales. In M.O. Martin, D.G. Kelvin, S.E. Stemler (Eds.), *TIMSS 1999. Technical report* (pp. 237-263). Chestnut Hull, MA: Boston College.

Summary

International large-scale assessments (ILSAs) play a major role in the evaluation of educational systems. These projects are characterized by the standardized assessment of student achievement and the collection of contextual data by means of curriculum, student, teacher, school, and home questionnaires. Together, the resulting high-quality data on student achievement and contextual factors provide great opportunities for more theory-oriented educational effectiveness research, particularly in international contexts.

To ensure the validity of analyses based on these data, particularly relating to measurement invariance across (sub)populations, efforts must be made to evaluate response behaviour across (sub)populations of interest. Potential differences in item response behaviour for respondents of equal ability can complicate inferences regarding proficiency differences between countries or specific subpopulations. A lack of measurement invariance, characterized by these differences in response behaviour is called differential item functioning (DIF). DIF can be present in both test and questionnaire data. Studies in this thesis intended to contribute to research in the field of education by deploying ILSA data in research areas where the availability of standardized data from multiple countries offers new research opportunities, and to explore methodologies for identifying and handling potential DIF in the framework of item response theory (IRT).

The first study in this thesis utilized data from the International Computer and Information Literacy Study (ICILS) 2013. The international results were reported based on a single, unidimensional scale and showed that girls outperformed boys. However, computer and information literacy is complex and potentially multidimensional. This study explored whether the achievement test in ICILS addressed specific dimensions of computer and information literacy and if so, whether the performance on these subscales differed between boys and girls. Also, two methods of investigation were applied to evaluate potential gender-related item bias and bias across countries: one was based on the differences between observed mean item scores in the gender and country groups and their expected values under the proposed model; the other method was based on comparing parameter estimates obtained for the subgroups.

Results show that the proposed three-dimensional model, with the clearest distinction between a computer literacy dimension and two information-oriented dimensions, indeed fits the data better. The postulated hypothesis that girls have a performance advantage on items in the more information-oriented dimensions and that for performance on items assessing computer literacy the advantage is reversed or even non-existent is supported. The investigation into potential DIF did not show substantial evidence for item bias: results support the use of one model across the nine countries and confirm the validity of the group comparisons based on this model.

The thesis continues with research on parental involvement. Although parental involvement is seen as a malleable factor in the student's home situation, which can enhance student achievement, evidence regarding these effects is not univocal. The inconsistent results may be caused by differences between educational systems and cultural differences, or by the great variation in the methods used to assess student achievement and the multifaceted parental involvement construct across studies. This research started by constructing five parental involvement scales from the Progress in International Reading and Literacy Study (PIRLS) 2011 questionnaire data across the 43 participating countries. The scales are modelled using the partial credit model (Masters, 1982) and the generalized partial credit model (GPCM, Muraki, 1992). Several ways to identify and model cultural DIF were explored: (1) the GPCM with random item parameters, (2) residual modelling using a Lagrange multiplier test statistic, and (3) a bi-factor approach. The latter is an extension of the bi-factor model (Gibbons & Hedeker, 1992) with a structure characterized by a main component across all countries, on which consequential comparisons can be made, and a country-specific, secondary component. All models clearly and consistently support the identification of cultural DIF (CDIF). However, results do vary over the methods as the first two methods focus on identification of uniform CDIF and the bi-factor application on non-uniform CDIF.

The research continues by exploring the relationship between different dimensions of parental involvement and student reading literacy across all participating countries in PIRLS-2011. Reading literacy is then regressed in latent multilevel models on the previously created parental involvement scales. Scores based on scaling both with and without the corrections for country-item interactions resulting from the first research objective are used. Results show that early literacy activities and help with homework

are both related to student achievement, but the observed impact of parental involvement on reading literacy was limited. Impact differences across countries do prove to be quite large, especially for helping with homework, where the effect of this dimension appears to be stronger in high-achieving countries. The comparisons between analyses with and without corrections for cultural differential item functioning in the parental involvement scales indicate that CDIF does not influence the inferences.

The final two studies in the thesis focus on the mathematics items in Trends in International Mathematics and Science Study (TIMSS) for second language learners (SLLs) and the role of item reading demand. The TIMSS test requires a basic level of reading ability, which may hinder students who do not speak the language of the test at home. The first of these studies presents an IRT model that combines multiple approaches to detect and model DIF in ILSAs. Two generalizations of the GPCM are described, which combine into the overall model. The first generalization is a bi-factor application related to one previously mentioned. The second generalization is a model where item parameters are regressed on student and item characteristics. The modelling steps are illustrated on data from the TIMSS-2015 mathematics test from four European countries, with reading demand classifications as the item properties of interest for students not speaking the test language at home. Results show that the full model fits the data best. This indicates that reading demand should be accounted for when students do not speak the test language at home.

The last study used data from the combined administration of TIMSS-2011 and PIRLS-2011 to further investigate the functioning of TIMSS mathematics items for SLLs, while controlling for their reading skills. This was done by estimating several latent regression models concurrently with response models on both the TIMSS and PIRLS response data. The results of this study confirm the often-reported relationship between student achievement in reading and mathematics. In addition, this enables testing the extent to which the often-observed DIF in mathematics tests for SLL students can be attributed to their potentially limited reading achievement. Results show a small but negative effect of being an SLL student, even after controlling for reading proficiency. This shows that the achievement gap between SLL and non-SLL students cannot be fully explained by their level of reading literacy. Also, a small interaction effect was found between item reading demand and student language

background, while controlling for reading proficiency. Implications of these findings for the development of future ILSAs relate to taking potential negative effects of item wording into account.

The studies in this thesis show how DIF analyses can be insightful by benefiting from the synergy between a methodological focus on validity and a focus on more substantive research questions. More than simply a task to tick off before the "real" questions are investigated (i.e., a mere check), DIF analyses can lead to insights into effects underlying test results. Throughout the studies in this thesis, it is therefore shown how, in studies with a substantive interest in comparing groups, the study of validity on both test and questionnaire items should be integrated into the methodology. Though no clear-cut one-method-fits-all strategy is presented here, the thesis shows that there are many ways to approach the issue.

Samenvatting

Grootschalige internationale surveys, in dit proefschrift afgekort tot ILSAs naar de Engelse benaming *international large-scale assessments*, spelen een belangrijke rol in het evalueren van onderwijssystemen. Deze onderzoeksprojecten kenmerken zich door gestandaardiseerde toetsen voor het meten van leerlingprestaties. Naast de toetsafnames wordt contextuele data verzameld met behulp van curriculum-, leerling-, leerkracht-, schoolleider- en oudervragenlijsten. De data rondom leerlingprestaties en contextuele factoren bieden tezamen veel mogelijkheden voor onderzoek op het gebied van onderwijseffectiviteit, met name in internationale context.

Dit proefschrift richt zich op een belangrijke validiteitsvraag bij internationale vergelijkingen: wordt in de verschillende (sub)populaties daadwerkelijk hetzelfde gemeten? Wanneer er verschillen zijn in de wijze waarop vragen worden beantwoord tussen respondenten met eenzelfde vaardigheidsniveau, bijvoorbeeld door een andere interpretatie van de vraag, is er sprake van *differential item functioning* (DIF). DIF kan zowel optreden in de toetsen als in de vragenlijsten. Wanneer er sprake is van DIF kan dit leiden tot vertekende resultaten bij het vergelijken van (sub)populaties.

Dit proefschrift presenteert onderzoeken die enerzijds bijdragen aan inhoudelijke kennis op het gebied van onderwijskunde. Dit door de inzet van ILSA-data bij vraagstukken waar de beschikbaarheid van gestandaardiseerde testdata uit meerdere landen nieuwe onderzoeksmogelijkheden biedt. Anderzijds verkennen de studies methodes binnen de item response theorie (IRT) voor het in kaart brengen van en het omgaan met mogelijke DIF.

Het eerste onderzoek in dit proefschrift is gebaseerd op data van een internationale survey naar computer- en informatievaardigheden (ICILS-2013). In de internationale rapportage werd er een enkele, unidimensionale schaal gerapporteerd, waarop meisjes gemiddeld hoger scoorden dan jongens. Het domein van computer- en informatievaardigheden is echter complex en wellicht multidimensionaal. In deze studie wordt gekeken of een multidimensionale representatie van de gemeten vaardigheid gerechtvaardigd is en zo ja, in welke gebieden de verschillen tussen jongens en meisjes zich dan het meest manifesteren. Daarnaast is er, ter validatie van de vergelijking tussen landen en de beide geslachten, op twee manieren onderzocht of er indicaties waren van DIF. Bij de ene methode door te kijken naar verschillen in de

waargenomen gemiddelde item scores in de subgroepen (naar land en geslacht) en hun verwachte waarde onder het model; bij de andere methode door het vergelijken van item parameterschattingen voor beide subgroepen.

Resultaten laten zien dat het voorgestelde driedimensionale model inderdaad beter past bij de data, waarbij vooral de dimensie computervaardigheid en de meer op informatievaardigheden gerichte dimensies het duidelijkst te onderscheiden zijn. De hypothese dat meisjes beter scoren op de meer informatie-georiënteerde dimensies en dat voor de computervaardigheid items dit niet het geval is – of dat ze zelfs lager scoren dan jongens – blijft overeind. Uit de analyses naar mogelijke DIF voor geslacht of land kwamen geen ernstige gevallen van DIF aan het licht: de resultaten ondersteunen het gebruik van één model voor de negen onderzochte landen en bevestigen de validiteit van de vergelijkingen tussen groepen op basis van dit model.

Het proefschrift vervolgt met onderzoek naar ouderbetrokkenheid. Hoewel breed verondersteld wordt dat ouderbetrokkenheid samengaat met betere leerprestaties, schetsen studies hiernaar geen eenduidig beeld. Mogelijk is dit te verklaren door culturele verschillen, verschillen in onderwijssystemen of door de grote variëteit in het meten van zowel de leerlingprestaties als het veelzijdige begrip ouderbetrokkenheid. In dit onderzoek is gestart met het samenstellen van vijf schalen voor ouderbetrokkenheid op basis van data uit de vragenlijstdata van 41 landen uit een internationale survey naar leesprestaties, PIRLS-2011. De schalen zijn gemodelleerd met het p*artial credit model* (Masters, 1982) en het *generalized partial credit model* (GPCM; Muraki, 1992). Om een beeld te krijgen van de mogelijke DIF over landen en hiervoor te corrigeren, wordt een aantal strategieën toegepast: (1) GPCM met stochastische item parameters, (2) toepassen van Lagrange Multiplier toetsingsgrootheid en (3) een bi-factor structuur. Deze bi-factor structuur is een uitbreiding van het bi-factor model (Gibbons & Hedeker, 1992) en kent een hoofdcomponent op basis waarvan scores tussen landen vergeleken kunnen worden en een tweede component om land-specifieke ruis te meten. Alle modellen ondersteunen duidelijk en consistent de identificatie van culturele DIF (CDIF). Wel zijn er verschillen in de resultaten van de verschillende modellen, omdat de eerste twee methodes zich richten op de identificatie van uniforme CDIF en de bi-factor toepassing op niet-uniforme CDIF.

Het onderzoek gaat verder met het verkennen van de relatie tussen de verschillende componenten van ouderbetrokkenheid en leerlingprestaties in begrijpend lezen over alle deelnemende landen in PIRLS-2011 heen. De relatie tussen leesvaardigheid en de eerder geconstrueerde ouderbetrokkenheidsschalen is onderzocht met multilevel analyses. Voor deze schalen zijn zowel scores zonder als met correcties voor CDIF gebruikt, zoals deze uit het eerste deel van het onderzoek naar voren kwamen. Resultaten laten zien dat vroege alfabetiseringsactiviteiten en ouderhulp bij huiswerk beide gerelateerd zijn aan leerlingprestaties, maar de samenhang tussen ouderbetrokkenheid en leesprestaties is beperkt. De samenhang varieert wel sterk tussen landen, met name bij ouderhulp bij huiswerk, waarvan het effect sterker blijkt in hoog-presterende landen. De vergelijkingen tussen de analyses met en zonder correcties voor CDIF in de ouderbetrokkenheidsschalen laten zien dat CDIF geen invloed heeft op de conclusies.

De laatste twee onderzoeken in het proefschrift gaan in op de wiskunde items in TIMSS: een internationale survey naar prestaties in reken- en natuuronderwijs. In deze twee onderzoeken staat de vergelijking centraal tussen leerlingen die de toetstaal thuis niet spreken (*second language learners*; SLLs) en leerlingen die dit wel doen en wat daarbij de rol van taligheid in de items is. De TIMSS-toets vereist een basale leesvaardigheid, wat wellicht lastig is voor SLL-leerlingen.

Het eerste onderzoek presenteert een IRT-model waarin meerdere benaderingen voor het detecteren en modelleren van DIF worden gecombineerd. Twee generalisaties van het GPCM worden beschreven, die samenkomen in een totaalmodel. De eerste generalisatie is een bi-factor applicatie, vergelijkbaar met de eerdergenoemde bi-factor applicatie. De tweede generalisatie betreft een model waarin item parameters afhankelijk zijn van leerling- en itemkenmerken. De modellen worden geïllustreerd aan de hand van data van leerlingen uit vier Europese landen op de TIMSS-2015 rekentoets, waarbij taligheid wordt ingezet als itemkenmerk voor leerlingen die thuis niet de toetstaal spreken. Uit de resultaten komt naar voren dat het volledige model het best bij de data past. Hieruit kan worden afgeleid dat taligheid in de rekentoets een factor is om rekening mee te houden voor leerlingen die thuis de toetstaal niet spreken.

De laatste studie in dit proefschrift gebruikt data van de gecombineerde afnames van TIMSS-2011 en PIRLS-2011 om verder te onderzoeken hoe de TIMSS-rekenitems uitpakken voor SLL-leerlingen wanneer er wordt gecontroleerd voor hun vaardigheid in begrijpend lezen. Dit is onderzocht met behulp van latente regressiemodellen die tezamen zijn geschat met de responsiemodellen op de TIMSS- en PIRLS-data. De resultaten van deze studie bevestigen in de eerste plaats de vaak gerapporteerde relatie tussen leerlingprestaties in lezen en rekenen. Daarnaast komt uit de resultaten naar voren dat er een klein maar negatief effect is op rekenprestaties voor SLL-leerlingen, ook wanneer er gecontroleerd is voor hun leesvaardigheid. Dit laat zien dat de prestatieverschillen tussen SLL- en niet-SLL-leerlingen niet volledig wordt verklaard door hun leesvaardigheidsniveau. Daarnaast is er een klein interactie-effect gevonden tussen de taligheid van een item en de taalachtergrond van leerlingen, gecontroleerd voor leesvaardigheid. Uit deze resultaten volgt de aanbeveling om bij toekomstige ILSAs rekening te houden met de mogelijke nadelige effecten van taligheid in de items.

De studies in dit proefschrift laten zien hoe DIF-analyses van meerwaarde kunnen zijn door te profiteren van synergie tussen de methodologische focus op validiteit en een focus op meer inhoudelijke vragen. Wanneer DIF-analyses goed worden ingezet – niet alleen als een actie die afgevinkt moet worden voordat het "echte" onderzoek kan beginnen – dan kunnen ze leiden tot inzichten in effecten die de onderzoeksresultaten beïnvloeden. Met de studies in dit proefschrift wordt daarom aangetoond hoe – wanneer vergelijkingen van (sub)populaties centraal staan – onderzoek naar de validiteit van zowel vragenlijst- als toetsdata kan worden geïntegreerd in de onderzoeksaanpak.