UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Mojca Rožman

Učinek sestave vzorca pri ocenjevanju parametrov postavk in dosežkov v mednarodnih raziskavah znanja

Effect of Sample Composition in the Estimation of Item Parameters and Proficiency Estimation in International Large-scale Assessments

Doktorska disertacija

Ljubljana, 2014

UNIVERZA V LJUBLJANI

FAKULTETA ZA DRUŽBENE VEDE

Mojca Rožman

Mentor: doc. dr. Gregor Sočan

Učinek sestave vzorca pri ocenjevanju parametrov postavk in dosežkov v mednarodnih raziskavah znanja

Effect of Sample Composition in the Estimation of Item Parameters and Proficiency Estimation in International Large-scale Assessments

Doktorska disertacija

Ljubljana, 2014

Acknowledgements

There are many people who contributed to this thesis. Firstly, I would sincerely like to thank my supervisor Gregor Sočan for the helpful comments and suggestions when I was encountering dead ends. Furthermore, I would like to thank the committee members Vasja Vehovar and Mojca Štraus for insightful comments and remarks provided during the process of defense. I am also grateful to Alenka Gril for the time spent on discussions, which helped me to elaborate my ideas. Many thanks to Eugenio Gonzales for the help in the initial phases of building the simulation ideas and for help when encountering technical difficulties.

Furthermore I would like to express my gratitude to Ana who helped to make parts of the confused texts more readable and understandable. In addition, many thanks to Eva who took the role of a deadline reminder and to Uroš who helped me to make the simulations work on their own.

Especially I would like to thank my parents and my brother, who always supported me in my work and encouraged me in my ideas. Above all, very special thanks to Diego and Anaiz who experienced my challenging moods especially in the final phases of this thesis and spread positive energy that made it easier for me to finish.

Univerza
v Ljubljani
Fakulteta
za družbene vede

Kardeljeva ploščad 5
1000 Ljubljana, Slovenija
telefon 01 58 05 120

I Z J A V A O A V T O R S T V U
doktorske disertacije

Podpisani/-a Mojca Rožman, z vpisno številko 74060811, sem avtor/-ica doktorske disertacije z naslovom:
Effect of Sample Composition in the estimation of Item Parameters and Proficiency Estimation in International Large-scale Assessments (Učinek sestave vzorca pri ocenjevanju parametrov postavk in dosežkov v mednarodnih raziskavah znanja).

S svojim podpisom zagotavljam, da:

- je predložena doktorska disertacija izključno rezultat mojega lastnega raziskovalnega dela;
- sem poskrbel/-a, da so dela in mnenja drugih avtorjev oz. avtoric, ki jih uporabljam v predloženem delu, navedena oz. citirana v skladu s fakultetnimi navodili;
- sem poskrbel/-a, da so vsa dela in mnenja drugih avtorjev oz. avtoric navedena v seznamu virov, ki je sestavni element predloženega dela in je zapisan v skladu s fakultetnimi navodili;
- sem pridobil/-a vsa dovoljenja za uporabo avtorskih del, ki so v celoti prenesena v predloženo delo in sem to tudi jasno zapisal/-a v predloženem delu;
- se zavedam, da je plagiatorstvo – predstavljanje tujih del, bodisi v obliki citata bodisi v obliki skoraj dobesednega parafraziranja bodisi v grafični obliki, s katerim so tuje misli oz. ideje predstavljene kot moje lastne – kaznivo po zakonu (Zakon o avtorski in sorodnih pravicah (UL RS, št. 16/07-UPB3, 68/08, 85/10 Skl.US: U-I-191/09-7, Up-916/09-16)), prekršek pa podleže tudi ukrepom Fakultete za družbene vede v skladu z njenimi pravili;
- se zavedam posledic, ki jih dokazano plagiatorstvo lahko predstavlja za predloženo delo in za moj status na Fakulteti za družbene vede;
- je elektronska oblika identična s tiskano obliko doktorske disertacije ter soglašam z objavo doktorske disertacije v zbirki »Dela FDV«.

V Ljubljani, dne 20.01.2014 Podpis avtorja/-ice: _____

Abstract

The focus of the thesis lies in investigating invariance in large-scale assessment (LSA) studies. In LSA studies, item response theory (IRT) models are used to determine the achievements of students. If the IRT model fits the data, trait level estimates with invariant meaning may be obtained from any set of items, and item parameters do not depend on sample characteristics. Parameter invariance is an ideal state and is violated if any of the item parameter estimates fail to be identical (up to the same linear transformation) across different examinee populations or measurement conditions.

In general, the number of participating countries in international LSA studies increases in each data collection cycle. Since the number of countries differs from cycle to cycle, there are also different common countries in sequential cycles used for item parameter estimation. For this reason and because a model never perfectly fits the data, questions about the effect of the composition of the sample used to calculate the item parameters arise. Would the same achievement score estimates be obtained if different countries were included in item parameter estimation? Theoretically, one country in LSA studies would give sufficient information to estimate item parameters. This is true provided the population covers the range of abilities; however, the uncertainty will differ because there would be differing amounts of information at different points of the distribution.

The participation of countries in international studies is mostly increasing and voluntary. By manipulating conditions to obtain the calibration sample, we attempted to determine whether the countries (with their characteristics) that participate have an effect on the methodology used to scale the achievement results for students in various countries. In general, four research questions were investigated (with a single country as a unit in the sample): firstly, the sample size of the calibration sample; secondly, the ability of the calibration sample; thirdly, the model used in the item calibrations; and finally, the content domain assessed (either mathematics or reading).

Data from Progress in International Reading Literacy Study (PIRLS) 2006 and Trends in International Mathematics and Science Study (TIMSS) 2007 were used and rescaled under different conditions. In both studies, student achievement is measured by administering objective tests. To assess students' knowledge, a complex rotated booklet design is used, and individual students respond only to a subset of items. This is done to achieve a broader content coverage in limited testing time, but also poses challenges for generating individual achievement estimates. For estimating the achievement scores, a combination of scaling with IRT and multiple imputation is used (multiple imputations are in LSA referred to as plausible values). With this approach, however, the advantage of estimating population characteristics is more efficiently offset by the inability to make precise statements about individuals.

Reference item parameters were obtained from full data sets when all countries were included in such a way that each country contributed equally to parameter estimation. This served as a reference to which all other obtained results were compared. In practice, five plausible values were reported for each student.

In order to address the research questions, we estimated new item parameters based on including different sets of countries in the calibration sample (the sample used to calibrate the items). Firstly, we included a different number of countries. The number varied from two to ten of the participating countries, since we assumed that after a certain number of

countries the item parameter estimates would not change dramatically. Countries were selected at random and the procedure was repeated several times to obtain information about the variation of the results.

The results show that in comparing the achievement scores across conditions, significantly lower differences to the reference scores were observed in the ten-country condition compared to other conditions. The general conclusion from this empirical evidence is that including more countries (in our case ten) is a desirable recommendation for a calibration sample in LSA studies, in order that the obtained results be invariant. In other words, the average achievement of countries shows negligible differences to the reference scores when more countries were included in the calibration sample. When ten countries were included in the item parameter estimations, we obtained remarkably similar item parameters as well as similar achievement scores in comparison to the reference.

To address the second research question, two different ability samples were used in the item calibration process. Estimation of item parameters was based on the inclusion of countries regarding their mean achievement; to determine country achievement, the achievement score estimates from the reference condition were used. We sorted the countries regarding their achievement (from the country with the highest achievement to the country with the lowest achievement) and then selected the upper third of countries and the lower third of countries. We then repeated the item parameter estimation for each condition (upper and lower set of countries), each time selecting ten of the 15 countries in the respective group. In this respect, a certain number of replicates were obtained for each condition. The new results were again compared against those obtained by inclusion of all countries.

From the results of including different ability samples in item calibration, we can conclude the following. The lower achieving countries seemed to be the more efficient calibration sample in use when observing achievement scores, including across subgroups. The differences based on higher achieving countries were in most categories at least three times higher than the differences based on lower achieving countries.

The next research question dealt with invariance across different models. The same procedure was repeated using different models (Rasch family models vs. three- (3PL), two-parameter logistic (2PL), and generalized partial credit models). The results of comparing different models in item calibration show that there are no substantial differences in the model used in achievement scores of countries when their average achievement score is above a certain number of points. As soon as a country's achievement is smaller, there are greater differences between models. Rasch models provide highly invariant item and consequently person parameters. In the middle range of achievements, the results based on Rasch models in comparison to 3PL, 2PL and generalized partial credit models provide similar results. In the lower and higher points of achievement distributions, the differences are greater and consistent.

In the last research question, invariance was observed across content domains. From different content domains, we selected reading and mathematics, because we considered these two skills to be basic and also more comparable across countries (science is taught as different subjects in different countries).

In observing different content domains, smaller invariance was expected in reading domain in comparison with mathematics. The results showed that in the domain of mathematics the differences to the reference were almost doubled compared to differences found in reading.

Although the differences in absolute values were extremely small, and in mathematics they were on average one score point (on a scale with mean 500 and standard deviation of 100), the effect size was important.

With this study, we obtained applied evidence of invariance of item parameter estimates and achievement score estimates in international LSA studies. The absolute differences in achievement scores compared to the reference across conditions were found to be small. Nevertheless, some surprising findings were made, and practical suggestions can be given based on the obtained results. This research contributes to a better understanding of the property of invariance in IRT models in real data, especially valuable for international LSA studies and other studies using IRT models.

Keywords: item response theory, parameter invariance, PIRLS, TIMSS

Povzetek

V doktorskem delu smo se osredotočili na opazovanje invariantnosti parametrov postavk in dosežkov učencev v mednarodnih raziskavah znanja. Za izračun dosežkov v mednarodnih raziskavah znanja uporabljajo modele teorije odgovora na postavko (TOP). Če se model prilega podatkom, je za modele TOP značilno, da so ocene lastnosti neodvisne od postavk na katerih so bile izračunane in parametri postavk so neodvisni od lastnosti vzorca na katerem so bili pridobljeni. Invariantnost parametrov predstavlja idealno stanje in ji ne moremo zadostiti, če katerakoli ocena parametra ni identična (oziroma linearno povezana) v različnih populacijah in merskih pogojih.

Število sodelujočih držav se v mednarodnih raziskavah znanja spreminja, v vsaki novi ponovitvi raziskave se število sodelujočih držav večinoma povečuje. V ocenjevanje parametrov postavk so vključene samo države, ki so sodelovale v zaporednih ponovitvah in ki s kakovostjo izvedbe raziskave sledijo mednarodnim postopkom. Glede na to, da v zaporednih izvedbah sodelujejo različne države, so v ocenjevanje parametrov postavk vključene vsakič druge države. Zaradi tega se zastavlja vprašanje invariantnosti dosežkov učencev in parametrov postavk. Bi bili izsledki oziroma zaključki enaki, če bi bile vključene druge države? Teoretično bi lahko ena država nudila dovolj informacij za oceno parametrov postavk. Vendar to velja le v primeru, če bi porazdelitev dosežkov učencev znotraj te države pokrivala celoten možen razpon dosežkov. Hkrati bi se negotovost ocenjenih dosežkov razlikovala glede na to, ali bi imeli več ali manj informacij o različnih točkah v porazdelitvi.

Sodelovanje držav v mednarodnih raziskavah znanja se povečuje in je prostovoljno. S spreminjanjem sestave kalibracijskega vzorca (vzorec na katerem se umerijo postavke) smo skušali ugotoviti, ali imajo države, ki sodelujejo (s svojimi lastnostmi), kakšen učinek na metodologijo, ki se uporablja za lestvičenje dosežkov v sodelujočih državah. Na splošno smo opazovali ocene dosežkov in parametrov postavk glede na štiri dejavnike (pri vseh je enoto v vzorcu predstavljala država): velikost vzorca; stopnja sposobnosti v vzorcu; razlike med modeli uporabljenimi za ocenjevanje parametrov postavk, in na zadnje še med različnimi vsebinskimi področji (na področju matematike in branja).

Uporabili smo podatke dveh mednarodnih raziskav, Mednarodne raziskave bralne pismenosti (PIRLS) 2006 in Mednarodne raziskave trendov znanja matematike in naravoslovja (TIMSS) 2007 ter ponovno izračunali dosežke učencev glede na različne pogoje. Znanje učencev v teh dveh raziskavah merijo s pomočjo objektivnih preizkusov znanja. Da bi ocenili znanje učencev, uporabljajo metodo matričnega razvrščanja nalog, kjer vsak učenec reši le določen nabor postavk (nalog). Postavk je namreč veliko več kot jih lahko v razumnem času reši posamezni učenec. Ker vsak učenec ne rešuje vseh postavk, pri vsakem učencu manjkajo podatki za določene postavke. Za izračun dosežkov se uporablja kombinacija lestvičenja s TOP ter metodologijo večkratnega vstavljanja (angl. multiple imputation; vrednosti večkratnega vstavljanja se v mednarodnih raziskavah znanja imenujejo verjetnostne vrednosti, angl. plausible values). S to metodologijo zanesljivo ocenjevanje dosežka posameznika ni možno, so pa ocene dosežkov zanesljive za posamezne skupine udeležencev.

Referenčne parametre postavk smo dobili tako, da smo v ocenjevanje vključili vse sodelujoče države in sicer na način, da je vsaka država enako prispevala k parametrom postavk. Dobljeni rezultati so nam služili kot osnova za primerjavo na novo izračunanih parametrov postavk in dosežkov učencev.

V iskanju odgovora na prvo raziskovalno vprašanje smo v ocenjevanje vključili različno število držav. Posamezni pogoji so se razlikovali glede na različno število vključenih držav v kalibracijski vzorec. In sicer smo se odločili za 2, 3, 4, 6 in 10 držav (saj smo predpostavljali, da se po določenem številu vključenih držav ocenjeni parametri postavk več ne bodo spreminjali). Države smo izbrali naključno in postopek znotraj vsakega pogoja večkrat ponovili ter tako pridobili informacijo o variabilnosti ocen.

Rezultati kažejo, da so absolutne razlike v dosežkih učencev v primerjavi z referenčnimi mnogo manjše, če v kalibracijski vzorec vključimo večje število držav. Ugotovili smo, da so te razlike najmanjše v pogoju z 10 državami v primerjavi z ostalimi pogoji. V splošnem lahko ugotovimo, da velikost vzorca ni tako pomemben dejavnik pri izračunu dosežkov učencev, če je vzorec držav dovolj velik. Zato je potrebno v izračun dosežkov vključiti čim večje število držav (ali vsaj deset kot kažejo naši rezultati). V tem primeru namreč ne ugotovimo bistvenih razlik v primerjavi z referenčnimi dosežki držav. Podobno je z ocenami parametrov postavk, ki so prav tako v pogoju z vključenimi 10 državami zelo podobni referenčnim.

V naslednjem koraku smo ocenjevali parametre postavk glede na vključevanje držav po njihovem povprečnem dosežku. Države smo najprej razvrstili po povprečnem dosežku (od države z najvišjim dosežkom do države z najnižjim dosežkom) ter nato za primerjavo izbrali zgornjo tretjino držav in spodnjo tretjino držav. Nato smo ponovno ocenili parametre postavk na podvzorcu vsake skupine držav (višji in nižji dosežek) in te parametre uporabili za izračun dosežkov vseh sodelujočih držav. Ponovno smo na novo izračunane oziroma ocenjene parametre postavk ter dosežke primerjali z referenčnimi.

Iz rezultatov dobljenih na podlagi povprečnega dosežka držav vključenih v kalibracijski vzorec lahko zaključimo naslednje. Izkazalo se je, da na podlagi držav z nižjim dosežkom učinkoviteje ocenimo dosežek vseh držav. Razlike v dosežkih izračunanih na podlagi držav z višjim dosežkov in referenčnimi dosežki so vsaj trikrat večje kot tiste dobljene na podlagi držav z nižjim dosežkom.

Naslednje raziskovalno vprašanje obravnava invariantnost parametrov postavk in dosežkov učencev glede na uporabo različnih modelov TOP. Postopek izračuna dosežkov smo ponovili s pomočjo uporabe različnih modelov (modeli iz Rascheve družine nasproti logističnim modelom z dvema in tremi parametri, vključno s posplošenim modelom z delnim točkovanjem). Rezultati ne kažejo večjih razlik med modeli, če imajo države povprečni dosežek višji od določene meje. Večje razlike med modeli se pojavijo v primeru držav z nižjimi dosežki. Prav tako kažejo rezultati dobljeni z Raschevimi modeli zelo malo variabilnosti oziroma so se izkazali za zelo invariantne v primerjavi z logističnimi modeli. Razlike med uporabljenimi modeli so na spodnjem in zgornjem delu porazdelitve dosežkov večje in hkrati stabilne.

Opazovali smo še invariantnost ocenjenih parametrov postavk ter dosežkov na različnih vsebinskih področjih (matematika in bralna pismenost). Izmed različnih vsebinskih področij smo izbrali branje in matematiko, saj ti dve spretnosti smatramo za osnovni in sta tudi bolj primerljivi med državami (naravoslovje namreč v različnih državah poučujejo pri različnih predmetih).

Pri opazovanju različnih vsebinskih področij smo pričakovali manjšo invariantnost na področju branja. Vendar so rezultati pokazali, da je manjša invariantnost dosežkov prisotna na področju matematike. Čeprav so bile razlike v absolutnih vrednostih v primerjavi z

referenčnimi zelo majhne, na področju matematike v povprečju za eno točko (na lestvici s povprečjem 500 in standardnim odklon 100), so bile velikosti učinka poemembne.

Rezultati doktorskega dela nudijo uporabne informacije v zvezi z invariantnostjo ocen parametrov postavk ter dosežkov v različnih pogojih kalibracijskega vzorca. Absolutne razlike v dosežkih znotraj pogojev niso bile velike v primerjavi z referenčnim pogojem. Kljub temu pa so nekateri dobljeni rezultati presenetljivi in nudijo nekaj praktičnih nasvetov, ki bodo uporabni v nadaljnjih ponovitvah mednarodnih raziskav znanja in drugih raziskavah, ki uporabljajo modele TOP. Doktorsko delo torej prispeva k boljšemu razumevanju pojma oziroma lastnosti invariantnosti modelov TOP na realnih podatkih.

Ključne besede: teorija odgovora na postavko, invariantnost parametrov, PIRLS, TIMSS

Contents

List of Figures

List of Tables

1 Introduction

Education is an important factor in both individual and societal development. A right to education has been created, recognized and mentioned in the European Convention on Human Rights and many other conventions during recent decades. All signatory parties should guarantee the right to education for all. In this context the quality of education also plays a significant role. In the past, indicators of the quality of educational systems were formed, for example, the number of schools and students at a particular school level or the average number of teachers in proportion to the number of students etc. In recent decades, educational (school) access has no longer been the most important indicator of educational quality (Gray 1997; Štraus et al. 2006; Klemenčič and Rožman 2009). It is now clear that the mere assurance of access to the educational system cannot guarantee educational effectiveness.

One of the most important factors in assuring quality of education is the evaluation of educational outcomes (Štraus 2004). In the evaluation of educational outcomes, national and international perspectives are important (Mislevy 1995; Bela knjiga 2011). Many countries carry out national examinations. Besides national evaluation, international evaluation also plays an important role. National and international assessments follow different goals. National assessments are usually carried out on a population of students (for example elementary or secondary) and deal with a national framework. International assessments provide comparisons of different educational systems and are carried out on a representative sample of a target population in the countries that choose to participate in a specific study. They have an international framework that may extend beyond the national scope and may or may not cover it completely. Based on comparative results, countries can better understand their educational context and outcomes from an international perspective (Porter and Gamoran 2002).

The data collected from international studies is used to inform policy makers, researchers, teachers, parents, students, media etc., and even important decisions about a country's educational system are sometimes supported by or arise from international assessment data. This is the reason why it is of specific importance that results obtained from these studies are valid and reliable on both a national and an international level. Of course, every

undertaken study has some theoretical and methodological limitations which do not necessarily have a major impact on the general quality of results as long as the users are aware of them. In this doctoral thesis we attempt to evaluate a part of the methodology used in large-scale assessments (LSAs) to estimate the achievement of students or the country. There are very strict guidelines that participating countries have to follow, which ensure standardized procedures; therefore differences in outcomes cannot be attributed to different procedures between countries. International study centers have also developed special procedures for estimating students' knowledge in a specific content domain. In general, these procedures are similar across different studies but they also have specific differences. Therefore we limited the focus to international LSAs conducted by the International Association for the Evaluation of Educational Achievement (IEA) and even more specifically on their two most well-known studies that both follow very similar processes and methodologies. These are the Progress in International Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science Study (TIMSS). Both studies are repeated in regular cycles: the first cycle of PIRLS was conducted in 2001 and the first cycle of TIMSS was conducted in 1995. Since both studies also report trends in achievement, the procedures and methodologies should be consistent throughout one study in order for the results of the cycles to be comparable. In other words, the methods that were chosen for TIMSS in 1995 also have to be used in 2011 for the data to be comparable, regardless of new trends and developments in the field. However, there are still enough arguments for using the specific methodology and it has proven to work well also in the most current studies.

International studies which assess students' knowledge in different content domains use item response theory (IRT) models. These models are also used because of the matrix sampling test design in these studies where students respond only to a subset of all items. IRT models ensure comparable achievement score estimates through the overlap of items between booklets. In recent decades, more and more countries are participating in international LSA studies. For example, in the last TIMSS 2011 cycle over 60 countries and educational systems participated (most with target populations of both 4^{th} grade and 8^{th} grade students and some only with one). However, the participation of countries in studies varies from cycle to cycle. Some of the countries do not meet the expected criteria that ensures good quality of data and are therefore excluded from the international reports and also from item parameter estimations (e.g. Mongolia in TIMSS 2007). Usually, in TIMSS

and PIRLS, data from benchmarking participants (regional entities that follow the same assessment procedures as the countries) are not included in trend (or item parameter) estimation. Some countries also decide to skip a cycle. In addition, only countries participating in subsequent cycles are included in the item parameter estimation. In general, the number of participating countries increases in every data collection cycle. Since the number of countries differs from cycle to cycle there are also different common countries in sequential cycles used for item parameter estimation. Because of this, the question about the effect of the composition of the sample used to calculate the item parameters arises. Would we get the same achievement score estimates if different countries were included in the item parameter estimation?

The focus of this thesis lies in the sample of the participating countries. More specific, the focus is on the subsample of countries that are participating in subsequent cycles. Participation of countries in international studies is voluntary. In that sense our interest is whether the subsample of countries that participate (with their specific characteristics) has an effect on the methodology used to scale the achievement results for students in all participating countries. Furthermore the investigation expands to the IRT models used and content domains that are assessed in TIMSS and PIRLS.

The first part of the thesis is theoretical and starts with a short introduction of the current state of international LSA studies and very briefly describes their history. Next IRT is presented with the focus on models which are used in international studies to scale students' achievement. Specific methods to obtain plausible values (student achievement scores) that were specifically designed for LSA studies are described. The theoretical part ends with the definition of the term invariance (which is the central part of the thesis) and summarizes previous research on parameter invariance in different fields with the specific focus on invariance obtained with different samples.

The second chapter focuses only on the research problem and presents arguments for the research questions that are investigated in the empirical part of the thesis. This is followed by a detailed description of the methods used for testing the research questions. The fourth chapter presents the results that follow the order of the research questions and the description of the methodology.

In the discussion the results are evaluated and linked to the previous research in the field. Finally, the thesis ends with conclusions, practical implications of the results for the LSA

studies and includes the original contribution of this work to the development of the scientific field.

1.1 International large-scale assessment studies

Knowledge is the central element of current society. Within an increasingly globalized world, a debate on how nations should educate students for a global world has arisen. Comparative education draws on multiple disciplines to examine education in different countries. It centers on the study of education from cross-cultural and cross-national perspectives. Although comparative education provides an opportunity to explore foreign cultures and their educational systems, it can also provide a refreshed capacity to appraise a person's own culture and educational values (Kubow and Fossum 2007).

The term *international testing* is a broad concept but is generally limited to the kinds of LSAs and studies that are administered in multiple countries and provide both between and within country comparative information (Wieseman 2010). LSAs are surveys of knowledge, skills or behaviors in a given domain. The goal is to describe a population, the main focus is on the group scores and not individuals as in large-scale testing programs (Kirsch et al. 2013). Assessments usually focus on student academic achievement. Academic achievement reflects the extent to which students attain learning objectives as defined in curricula and syllabuses for specific subjects (Puklek Levpušček, Zupančič and Sočan 2012). The international studies serve to identify strengths and weaknesses of an educational system and therefore to inform policy debates in education (Phillips and Schweisfurth 2007).

Kirsch et al. (2013) state that the development of LSAs represents a cycle. The initial work is motivated by policy questions which then drive the development of assessment frameworks and the design of instruments to address those questions. The findings then create different policy questions and the cycle continues.

Wieseman (2010) reports that the most well-known and longest running international assessment organization is the IEA. More recently, the Organization for the Cooperation and Economic Development (OECD) has become involved in international educational testing at the secondary educational level. The IEA and OECD studies also collect a rich

array of background information about students' attitudes and other factors relevant and related to the students' achievement.

The history of LSAs goes back to the early 1960s, but a significant development toward a more systematic focus on national monitoring began, in opinion of Wagemaker (2011), with the release of the report *A Nation at Risk: The Imperative for Educational Reform* (A Report to the Nation and the Secretary of Education United States Department of Education by The National Commission on Excellence in Education, published in April 1983) and results of the *Second International Science Study* (Science achievement in seventeen countries: A preliminary report, released by the IEA in 1988). Before the 1980s, the question as to how and on what basis policymakers, administrators, and teachers made decisions in the field of education was to become the concern of comparative studies of education in general and the work of the IEA in particular (Keeves 2011). The IEA has brought an international perspective to the work of educational policy analysis and research (Wagemaker 2011).

The IEA started in 1958 when the founding meeting took place in Hamburg. The initial goal of the IEA was to create a study of the process of youth education in a changing world, especially the assessment, evaluation and investigation of the learning that resulted from teaching in schools and a comparison across countries. The first study undertaken by this organization, the *Pilot Twelve-Country Study* (Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961 by Foshay, Thorndike, Hotyat, Pidgeon and Walker released in 1962), sought to investigate the outcomes of educational achievement in reading comprehension, mathematics, science, geography and non-verbal ability in 12 countries (the French part of Belgium, England, Federal Republic of Germany, Finland, France, Israel, Poland, Scotland, Sweden, Switzerland, United States, and Yugoslavia). The target population was the last age level at which nearly all of an age group remained at school in the countries involved (13 years). This study established that a cross-national investigation was feasible in spite of the problems with translation and administration, since the findings were considered meaningful (Keeves 2011). In the next study (a cross-national study of mathematics in 1964, the *First International Mathematics Study*) quantitative and psychometric techniques were used, as well as random probability (two-stage in most countries) sampling at several levels of secondary education (Postlethwaite 1967).

For the *First International Mathematics Study* mathematics was chosen because it was accepted that there was more in common in the field of mathematics across countries than in any other subject. The IEA studies initially advanced and tested a large number of hypotheses and the questions addressed became increasingly more complex (Keeves 2011).

The design of probability samples at the different levels of schooling chosen was a challenging task. The samples were necessarily stratified by region, with schools as the primary sampling unit and with a specified number of students selected randomly from within each school. Four questionnaires were constructed (concerning the student, teacher, school and national information). Keeves (2011) reports that the tests and other instruments were strongly criticized by both mathematicians and psychologists, who opposed the whole enterprise of assessing outcomes of education by employing objective tests and attitude scales that could be processed by computer through the use of optical mark scored answer sheets.

The decision to formally establish a regular cycle of studies in mathematics, science and later reading was also a result of the expansion of number of participating countries. In addition to high-income countries (which formed the majority of participant countries up to the 1980's), many low and middle-income countries joined the studies. Their social, political and economic circumstances are distinguished markedly from today's OECD counterparts. The inclusion of a broader range of countries with distinctive local circumstances has led to the development of new ways of working to ensure that all countries can participate and that studies continue to achieve the highest technical standards (Kijima 2010).

Compared to the early cycles of international studies, more and more countries are participating with every new cycle (the numbers of participating countries in two of the most widespread studies are stated in future text). Over the past decade there has been an increase in the number of countries that assess student performance against peers of a similar age from other countries. Among the participating countries there has been a significant rise in the number of developing countries. Reasons for test participation among industrialized nations are easier to identify than those for less industrialized nations, as industrialized nations need to fulfill their obligations as members of organizations, such as the OECD. On the other hand, reasons for participation among developing countries are

harder to discern in part because of the specific challenges countries face when administering these tests. Today, assessments are widely considered to be a necessary tool in national education policy making that helps nation-states to adopt better decisions for education policy. International assessments are now an internationally accepted mechanism that is sought after by both developed and developing countries. These international assessments are widely considered to be some of the most legitimate tools for comparing the performances of children from various countries. For this reason, in countries where national assessments are underdeveloped, countries will rely on the results of cross-national assessment to inform policy decisions (Kijima 2010). Braun (2013) points out that as the number of participating jurisdictions (countries) grows, an increasing burden is placed on program staff, particularly if the additions involve new languages or nations with poor infrastructure. The question is whether the staff could continue to achieve broad consensus, preserve quality, and meet tight timelines. Failure to plan for the operational implications could lead international LSAs to become victims of their own success.

The program of research and evaluation conducted by the IEA responded to the need for greater accountability in education systems in countries across the world and also contributed to the transformation of the field of comparative education. The IEA has transformed empirical research in education into a scholarly enterprise across countries. The overriding goal of IEA studies (Mullis et al. 2009) is to learn more about the factors that influence student attitudes and achievement which may be manipulated to bring improvements in attitudes and achievement, or efficiencies in the educational enterprise. Keeves (2011) states that the greatest shortcoming of the first 30 years of research was the failure to take into consideration the cultural differences that operated between and within countries and national systems of education. The observed cross-national differences in achievement represent what needs to be explained by the very same organizational and cultural features often inappropriately raised as barriers to valid comparisons in the first place (Baker 1997).

It is obvious that international studies have an impact on national policymaking. Many changes in education have been triggered by the results of international studies. However, best practice should always be inferred with caution. Successful implementation of any educational policy or practice depends on the cultural, historical, and socioeconomic forces operating within and among countries. Wholesale adoption of education is seen to be

shortsighted and adaptations in overtaking practices which are put in an environmental context are being encouraged (Kubow and Fossum 2007).

Data from international comparative research is increasingly used in the definition of quality indicators of national educational systems. The aim of these studies is also to provide descriptions of the different activities performed in the education systems and their connection with students' achievements (Klemenčič 2010). Klemenčič (2010) exposes that indirect impacts of international LSAs on national policy are not measurable at all, so estimations about impacts could be misleading, or in other words, it is very difficult to determine them. Although there is no doubt that information of international LSAs is used for national purposes it is impossible to determine the amount of their impact because of all the numerous other national and international factors that play a role in national policy making.

In every study effort is undertaken to ensure reliability and validity of tests. In addition, an international study must have comparative validity. For comparative validity, the classical concerns of reliability and validity still apply, but the concepts are extended to encompass the idea that the data should be internationally comparable. That is, inferences made about achievement differences between countries can be substantiated. Goldstein (2004) reports that in the Programme for International Student Assessment (PISA) item format may be of an important feature of country differences related to curriculum and teaching. In his study he found that the pattern of item responses varies across some countries (i.e. England and France), and because of that raises his doubts about making any comparisons across countries based upon a single scale.

There are many questions that relate to the comparability of the results from international LSA studies. The questions are especially related to the validity of cross-country comparisons. Baker (1997) summarized a few of them in his paper. One concern relates to the influence of differences in how schools operate from one country to another on achievement differences. Since schooling is extremely different cross-nationally, it is claimed to be incomparable. In Baker's (1997) opinion, the tendency of modem school systems is to converge on one basic model throughout the world. It is obvious from a steady stream of cross-national studies of the organization of schools that the basic design, curricular areas, and the structure of schools have all been taking a similar course

worldwide during past decades. In the end they do differ in some ways but in general they follow the same path and are comparable.

Another issue that might have an influence on the comparability of cross-national studies is school enrollment (Baker 1997). Especially in developing countries, where for example secondary school enrollment rates are still significantly below 100%, students are more likely to come from families of higher socioeconomic status whose children are more academically advantaged (this issue is specifically important for TIMSS Advanced, since this study focuses on the secondary school population).

In Baker's (1997) opinion, the concern about testing for bias in international studies is usually expressed in one of two ways. One concern is based on different languages or cultural understanding of words in test items. In certain subject areas, such as reading or civics, language and cultural differences can indeed be an obstacle to valid test construction. This is also true for mathematics and science. Mathematical word problems, for instance, can contain language that is subject to cross national misinterpretation, and the same can be said for science test items.

Another potential issue in cross-country comparisons pointed out by Baker (1997) is that countries that participated in past international studies are not a representative sample of all countries in the world, which triggers the question of what we can learn from the comparisons themselves. Countries have not been selected (i.e. sampled with equal probability) at random to take part in any of the international studies. All countries are invited, and participation is open to any of them that wish to and have the necessary resources to participate. There is no approximation of a representative sample of the world. Furthermore, there is no sampling of countries for a priori theoretical or policy comparisons. In a strict view one cannot generalize from past international studies to some notion of "world achievement" (Baker 2001). However, as already mentioned, over the past 30 years of international achievement studies, a sufficient number of different types of countries have participated and we can assume that most common types of education systems have, at some point, been included.

These are only some of the concerns that arise from international LSAs. However, LSAs have advanced methodological innovations (such as the use of IRT that offers comparable scales across multiple forms of a test, a version of matrix sampling balanced incomplete block design, as reported by Kirsch et al. 2013) which are theoretically based. Still, there

are some questions about the comparability of the countries' achievement scores that have to be studied in more detail. In this thesis we attempt to investigate one possible problem from a methodological perspective.

Currently, there are several international studies which assess students' knowledge in different content domains and some are repeated in cycles. The studies conducted under IEA which focus on students' knowledge, for example, TIMSS, PIRLS and International Civic and Citizenship Education Study (ICCS), are curriculum based. These studies differ in terms of the population and content knowledge they assess. TIMSS focuses on mathematics and science in 4th and 8th grade students, PIRLS assesses reading literacy in 4th grade students, and the ICCS focuses on civic and citizenship education in 8th grade students. There are also some studies carried out by the OECD, for example PISA, with the focus on mathematics, science and reading literacy in 15-year old students (testing knowledge acquired for life), and the Programme for the International Assessment of Adult Competencies (PIAAC) which assesses the literacy and numeracy skills of adults aged 16-65 years and their ability to solve problems in technology-rich environments. IEA studies focus on assessing whether the intended curriculum was achieved, whereas OECD studies focus on the yield of all educational activities that take place in the country, and also on the readiness of the population to make the transition into the workforce (PISA) or that is already part of the workforce (PIAAC) in OECD countries.

In the past few decades, the IEA completed more than 20 studies. The content assessed was very broad and included, for example, mathematics, science, reading literacy, information technology in education, classroom environment, civic education, foreign languages, and literature education. The covered population ranged from primary school children to their teachers and parents. Current studies under IEA are as follows (IEA 2013):

- ICILS 2013 – International Computer and Information Literacy Study 2013;
- TIMSS 2011 – Trends in International Mathematics and Science Study 2011;
- PIRLS 2011 – Progress in International Reading Literacy Study 2011;
- ICCS 2009 – International Civic and Citizenship Education Study 2009;
- TEDS-M – Teacher Education and Development Study in Mathematics.

The participation of countries in the two of most popular studies under IEA, namely PIRLS and TIMSS, are presented in the following tables. These two studies are repeated in regular cycles: PIRLS every five years and TIMSS every four years. The design and procedures of these two studies follow the same path and are therefore the focus of our interest. Participants include not only countries but also some distinct education systems within countries (e.g. the Dutch-speaking part of Belgium and Hong Kong Special Administrative Region (SAR)). In addition in TIMSS and PIRLS also entities are participating that are not treated the same as the regular countries and sub-national jurisdictions. They are in TIMSS and PIRLS referred to as benchmarking participants (for example Canadian provinces, US states, and emirates from the United Arab Emirates). Benchmarking participants have different sample size requirements and they are not used in scaling achievement scores.

Table 1.1: Country and sub-national jurisdiction participation in TIMSS

Country	Grade 4				Grade 8				
	2011	2007	2003	1995	2011	2007	2003	1999	1995
Algeria		•				•			
Argentina							•	•	
Armenia	•	•	•		•	•	•		
Australia	•	•	•	•	•	•	•	•	•
Austria	•	•		•					•
Azerbaijan	•								
Bahrain	•				•	•	•		
Bosnia and Herzegovina						•			
Belgium (Flemish)	•		•				•	•	•
Belgium (French)									•
Botswana						•	•		
Bulgaria						•	•	•	•
Chile	•				•		•	•	
Chinese Taipei	•	•	•		•	•	•	•	
Colombia		•				•			•
Croatia	•								
Cyprus			•	•		•	•	•	•
Czech Republic	•	•		•		•		•	•
Denmark	•	•							•
Egypt						•	•		
El Salvador		•				•			
England	•	•	•	•	•	•	•	•	•
Estonia							•		
Finland	•				•			•	
France									•
Georgia	•	•			•	•			

Country	Grade 4				Grade 8				
	2011	2007	2003	1995	2011	2007	2003	1999	1995
Germany	•	•							•
Ghana					•	•	•		
Hong Kong SAR	•	•	•	•	•	•	•	•	•
Hungary	•	•	•	•	•	•	•	•	•
Iceland				•					•
Indonesia					•	•	•	•	
Iran, Islamic Rep. of	•	•	•	•	•	•	•	•	•
Israel				•	•	•	•	•	•
Italy	•	•	•	•	•	•	•	•	•
Japan	•	•	•	•	•	•	•	•	•
Jordan					•	•	•	•	
Kazakhstan	•	•			•				
Korea, Rep. of	•			•	•	•	•	•	•
Kuwait	•	•		•		•			•
Latvia		•	•	•			•	•	•
Lebanon					•	•	•		
Lithuania	•	•	•		•	•	•	•	•
Macedonia, Rep. of					•		•	•	
Malaysia					•	•	•	•	
Malta	•					•			
Moldova, Rep. of			•				•	•	
Mongolia		•				•			
Morocco	•	•	•		•	•	•	•	
Netherlands	•	•	•	•			•	•	•
New Zealand	•	•	•	•	•		•	•	•
Northern Ireland	•								
Norway	•	•	•	•	•	•	•		•
Oman	•				•	•			
Palestinian Nat'l Auth.					•	•	•		
Philippines			•				•	•	
Poland	•								
Portugal	•			•					•
Qatar	•	•			•	•			
Romania	•				•	•	•	•	•
Russian Federation	•	•	•		•	•	•	•	•
Saudi Arabia	•				•	•	•		
Scotland		•	•	•		•	•		•
Serbia	•					•	•		
Singapore	•	•	•	•	•	•	•	•	•
Slovak Republic	•	•					•	•	•
Slovenia	•	•	•	•	•	•	•	•	•
South Africa							•	•	•
Spain	•								•
Sweden	•	•			•	•	•		•

Country	Grade 4				Grade 8				
	2011	2007	2003	1995	2011	2007	2003	1999	1995
Switzerland									•
Syrian Arab Republic					•	•	•		
Thailand	•			•	•	•		•	•
Tunisia	•	•	•		•	•	•	•	
Turkey	•				•	•		•	
Ukraine		•			•	•			
United Arab Emirates	•				•				
United States	•	•	•	•	•	•	•	•	•
Yemen	•	•	•						
Out of Grade Participants									
Botswana (6,9)	•				•				
Honduras (6,9)	•				•				
South Africa (9)					•				
Yemen (6)	•								
Benchmarking Participants									
Alberta, Canada	•	•		•	•			•	•
British Columbia, Canada		•				•		•	
Ontario, Canada	•	•	•	•	•	•	•	•	•
Quebec, Canada	•	•	•	•	•	•	•	•	•
Basque Country, Spain						•	•		
Abu Dhabi, UAE	•				•				
Dubai, UAE	•	•			•	•			
Alabama, US					•				
California, US					•				
Colorado, US				•	•				
Connecticut, US					•			•	
Florida, US	•				•				
Indiana, US			•		•		•	•	
Massachusetts, US		•			•	•		•	
Minnesota, US		•		•	•	•			•
North Carolina, US	•				•			•	

Note: • Indicates participation in that testing cycle. Sources: Martin (2005, 1-3 – 1-5), Foy and Olson (2009, 82-83), Mullis et al. (2012, 422-423).

From Table 1.1 is evident that the number of participants in cycles is increasing. In the first cycle in 1995 there were 29 4th grade participants and 43 8th grade participants. The numbers in 2011 increased to 59 and 59, respectively. Between 1995 and 1999 there were 28 common participants for the 8th grade and between 1995 and 2003, 18 common participants in the 4th grade. Between 1999 and 2003, 37 common participants participated with 8th grade students. Between 2003 and 2007 there were 39 common participants for the 8th grade and 24 common participants for the 4th grade. Finally, between 2007 and 2011

there were 41 common participants for the 8th grade and 34 for the 4th grade. Not only the number of participants in cycles but also the number of common participants between cycles is increasing. Moreover, 17 educational systems participated in all cycles of the 8th grade and 15 participated in all cycles of the 4th grade.

The TIMSS Advanced study is a part of TIMSS and was conducted in 1995 together with TIMSS. In 2008 it was conducted independently and not in the same year as TIMSS. TIMSS Advanced assesses student achievement in advanced mathematics and physics in the final year of secondary school which is usually the 12th grade. In Table 1.2 the country participation in both cycles is presented.

Table 1.2: Country participation in TIMSS Advanced

Country	2008	1995
Armenia	•	
Australia		•
Austria		•
Canada		•
Cyprus		•
Czech Republic		•
Denmark		•
France		•
Germany		•
Greece		•
Islamic Rep. of Iran	•	
Israel		•
Italy	•	•
Latvia		•
Lebanon	•	
Lithuania		•
Netherlands	•	
Norway	•	•
Philippines	•	
Russian Federation	•	•
Slovenia	•	•
Sweden	•	•
Switzerland		•
United States		•

Source: Foy and Arora (2009, 80).

As can be seen from Table 1.2, five countries participated in both cycles conducted in 1995 and 2008; in 2008 the number of participating countries had decreased from 19 to 10.

PIRLS has been repeated three times until now and the participating countries and sub-national jurisdictions are presented in Table 1.3.

Table 1.3: Country and sub-national jurisdiction participating in PIRLS

Country	2011	2006	2001
Argentina			•
Australia	•		
Austria	•	•	
Azerbaijan	•		
Belgium (Flemish)		•	
Belgium (French)	•	•	
Belize			•
Bulgaria	•	•	•
Canada	•		
Chinese Taipei	•	•	
Colombia	•		•
Croatia	•		
Cyprus			•
Czech Republic	•		•
Denmark	•	•	
England	•	•	•
Finland	•		
France	•	•	•
Georgia	•	•	
Germany	•	•	•
Greece			•
Hong Kong SAR	•	•	•
Hungary	•	•	•
Iceland		•	•
Indonesia	•	•	
Iran, Islamic Rep. of	•	•	•
Ireland	•		
Israel	•	•	•
Italy	•	•	•
Kuwait		•	•
Latvia		•	•
Lithuania	•	•	•
Luxemburg		•	
Macedonia		•	•
Malta	•		
Moldova, Rep. of		•	•
Morocco	•	•	•
Netherlands	•	•	•

Country	2011	2006	2001
New Zealand	•	•	•
Northern Ireland	•		
Norway	•	•	•
Oman	•		
Poland	•	•	
Portugal	•		
Qatar	•	•	•
Romania	•	•	•
Russian Federation	•	•	•
Saudi Arabia	•		
Scotland		•	•
Singapore	•	•	•
Slovak Republic	•	•	•
Slovenia	•	•	•
South Africa		•	
Spain	•	•	
Sweden	•	•	•
Trinidad and Tobago	•	•	
Turkey			•
United Arab Emirates	•		
United States	•	•	•
Fifth Grade Participants			
Iceland		•	
Norway		•	
Sixth Grade Participants			
Botswana	•		
Honduras	•		
Kuwait	•		
Morocco	•		
Benchmarking Participants			
Alberta, Canada	•	•	
British Columbia		•	
Ontario, Canada	•	•	•
Quebec, Canada	•	•	•
Maltese - Malta	•		
Andalusia, Spain	•		
Abu Dhabi, UAE	•		
Dubai, UAE	•		
Florida, US	•		

Note: • Indicates participation in that testing cycle. Sources: Foy and Kennedy (2008, 9-10), Foy and Drucker (2013, 78-79).

In 2001 there were 37 participants; the number increased to 58 in 2011. There were 30 common participants between 2001 and 2006, and 35 between 2006 and 2011. Moreover, 24 participants participated in all three cycles of PIRLS.

International studies rely mainly on cross-sectional non-experimental designs, with data collection through sample survey methods. Much effort is undertaken to ensure comparability, reliability and validity of the outcomes. Although the procedures in international LSAs (IEA and OECD) are similar in general, there are several specific differences between studies carried out by IEA or OECD and even among different studies under IEA. One of the common characteristic used is scaling procedures of the data which rely on IRT.

1.2 Item response theory

One of the important characteristics of measurement in social and behavioral sciences is that the abilities or traits of interest are not directly measureable. We can measure the unobservable (latent variable) traits with numerous observable behaviors (manifest variables). The focus is usually on behavior and attributes which characterize an individual's behavior in home, work, school, or social settings (in general we can refer to them as psychological traits). These constructs are hypothetical concepts and in the first stage of measurement they have to be defined operationally. First, the correspondence between the theoretical construct and observable behaviors that are legitimate indicants of that construct has to be made. Measurement of the psychological attribute or trait occurs when a quantitative value is assigned to the behavioral sample collected using a test (Stevens 1946). De Ayala (2009) explains that measurement can be considered a process by which one attempts to understand the nature of a variable by applying mathematical techniques. The result then is not necessarily a number or a continuous variable (for example mathematical techniques that results in individuals being classified into latent classes and an assessment of how well the class structure describes the manifest data).

From measurements of observable behavior one can draw inference about the amount of theoretical construct that characterizes an individual. Assignment of numbers to the properties of objects must be made according to specified rules. The development of systematic rules and meaningful units of measurement for quantifying empirical

observations is known as scaling (Crocker and Algina 1986). Scaling is the process of associating numbers or other ordered indicators with the performance of examinees. These numbers or ordered indicators are intended to reflect increasing levels of achievement or ability (Kolen and Brennan 2004). However, in the real world we cannot obtain the value of the examinee's ability parameter. The best we can do is to obtain an estimate.

The study of pervasive measurement problems and methods for their resolution has evolved into the specialized discipline in education and psychology known as test theory. It provides a general framework for viewing the process of instrument development. Mathematical models and methods do not rest on any particular psychological or educational theory and may be equally useful for measurement of many different attributes (Crocker and Algina 1986).

Classical test theory was the mainstay of psychological test development for most of the 20th century (Embretson and Reise 2000). Van der Linden and Hambleton (1997, 2) state that the classical test theory starts from the assumption "that systematic effects between responses of examinees are due only to variation in the ability (true score) of interest. All other potential sources of variation existing in the testing materials, external conditions, or internal to the examinees are assumed either to be constant or to have an effect that is nonsystematic or 'random by nature'". The observed score is decomposed into a true score and an error score (error of measurement):

$$X=T+E$$

where X represents the observed test score, T the individual's true score and E a random error component.

The weak theoretical assumptions of classical test theory make it applicable to the development of various tests and test score analysis problems. The assumptions involve a random and normally distributed error around the true score (where the expected value of the error is 0). Random errors are uncorrelated with each other and also to the true score. In other words, in the population the errors are uncorrelated with the trait scores for an instrument, the errors on one instrument are uncorrelated with the trait scores on a different instrument, and the errors on one instrument are uncorrelated with the error scores on a different instrument (de Ayala 2009).

Statistical inferences from the data collected cannot be generalized beyond the standardized levels of its error or nuisance variables (van der Linden and Hambleton 1997). Therefore, test scores are not comparable between different tests (unless they are parallel), and person scores cannot be directly compared unless the same test has been taken. The need for obtaining invariant parameters drove the development of a new test theory which was first extensively described and summarized by Lord and Novick (1968). IRT has rapidly become mainstream and a basis for psychological measurement. But still there is a number of psychological tests that are developed based on classical test theory.

IRT, also known as latent trait theory, is a model-based measurement framework in which trait level estimates depend on both persons' responses and the properties of the items that were administered (Embretson and Reise 2000). A person's latent trait level (which is usually denoted as θ) is estimated from responses to the (test) items. An IRT model includes trait level and item properties, which are related to a person's item responses. IRT is based on a mathematical model of how examinees at different ability levels for the trait should respond to an item. It is because of this characteristic that the performance of examinees who have taken different tests can be compared. It also permits one to apply the results of an item analysis to groups with different ability levels from those of the groups used for the item analysis (Crocker and Algina 1986).

IRT models involve two key assumptions (Embretson and Reise 2000): the item characteristic curves (ICCs) have a specified form, and local independence has been obtained. An ICC describes how changes in the latent trait level relate to changes in the probability of a specified response. ICC is the graphical representation of item response function and represents the probability of success on item i, usually denoted as $P_i(\theta)$, as a function of the trait level θ. ICCs plot the probability of a correct response as a monotonic increasing function of a trait level (logistic functions and normal ogive functions are the most prevalent).

In Figure 1.1 and Figure 1.2, ICCs for two different items are presented. In the first figure a three parameter logistic model (3PL) was used, whereas the second figure is the result of a generalized partial credit model.

Figure 1.1: Item characteristic curve for a three parameter logistic model (3PL) with a=1.580, b=0.085 and c=0.127. Parameter estimates are derived from the 3PL model using PARSCALE software.

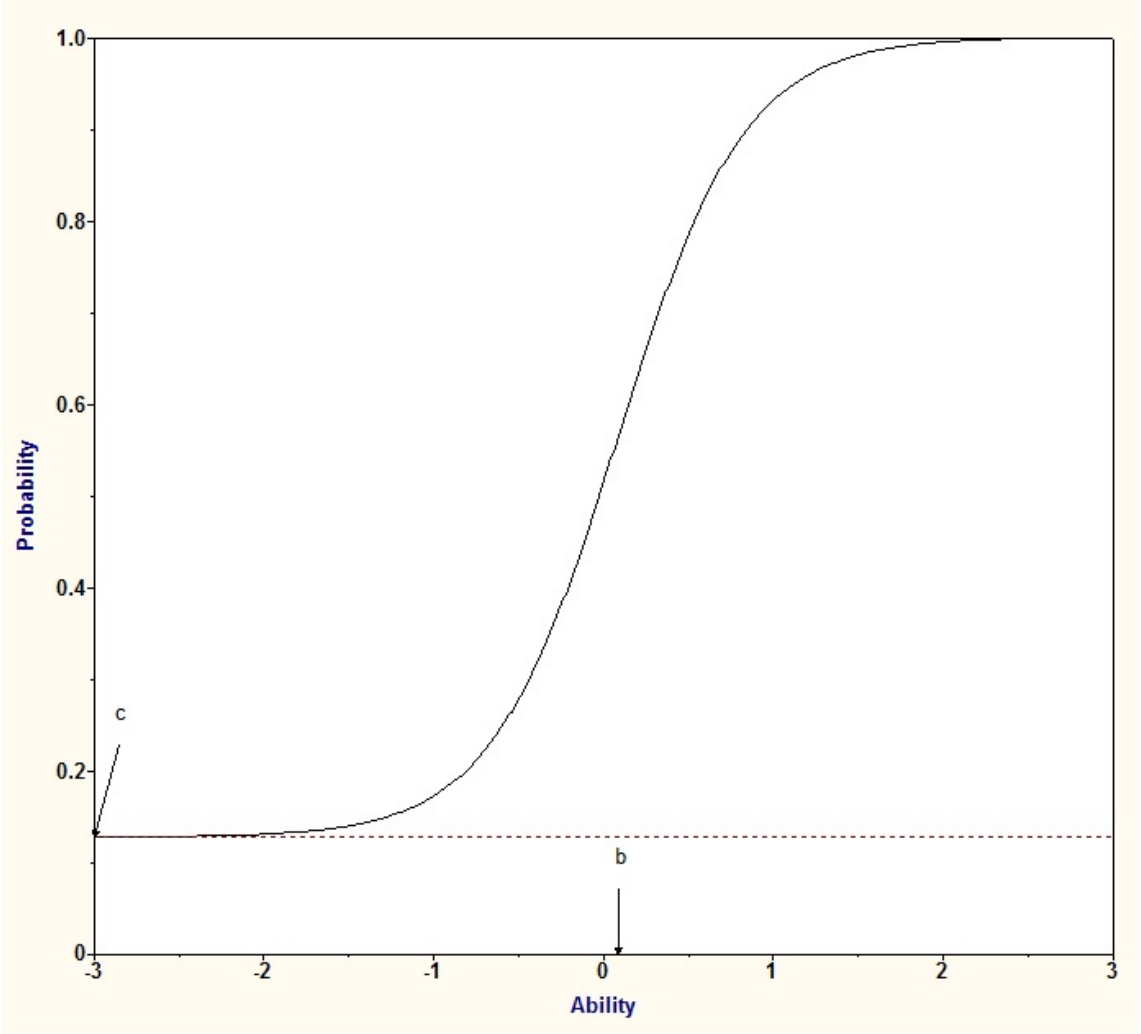
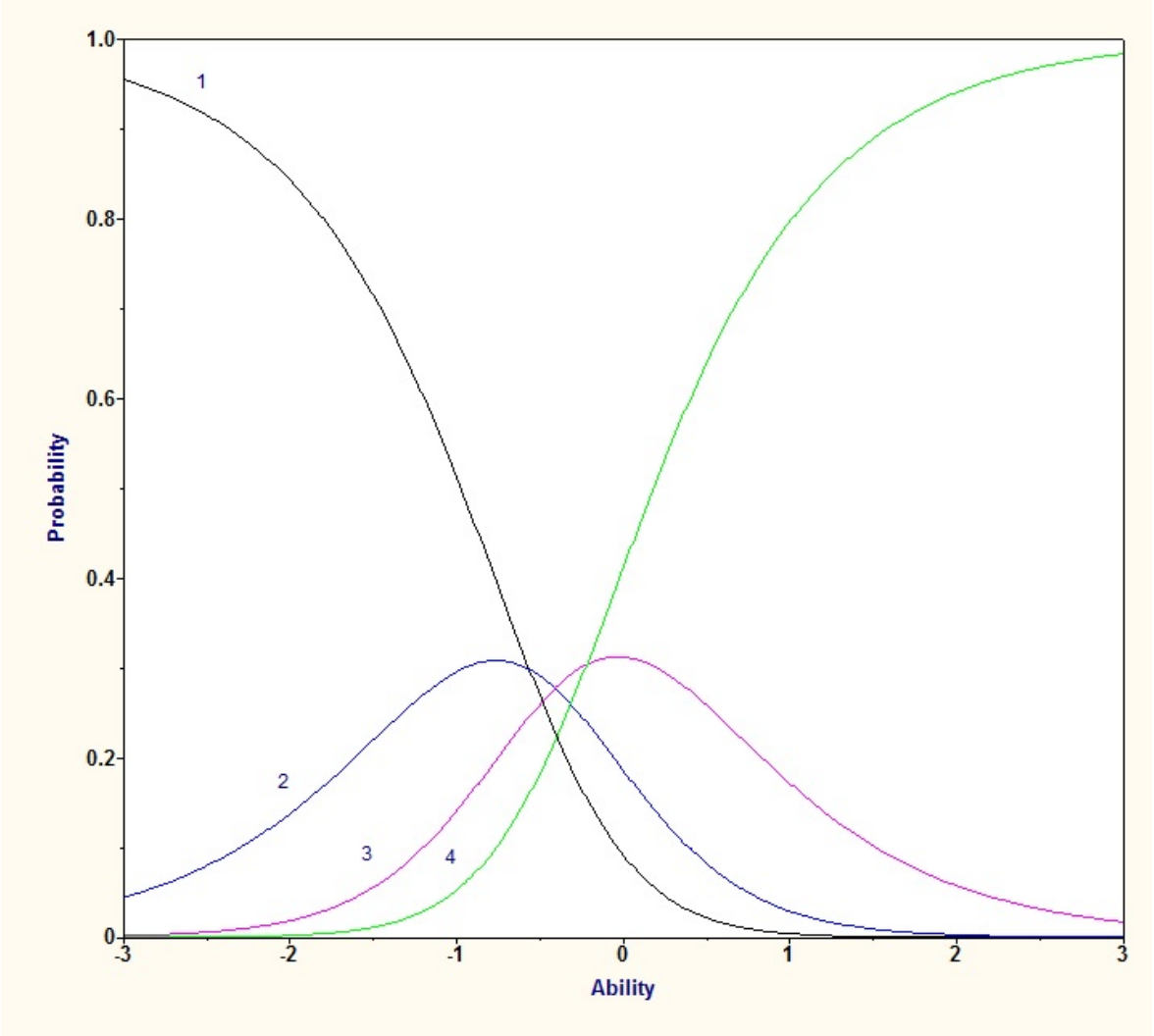


Figure 1.2: Item characteristic curves from a generalized partial credit model (normal metric) with a=0.74, b=-0.40, d_1=0.17, d_2=0.01 and d_3=-0.18. Parameter estimates are derived from a generalized partial credit model using PARSCALE software.



Local independence of items is reached when the relationship among items is fully characterized by the IRT model, or in other words, if solving an item is independent of the outcome of any other item when we control for the latent trait. Local independence is also evidence for unidimensionality, if the IRT model contains person parameters on only one dimension. However, local independence and unidimensionality are two different constructs (Crocker and Algina 1986). The dimensionality of a test is equal to the number of latent traits required to achieve local independence.

The assumption of a unidimensional latent trait is common for test construction since unidimensional constructs enhance interpretability. It is clear that the assumption cannot be strictly met because there are always some cognitive, personality, and test taking factors

that impact the test performance at least to some extent (Hambleton and Swaminathan 1985). Thus, local independence and the number of latent traits are mostly a matter of assumption (Crocker and Algina 1986). If tested in practical situations usually a degree of local independence and a domination of one dimension are examined.

IRT is used for scale construction to measure the latent trait. IRT models can be applied to measure personality traits, moods, behavioral dispositions, situational evaluations, and attitudes as well as cognitive traits (Embretson and Reise 2000). "Measurement instruments must first be created, and the units calibrated, so that we all agree on the reproducibility of their location" (Bond and Fox 2001, 3); only then we can use the instruments to measure the desired trait. Before applying IRT we have to be sure that the instrument (or test) measures the desired trait. Then it is important to apply it to a representative sample of the desired population. The last part is of specific importance, if we would like to make generalizations of outcomes on the population.

The latent trait scale in IRT has an arbitrary origin and unit of measurement. The arbitrary origin means that any one of the homogeneous subpopulations can be assigned a score of 0 since none of the subpopulations is characterized by a complete absence of the latent trait. The arbitrary unit of measurement means that after one group is assigned a score of 0, any other homogeneous subpopulation, whose members have more latent ability than members of the zero subpopulation, can be assigned a score of 1. The ability difference between these two subpopulations is then the unit of measurement. Since the unit and origin is arbitrary, it is very common to choose the origin and unit so that the mean latent trait score is 0 and the standard deviation is 1 for some population of interest (Crocker and Algina 1986).

In IRT, the true score of an examinee is defined as follows:

$$TS_j = \sum_{i=1}^{N} P_i(\theta_j)$$

where TS_j is the true score for examinees with ability level θ_j. i denotes an item, and $P_i(\theta_j)$ depends upon the particular item characteristic curve model employed.

1.2.1 IRT models

There are many ways (models) in which the relationship between item responses and underlying abilities can be specified. The person's response pattern to a particular set of items (data matrix) provides the basis for estimating trait level. One way of classifying IRT models is on the basis of the examinee responses. Models then can be applied to dichotomously scored items (true-false, short answer, sentence completion, matching items, forced choice etc.), which are most commonly used. Although the normal ogive was the predominant function for the ICC in early research on latent trait theory, in the mid-80s it has largely been replaced by logistic models which require simpler computations (Crocker and Algina 1986).

There are many IRT models differing in the mathematical form of the ICC and (or) the number of parameters specified in the model. The logistic IRT models are based on the logistic distribution. For example, the Rasch model (or one parameter logistic model – 1PL) transforms raw data into abstract (latent), equal interval scales. Equality of intervals is achieved through log transformations of raw data odds, and abstraction is accomplished through probabilistic equations (Bond and Fox 2001, 7). With the Rasch model, all items are assumed to have the same discriminating power, while the two parameter logistic model (2PL) and 3PL provide an extra item parameter to account for differences among items in discriminating power (2PL and 3PL) and one to account for guessing (3PL).

The ICCs can differ in location (the location of the inflection point of the curve), which describes the extent to which items differ in probabilities across trait levels or item difficulty; in slope, which describes how rapidly the probabilities change with trait level (due to the S shape of the ICC; the slope of the curve changes as a function of the ability level and reaches a maximum value when the ability level equals the item's difficulty and thus, the item is doing its best in distinguishing between examinees in the neighborhood of this ability level, Baker 2001) or item discrimination; and in lower asymptote, which changes the lower limit of the item probability range (the lower bound is greater than zero, Embretson and Reise 2000) or item guessing. The lower asymptote denotes the probability of getting the item correct by guessing alone. It is important to note that by definition, the value of this parameter does not vary as a function of the ability level. A side effect of using the guessing parameter is that the definition of the difficulty and discrimination parameters is changed. Under the Rasch and 2PL models, the location parameter

(representing item difficulty) is the point on the ability scale at which the probability of a correct response is 0.5. Now this probability is halfway between the value of the lower asymptote and 1.0. When including the guessing parameter in the model the slope parameter slightly changes. The slope parameter in 2PL (without the lower asymptote) is negatively correlated with the slope parameter in 3PL (with the lower asymptote). More specific, under the 3PL, the slope of the ICC at $\theta = b$ is actually a (1 - c)/4 (where b represents the difficulty parameter, a the slope parameter and c the guessing parameter, Baker 2001).

The partial credit model is a unidimensional model and can be considered as an extension of the 1PL model; it has all the standard Rasch model features such as separability of person and item parameters. It can be used with items where partially correct answers are possible and is appropriate also for analyzing attitude or personality scale responses where subjects rate their beliefs or respond to statements on a multi-point scale (Embretson and Reise 2000). When an item provides more than two (ordinal) response categories, for example 0, 1 and 2, a score of 1 is not expected to be increasingly likely with increasing ability. From some point on, a score of 2 becomes more probable and a score of 1 becomes less probable. It follows from the order $0<1<2<\ldots.<m_i$ that the conditional probability of scoring x rather than x-1 on an item should increase monotonically throughout the ability range. By conditioning on a pair of adjacent categories (and so eliminating all other response possibilities from consideration), the model focuses on the local comparisons of categories x-1 and x (van der Linden and Hambleton 1997). The term d can also be directly interpreted as the point on the latent trait scale at which two consecutive category response curves intersect. The d intersection parameters can be considered as step difficulties associated with the transition from one category to the next and there are m step difficulties (intersections) for an item with m_i+1 response categories. The d parameters represent the relative difficulty of each step (Embretson and Reise 2000).

A generalized partial credit model is based on the partial credit model, relaxing the assumption of uniform discriminating power of test items. This model can attain some of the objectives that the Rasch model achieves and can also provide more information about the characteristics of test items than the Rasch model. If the number of response categories is m_i then m_i-1 category threshold parameters can be arbitrarily defined as any value. The parameters b_{ih} can be decomposed to b_i-d_h. The values of d_h are not necessarily ordered sequentially within an item. The parameter d_h is interpreted as the relative difficulty of

category h in comparing other categories within an item or the deviate of each categorical threshold from the item location b_i. The location constraint ($\sum d_h=0$) is imposed to eliminate an indeterminacy (van der Linden and Hambleton 1997). An example of an ICC for the generalized partial credit model is presented in Figure 1.2.

As already mentioned, different models are used since there are different item types in the test. A 3PL is used for multiple choice items where there are only correct and incorrect responses, a 2PL is used for constructed-response items with just two response options, and a (generalized) partial credit model is used for polytomous constructed response items (more than two response options).

Below the equations for IRT models are presented. Only equations for already described models, which are also used in international LSAs (PIRLS and TIMSS), and in the empirical part of the thesis are presented:

3PL model:

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$$

2PL model:

$$P_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$$

1PL or Rasch model:

$$P_i(\theta) = \frac{e^{D(\theta-b_i)}}{1+e^{D(\theta-b_i)}}$$

Partial credit model:

$$P_i(\theta) = \frac{e^{D(\theta-b_{i,x})}}{1+e^{D(\theta-b_{x,i})}}, \text{ or}$$

$$P_i(\theta) = \frac{e^{\sum_{v=0}^{l} D(\theta_k-d_{i,v})}}{\sum_{g=0}^{m_i-1} e^{\sum_{v=0}^{g} D(\theta_k-d_{i,v})}}$$

Generalized partial credit model:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{\sum_{g=0}^{m_i-1} (e^{\sum_{v=0}^{g} Da_i(\theta_k - b_i + d_{i,v})})}$$

b_i (-∞, ∞) – difficulty parameter (the location parameter of item i) – the point on the ability scale where an examinee has a probability of success on the item i of 0.5 or halfway between the value of the lower asymptote and 1.0 (in case of a 3PL)

a_i (0, ∞) – discrimination parameter (the slope parameter of item i) – proportional to the slope of the tangent to the response function at point b

c_i (0, 1) – a lower asymptote parameter (the guessing parameter of item i)

$d_{i,l}$ (l=0, m_i-1) – category l threshold parameter or step parameter of item i (m is number of response categories for item i)

θ (-∞, ∞) – latent trait or trait level of an examinee

In all equations, D is a constant which is set to 1.7 because $P_i(\theta)$ for the normal and logistic ogives do not differ by more than 0.01 for any value of θ (Lord and Novick 1968). In this case item parameters are interpreted in terms of the logistic function. Thus, the reported values could be divided by 1.7 to obtain the corresponding normal ogive values (Baker 2001).

It should be noted at this point that these properties say nothing about whether the item really measures some facet of the underlying ability or not; that is, a question of validity (Baker 2001).

1.2.2 Parameter estimation

Depending on the logistic model chosen, the ICC parameters have to be estimated for every item. This procedure is an iterative process. The most common estimation procedures are maximum likelihood and so-called heuristic or approximate procedures (under this approach, initial values for the item parameters, such as b = 0.0, a = 1.0, are

established a priori, Baker 2001). When the latent trait scores and the item parameters are estimated simultaneously, we call the procedure the joint maximum likelihood procedure. The joint maximum likelihood procedure jointly solves the equations for the values of unknown parameters in an iterative scheme, which starts with initial values for the ability parameters, fixes the item parameters, and solves the equations for improved estimates of the values of the ability parameters etc. (van der Linden and Hambleton 1997). Crocker and Algina (1986) report several drawbacks of the procedure when a 3PL is used. Firstly, the procedure needs a substantial number of examinees for accurate estimation (at least 1000). If the purpose is to get a stable estimate, an even larger sample might be required. Secondly, it is not well investigated whether the estimates of the item parameters are consistent. Another alternative is to use the marginal maximum likelihood procedure or conditional maximum likelihood procedure. The later can be only used for a 1PL or Rasch model.

The 3PL model suffers from multicolinearity. Estimates of *a* and *c* are sensitive to minor fluctuations in the responses used to produce these estimates. Unless huge samples of examples or tight priors around the true parameter values are used, the estimates are unstable (van der Linden and Hambleton 1997).

There are two cases for which the maximum likelihood estimation procedure fails to yield an ability estimate. First, when an examinee answers none of the items correctly, the corresponding ability estimate is negative infinity. Second, when an examinee answers all the items in the test correctly, the corresponding ability estimate is positive infinity. In both of these cases it is impossible to obtain an ability estimate for the examinee (Baker 2001). Without some distribution assumption it is in general impossible to determine the ability level of a student who answered every item or no item correct. His ability could even be very close to minimum or maximum or very far away.

Since the limitations of the mentioned method, Bock and Aitkin (1981; in van der Linden and Hambleton 1997, 15) reformulated the marginal maximum likelihood method with the expectation maximization method. In the first (expectation) step, the provisional expected frequency and the provisional expected sample size are computed. In the next (maximization) step, the marginal maximum likelihood estimates are obtained by Fisher's scoring method (Kendal and Stuart 1973, in van der Linden and Hambleton 1997, 156).

This method is also implemented in IRT model scaling programs (see, for example, PARSCALE).

Item parameters are estimated based on the examinees responses, which reflect the characteristics of items and persons (difficulty, discrimination etc.). Item parameters and examinees' responses are then combined in an estimation of the latent trait. The trait level is estimated in the context of an IRT model, and therefore we refer to IRT as model-based measurement.

1.2.3 Choosing the model

The choice of model to use in a particular situation is a rather complex one. One factor in the choice concerns how realistic the assumptions of the models are. The 3PL model can accommodate for guessing, and guessing must be considered as a possibility on multiple-choice and true-false items. It may seem that the three-parameter model should be used with multiple-choice and true-false tests, whereas the two-parameter model should be used with other types of tests. However, guessing may be negligible on some multiple-choice and true-false tests and variation in item discrimination may be negligible for any type of test. In this case, the one-parameter model will be entirely adequate. This is an important consideration since the use of an unnecessarily complex model will probably result in less-accurate estimates and less-adequate applications than use of an adequate, simpler model. The unnecessarily complex model will require estimation of parameters that really do not need to be estimated (Baker 2001).

Furthermore, the choice of model depends on the extent to which an application of a simpler model is robust to violations of its assumption. The issue of robustness to violations arises because estimation of more complex models tends to be less practical (Baker 2001). If the number of parameters increases, fewer data per parameter are available and parameter estimates may show serious instability. The problem is aggravated if the model has a tradeoff between some of the parameters in their effects on the response probabilities. As already mentioned for the 3PL model (and also other statistical models), this condition is reminiscent of the problem of multicollinearity in multiple linear regression, and a prohibitively large amount of data may be needed to realize stable parameter estimates. Another likely problem with complicated models is the lack of

identifiability of parameter values. Even if parameters exist, attempts to estimate them may result in maximum likelihood estimates which are not unique or it is possible that no maximum likelihood estimates exist at all. Finally, if unique parameter estimates are known to exist, they may still be hard to calculate (van der Linden and Hambleton 1997).

As Rupp and Zumbo (2003) point out, the choice of appropriate model can also be a matter of training or tradition. However, in practical situations, model choice is usually driven by real constraints placed by models on the data input and the consumers on their output. Simpler models have fewer parameters to estimate and thus make less stringent sample size requirements for stable parameter estimation. Even if parameters in more complex models can be estimated to the desired degree of accuracy, it may be difficult to interpret them meaningfully from a substantive theoretical viewpoint, and this may be what is desired.

1.2.4 Evaluation of model fit

In any application of latent trait theory, several interrelated issues must be addressed. One of them is the goodness of fit of a model to the data. No single statistical test is an adequate indication that a model fits the data. Instead, a series of tests to explore a variety of ways that a model may misfit data should be conducted (Crocker and Algina 1986; Embretson and Reise 2000). In exploration of model fit also formal assessment of dimensionality and local independence might be helpful.

Sometimes assessment material is hierarchically structured, which means that several items relate to a single context (Monseur et al. 2011). As Monseur et al. (2011) warn, because multiple items are connected together to a common passage, items within a unit are not likely to be conditionally independent, so the assumption might be violated. In their research, Monseur et al. (2011) found that the consequence of local item independence violation in PISA is that the relative variability of low-performing countries is overestimated while the relative variability of high-performing countries is underestimated.

Van der Linden and Hambleton (1997) state that statistical tests should not be used solely to determine the adequacy of model fit. Also, checks on the presence of ability and item parameter invariance provide valuable information about model fit.

Since different models can be fit to different items the fit of a particular IRT model can be judged separately item by item. Embretson and Reise (2000) reported that there are basically two general approaches for item fit evaluation. One way is to graphically compare estimated item response curve with the empirical one and present it in a plot. In graphical procedures no real statistical tests are performed. Another way is to formalize these comparisons by a statistic that tests for the significance of the residuals (usually chi-square goodness of fit index is used). If the value of the obtained index is greater than a criterion value, the ICC specified by the values of the item parameter estimates does not fit the data. This can be caused by two conditions. Either the wrong ICC model may have been employed, or the values of the observed proportions of correct responses are so widely scattered that a good fit cannot be obtained (regardless of the model, Baker 2001).

Another option is to evaluate individual person fit. There are several approaches, as reported by Embretson and Reise (2000) that attempt to assess validity of the IRT model by producing person fit-indices. Both approaches, the item and person fit, can then provide information about a model fit (some item and person indices can be aggregated to provide a general indication of model fit).

Furthermore a model comparison approach can be used. A researcher can fit a 2PL and a 3PL model and then compare the log likelihoods of both models using chi-square statistics. This procedure is more used for model comparisons (which model fits better) than for absolute judgments of fit (Embretson and Reise 2000).

Well established statistical tests do not exist for the 2PL and 3PL model since the utility of statistical tests in assessing model fit is questionable especially for large samples (among others also because of the fact that chi square statistic is very sensitive to sample size). Since no model is likely to fit a set of test items perfectly, given sufficient amounts of data, the assumption of model fit or adequacy, whatever the model, is likely to be rejected (van der Linden and Hambleton 1997). Embretson and Reise (2000) report that for the majority of research applications, if the data is unidimensional and we have a large sample, it does not make much difference which particular IRT model is used.

The field of assessing model fit in IRT is still developing. There are a lot of attempts in using different methods (for example: a scaling correction for the chi-squared fit statistic, fit statistic based on posterior expectations, as reported by Stone and Zhang 2003, and goodness of fit framework based on logistic regression as proposed by Mair et al. 2008)

that are beyond the scope of this thesis. The above mentioned approaches are the only the very basic ones which are still widely used today.

1.2.5 Test equating and linking procedures

On some occasions it is required that tests are administered more than once. The purpose can be to track educational trends over time or more testing dates of the same content (so the examinees have some flexibility to choose the date). In any case it is not very wise for the test questions to be the same (because an examinee tested twice might be administered the same test). These issues can be addressed by administering a different collection of test questions (test form) to examinees on different dates. Equating is a statistical process that is used to adjust scores on test forms so that scores on the tests can be used interchangeably. Equating adjusts for differences in difficulty among forms that are meant to be similar in difficulty and content. Equating adjusts for differences in difficulty and not for differences in content (Kolen and Brennan 2004).

There are processes that are similar to equating and can be referred to as linking (or scaling to achieve comparability). The goal of linking is to put scores from two or more tests on the same scale. Although similar statistical procedures are often used in linking as in equating, their purposes are different. Tests that are purposefully built to be different are linked, whereas equating is used to adjust scores on test forms that are built to be as similar as possible in content and statistical characteristics. When equating is successful, scores on alternate forms can be used interchangeably (Kolen and Brennan 2004). In general equating can be seen as a specific (strongest) form of linking.

Holland and Dorans (Holland et al. 2007) divide linking methods into three basic categories called predicting, scale aligning, and equating. The goal of predicting is to predict an examinee's score on a test based on other information about that examinee. It can be applied to examinees who are similar to those in the population from which the prediction equations are derived, then they are likely to be useful. For examinees who are very different from them, these predictions are less likely to be accurate.

The goal of scale aligning is to transform the scores from two different tests onto a common scale and has many subcategories (battery scaling, anchor scaling, vertical scaling, calibration and concordance). The goal of equating is to establish an effective

equivalence between scores on two test forms such that the scores from each test form can be used as if they had come from the same test (Holland et al. 2007).

In LSA, it has been a common practice to include a set of items in the cognitive part for repeated use across years. These common items, referred to as anchor or linking items are used to equate or link test scores of multiple test forms at different time points. Using common items, test scores are linked and become comparable for different groups of examinees.

1.3 Overview of test design in LSA

Student achievement in international LSA studies is measured by administering objective tests to a sample of students who have been selected as representative of national populations. To assess students' knowledge, a wide variety of items is used but individual students respond only to a subset of items. This is done to achieve a broader content coverage in limited testing time. In order to ensure that items receive sufficient exposure in the sample and that sufficient items are administered to individual students to estimate population proficiency reliably, a complex rotated booklet design is used. Specifically, items are assembled into a non-overlapping set of blocks with 10 to 15 items per block. The assessment blocks are assembled to create a balance across blocks and booklets with respect to content domain, cognitive domain, and item format. Blocks are further paired into booklets. Booklets are randomly assigned to students and in this design they enable a comprehensive picture of the content assessed in the population of interest.

To enable linking between booklets, each block appears in two booklets. An example of the design is presented in Table 1.4. It consists of eight blocks (A-H) that are paired into seven booklets (1-7).

Table 1.4: An example of a complex matrix-sampling booklet design

Booklet	1	2	3	4	5	6	7
Block one	A	B	C	D	E	F	G
Block two	B	C	D	E	F	G	H

For example, in 2007, the 8th grade TIMSS assessment included 429 total mathematics and science items distributed across 14 mathematics blocks (M01–M14) and 14 science blocks (S01–S14), arranged into 14 booklets with four blocks each. Under this design, each block (and therefore each item) appears in two booklets. These 28 blocks of items represent more than 10 hours of testing time; however, the booklet design used by TIMSS reduced individual testing time to 90 minutes per student plus 30 minutes for the student background questionnaire (Rutkowski et al. 2010).

This procedure does not permit a precise estimation of examinees' ability (θ). The relatively small number of items per block and the relatively small number of blocks per test booklet mean that the accuracy of measurement at the individual level of these assessments is considerably lower than is the level of accuracy common for individual tests used for diagnosis, tracking, and/or admission purposes (von Davier et al. 2009). With this approach, the advantage of estimating population characteristics is more efficiently offset by the inability to make precise statements about individuals. The measurement of individual proficiency is achieved with a substantial amount of measurement error (von Davier et al. 2009).

1.4 Plausible value methodology

Results of early international assessments were commonly reported in terms of total number of correct scores or average percentage of correct scores. Such scores are reasonable as long as they are based on a common set of items (Linn 2002). With the use of a complex rotated booklet design, the methodology also had to be adjusted.

One way of taking the uncertainty associated with the estimates into account, and of obtaining unbiased group-level estimates, is to use multiple values representing the likely distribution of a student's proficiency. These so-called plausible values provide us with a database that allows unbiased estimation of the plausible range and the location of proficiency for groups of students. Plausible values are based on student responses to the subset of items they receive, as well as on other relevant and available background information (Mislevy 1991). Plausible values can be viewed as a set of special quantities, generated using a technique called multiple imputation. Plausible values are not individual

scores in the traditional sense, and should therefore not be analyzed as multiple indicators of the same score or latent variable (Mislevy 1991).

In order to overcome the challenge associated with the design, LSA studies adopted a population or latent regression modeling approach that uses marginal estimation techniques to generate population level achievements. In the following section we summarize the plausible value methodology as it is described in the paper of Mislevy et al. (1992).

If the θ values were available for every student, it would be possible to compute any statistic $t(\theta,Y)$, where Y represent responses of examinees to background questions, to estimate a corresponding population quantity T. Another function $U(\theta,Y)$ would be used to estimate sampling uncertainty as the variance of t around T in repeated samples from the population. Because IRT models are latent variable models, θ values are not observed even for the examinees in the sample. To overcome this problem, θs are treated as missing values and the approximation of $t(\theta,Y)$ is obtained by its expectation given (X,Y), the actual observed data, where X is the matrix of item responses for all examinees:

$$\hat{t}(X,Y) = E\left[t(\theta,Y|X,Y\right] = \int t(\theta,Y)\,p(\theta|X,Y)\,d\theta$$

In special cases it is possible to obtain an estimate of a population characteristic without ever obtaining a score estimate for a single individual (calculate the integral equation directly). However, closed-form solutions are not forthcoming with IRT models and alternative methods must be sought to evaluate the equation. Random draws from the conditional distributions $p(\theta \mid x_i,y_i)$ are performed for every examinee i. The random draws can be viewed as imputations in the missing data terminology and are referred to as plausible values in LSA terminology. Typically, five values are drawn for each examinee so that the uncertainty associated with the fact that θs are not observed can be quantified (if the measurement error is small, then multiple scores for an individual will be close together; if the measurement error is large, then multiple scores for an individual will be far apart; Wu 2005). Plausible values are drawn from distributions that already implicitly include the characteristics of the population through the factor $p(\theta \mid X,Y)$. The plausible values thus only reflect the population characteristics with which they are constructed. Even though precise scores are available for every examinee, plausible values are not test scores for individuals in the usual sense. They are offered only as intermediate calculations to compute estimates of population characteristics and not for inferences or decisions about

individual examinees. Wu (2005) states that any one set of plausible values will give unbiased estimates of group distributions and differences between subgroups. The average of these estimates across the subgroups will give us the best estimates of the group-level statistics of interest.

Using first Bayes' theorem and the IRT assumption of conditional independence ($P(x_i \mid \theta, y_i) = p(x_i \mid \theta)$),

$$p(\theta|x_i, y_i) \propto P(x_i|\theta, y_i)\, p(\theta|y_i)$$
$$= P(x_i|\theta)\, p(\theta|y_i)$$

where $P(x_i \mid \theta)$ is the likelihood function for θ and $p(\theta \mid y_i)$ is the distribution of θ given the observed value y_i of background responses. A normal distribution is assumed for θ and the following model is fit:

$$\theta = \Gamma' y^c + \varepsilon,$$

where ε is normally distributed with mean 0 and dispersion $\sum$ and y^c is the vector of complete background variables. Γ and $\sum$ are the parameters to be estimated. As in regression analysis, Γ is a vector or matrix of effects and $\sum$ is a scalar or matrix of variance residuals. Maximum likelihood estimates of Γ and $\sum$ can then be obtained. The conditional distribution $p(\theta \mid y_i)$ is assumed to be multivariate normal with mean $\mu_i^c = \Gamma' y_i^c$ and covariance matrix $\sum$. A plausible value is drawn at random from this normal distribution.

There are some differences between plausible values and the latent ability parameter θ as defined in the usual 1PL, 2PL or 3PL models. Instead of directly estimating a student's θ, a probability distribution for a student's θ is estimated as described before. Furthermore, instead of obtaining a point estimate for θ, a range of possible values for a student's θ with an associated probability for each of these values is estimated. Plausible values are random draws from the (estimated) distribution for a student's θ. This distribution is referred to as the posterior distribution for a student. Plausible values are meant to be used to estimate population characteristics, and as Wu (2005) shows, they perform better than point estimates of abilities. The sample mean and sample variance of the distribution of plausible values built from all students are unbiased estimates of the population mean and variance.

Wu (2005) reported that for the Rasch model, many point estimates of ability are possible, for example maximum likelihood estimates, weighted maximum likelihood estimates or

expected a-posteriori estimates. However, in contrast to these estimates, plausible values are used to construct the population distribution. This distribution is smoother, since examinees with the same total score (and same posterior distribution) will likely have different plausible values. Therefore the resulting distribution is a better representation of the underlying continuous population distribution ($g(\theta)$). According to Wu (2005), plausible values perform well in recovering the population mean, variance and percentiles, even when very short tests are administrated.

Beaton and Johnson (1992) report that the theory and use of plausible values was first developed for the analyses of 1983-84 US National Assessment of Educational Progress (NAEP) data based on Rubin's (1987, in Mislevy et al. 1992, 138) work on multiple imputations. Plausible values were used in all subsequent NAEP surveys, and are now also used in surveys such as TIMSS, PIRLS and PISA.

1.5 Scaling in international LSA studies

1.5.1 Scaling procedures in PIRLS and TIMSS

Assessment programs, like PIRLS, PISA, and TIMSS, use a complex two-stage clustered sampling design (OECD 2005; Martin et al. 2007; Olson et al. 2008; OECD 2009). In Stage 1, schools are chosen based on a probability proportional to the (school's) size, whereby larger schools are chosen with higher probability. In IEA studies, the second stage consists of choosing randomly one or two intact classes at the 4^{th} grade (TIMSS and PIRLS) or 8^{th} grade (TIMSS) level. All students in the selected classes are then assessed. Alternatively, the PISA approach results in the random selection of a set number of individual students (usually 35) from each sampled school's list of 15-year olds.

In addition to the cognitive items, students also respond to a number of background questions that provide information about their home and school environment. The scaling relies on IRT and combines students' responses to provide accurate estimates of achievement in each participating country as well as trends in achievement for countries that have participated in the previous cycle of the study.

In every cycle, some of the assessment blocks are released to the public and replaced by newly developed blocks. A number of assessment blocks are also kept secure to be used again in future assessments. These blocks establish the link to the previous cycle so the achievement scores can be made using the same metric as those used previously. The procedure enables measuring trends through time in countries that participate in subsequent cycles.

On the following pages the procedures used in PIRLS and TIMSS are described as they are reported in the technical reports of each study (Martin et al. 2007; Olson et al. 2008); however, the methods used in PISA are generally similar and originate from methods developed for the NAEP (Beaton and Johnson 1992, Mislevy et al. 1992).

Since a test is usually subject to different types of items, different models must be implemented within each assessment. IEA studies TIMSS and PIRLS use a 3PL for multiple choice items, a 2PL for constructed-response items with just two scoring options, and a generalized partial credit model for polytomous constructed response items (more than two response options). The models have already been presented and described in the section Parameter estimation.

In general, the application of IRT scaling and plausible value methodology involves four major tasks:

1. Calibrating the achievement test items – estimating item parameters;
2. Creating principal components based on the background information for use in conditioning;
3. Generating IRT scale scores;
4. Placing the proficiency scores on the metric that was used in the previous cycle.

The scales are based on an approach called concurrent item calibration. In concurrent item calibration all items from the current and previous cycle are used. Secured or common items between cycles ensure sufficient overlap between assessments to build the link. Typically around half of the items of previous assessment are repeated in the next assessment. The calibration sample consists only of countries that participated in both assessments. Also, in the calibration phase, data from the current assessment are used together with the data from the previous assessment. If, for example, scaling is carried out

for PIRLS 2006, all countries that participated in PIRLS 2001 and PIRLS 2006 are selected and both sets of data for these countries are used (i.e. data from PIRLS 2001 and PIRLS 2006, even though the scaling will be used for PIRLS 2006 only).

Item calibration usually consists of three steps to build a linkage between the current and previous calibration. In the first stage a set of item parameters for every item is established based on data from both the previous cycle and the current cycle (using all items from both assessments and only countries common to the current and the previous cycle). Since the sample sizes differ between countries, data are weighted to ensure that the data from each country and each assessment year contribute equally to the item calibration.

All background variables from the student questionnaire are used in conditioning. The amount of data from the questionnaire is reduced using principal component analysis (PCA). Typically, components accounting for 90% of the variance in the data are selected. The PCA is performed separately for each country and therefore different numbers of principal components are required to account for 90% of the common variance in each country's background variables.

The next step is to generate the IRT scale or achievement scores. Achievement scores are five random draws from the conditional (posterior) distribution of scale proficiencies, given the student's item responses, background variables, and model parameters for items. By including all available data in the model (conditioning), relationships between these background variables and the estimated proficiency scores are appropriately accounted for in the plausible values. Plausible values generated are initially on the same scale as the item parameters. This scale metric is arbitrary and ranges from approximately -3 to +3 with an expected mean of 0 across all countries.

When the scale is applied for the first time (as was the case in PIRLS 2001 and TIMSS 1995), the arbitrary constants for the origin and unit size are set to a mean of 500 and a standard deviation of 100. This scale avoids negative values for student scale scores and eliminates the need for decimal points in reporting student achievement (Gonzalez 1997). For comparisons between cycles, all the data from later cycles have to be placed on this metric.

After plausible values are generated, the mean and standard deviation of the latent ability distribution can be calculated and differences between distributions can be observed:

- difference in distribution of latent ability: previous cycle under concurrent calibration vs. current cycle under concurrent calibration - change in achievement;

- difference in distribution of latent ability: previous cycle under concurrent calibration vs. previous cycle under previous cycle calibration - change in item parameter estimates.

The next step is to find a linear transformation. The linear transformation is needed to adjust for the differences in item parameters arising from the fact that in the previous assessment, data were combined with the different assessment data in the calibration. The gap between both calibrations of the previous cycle data (previous and concurrent) is typically small and is due to slight differences in the item parameter estimations (because the previous assessment data was calibrated with other assessment data in the two calibrations).

The linear transformation removes this gap and transforms the distribution of the previous assessment data under concurrent calibration. However, it still preserves the gap between the previous and current cycle data under the concurrent calibration which represent the change in achievement. The final step is to apply this linear transformation to the current assessment data scaled using the concurrent calibration. With this transformation the current assessment data are placed on the same metric of the previous assessment (Olson et al. 2008).

Linear transformations are given by:

$$PV_i^* = A_i + B_i * PV_i$$

where PV_i is the plausible value i prior to transformation, PV_i^* is the plausible value i after transformation, and A_i and B_i are the linear transformation constants. The constants are obtained by the international means and standard deviations of the proficiency scores for a scale using the plausible values generated in the previous cycle for trend countries only (countries that participated in both cycles). The same calculations are carried out for the trend countries under the new (concurrent) calibration. Thus, the same data from the same countries are scaled in the previous cycle and in the current cycle. The linear transformation constants are then defined as:

$$B_i = \sigma_i / \sigma_i^*$$

$$A_i = \mu_i - B_i \mu_i^*$$

where μ_i is the international mean based on plausible value i released in the previous cycle, μ_i^* is the international mean based on plausible value i of the current cycle, σ_i is the international standard deviation based on plausible value i released in the previous cycle, and σ_i^* is the international standard deviation based on plausible value i of the current cycle.

With these constants, all of the proficiency scores from the later cycle are transformed by applying the same linear transformations for all countries. There are five sets of transformation constants for each scale, one for each plausible value. After applying this transformation the proficiency scores are on the same metric as proficiency scores in the previous cycle and are therefore directly comparable.

1.5.2 Some specific cases of scaling procedures in TIMSS and PIRLS

In TIMSS 1995, a one parameter (Rasch) model was initially used for scaling student achievement. Beaton and Robitaille (2002) argue that a 3PL could alternatively be used. This model could be expected to fit the data better than the Rasch model since multiple choice items were administrated. They also report that the Rasch model was used for TIMSS 1995 because of the availability of the required expertise and software at that time. In TIMSS 1999, data were scaled using the 3PL model, and the TIMSS 1995 data have been rescaled using the same model to achieve comparability with TIMSS 1999. Arguments can be made for either model, but Beaton and Robitaille (2002) state that the 3PL model became increasingly popular.

Nowadays (since 1999), TIMSS and PIRLS use the same IRT models and procedures for obtaining achievement scores. They use a 3PL model, a 2PL model and a generalized partial credit model. The procedure of scaling is described in more detail in the previous sections. Here we give a brief overview of the specific procedures used in the current cycles of the studies.

Foy et al. (2010) were investigating difficulties in estimating lower achievement in PIRLS 2006 for lower achieving countries. Based on observations of precision of student achievement scores they concluded that a minimum average percent correct (of 30%)

across all items of the assessment should be present otherwise there is a bias in reporting achievement. If a test is too difficult for a student, and the student cannot answer many of the items correctly, this results in overestimation of student achievement. In this situation the lowest possible score on the test may be an overestimate of the student's real achievement, compared to what the results would show on an easier test with items better suited to the student's ability. Based on their study, in the last cycle of PIRLS 2011 there was an additional module of PIRLS called prePIRLS, which was intended for populations of readers that would find the PIRLS assessment too challenging. Only three countries participated in prePIRLS, which presented challenges for the scaling.

A special scaling approach was required to make the best use of the limited data available. Because one country administered both PIRLS and prePIRLS to the same fourth grade students, it was possible to use this data as a link between the two assessments. Preliminary analyses revealed a high latent correlation (0.91) between the two assessments. Furthermore, this was considered to provide sufficient evidence of a single construct of reading achievement underlying both assessments to justify a combined scaling of PIRLS and prePIRLS. For prePIRLS the item calibration step involved a concurrent calibration of the prePIRLS data from its three countries together with the PIRLS data from all of the PIRLS 2011 countries. In this concurrent calibration, the PIRLS items had item parameters fixed at values previously estimated from the main PIRLS 2011 concurrent calibration. Based on the prePIRLS, item parameters could be placed on the same scale as the PIRLS items, and also robustness in the estimation of the prePIRLS item parameters was added (Foy et al. 2012).

Furthermore, the conditioning for prePIRLS was done in exactly the same way as for PIRLS (as already described in the previous chapter). The prePIRLS item calibration established a link between the PIRLS and prePIRLS scales, but this was done only on the basis of data from one country. For this reason, the use the PIRLS-prePIRLS link to establish the metric for the prePIRLS scale was considered insufficient. Instead, the linear transformations to determine the prePIRLS reading metric were set to produce an average of 500 and standard deviation of 100 across the three participating countries. These same linear transformations were also applied to the subdomains that were scaled separately (Foy et al. 2012).

As evident from Table 1.2 in TIMSS Advanced there were only four common countries participating in both cycles (1995 and 2008). Because of the small number of countries that participated in both TIMSS Advanced assessments, concurrent item calibrations were conducted using data from all the countries that participated in either the 1995 assessments or the 2008 assessments.

1.5.3 Item statistics and model fit evaluation in TIMSS and PIRLS

Before the application of IRT scaling an extensive item review is conducted. This is done to detect unusual item properties that could reveal a problem or error in a particular country. In case such items are found, the country's translation verification documents and printed booklets are examined for flaws or inaccuracies and, if necessary, the item was removed from the international database for that country. Furthermore an item by country interaction is observed as a graphical representation of the difference between each country's Rasch item difficulty and the international average Rasch item difficulty across all countries (Foy et al. 2012).

Specific attention is given to linking items that are common to the current and previous assessment. The main aim is to check that these items have statistical properties similar to those they had in the previous assessments. No special attention is needed if the difference between the Rasch difficulties across the two assessments for a particular country is smaller than 2 logits. Furthermore as one indicator of reliability, Cronbach's Alpha coefficient of reliability is calculated at the assessment booklet level (Foy et al. 2012).

In TIMSS and PIRLS the graphical methods to observe model fit are used. In the report about methods and procedures from TIMSS and PIRLS 2011, the procedures about model fit are described as follows:

> After the item calibrations are completed, checks were performed to verify that the item parameters obtained adequately reproduce the observed distribution of student responses across the proficiency continuum. The fit of the IRT models to the TIMSS and PIRLS assessment data is examined by comparing the item response function curves generated using the item parameters estimated from the data with the empirical item response functions calculated from the latent abilities estimated for each student that responded to the item. When the

> empirical results for an item fall near the fitted curves, the IRT model fits the data well and provides an accurate and reliable measurement of the underlying proficiency scale. Graphical plots of these response function curves are called item characteristic curves (Foy et al. 2012, 18).

1.6 Parameter invariance in item response theory

1.6.1 Definition

One of the advantages of IRT is that once its assumptions are met to a reasonable approximation by the item response data, proficiency estimates depend on neither the particular subset of items nor the particular subsample of the sample (Embretson and Reise 2000). This is noted as parameter invariance and is one of the most important features of IRT as highlighted in many books (van der Linden and Hambleton 1997; Embretson and Reise 2000). The goal of IRT is to provide both invariant item statistics and ability estimates (Hambleton and Swaminathan 1985). As the cornerstone of IRT, the importance of the invariance property of IRT model parameters cannot be overstated, because, without this crucial property, the complexity of IRT models can hardly be justified on either theoretical or practical grounds (Fan 1998).

As mentioned, the definition of invariance in IRT essentially represents two things. Firstly, an individual's latent trait level can be estimated based on the individual's responses to any set of items with known item response functions. In IRT the scaling of the latent trait does not depend on any particular set of items. The parameters of an item response function are defined with respect to the latent trait scaling. Consequently, a response to any set of items can be used to estimate an individual's location on the latent trait continuum. This aspect is usually referred to as person parameter invariance. The second aspect of invariance is item parameter invariance. It represents the idea that the difficulty and discrimination of an item does not depend on the characteristics of the sample. However, the invariance is defined only within a linear transformation (Morizot et al. 2007).

As Rupp and Zumbo (2006) report, the term invariance indicates that parameter values are identical in separate examinee populations or across separate measurement conditions. They also point out that parameter invariance denotes "an absolute ideal state that holds

only for perfect model fit" (Rupp and Zumbo (2006, 64). Brennan (2008) states that population invariance is a matter of degree.

The practical implication of the principle of invariance is that a test located anywhere along the ability scale can be used to estimate an examinee's ability. For example, an examinee could take a test that is "easy" or a test that is "hard" and obtain, on average, the same estimated ability. This is in sharp contrast to classical test theory, where such an examinee would get a high test score on the easy test, a low score on the hard test, and there would be no way of ascertaining the examinee's underlying ability. Under IRT, the examinee's ability is fixed (it has a particular value in a given context) and invariant with respect to the items used to measure it (Baker 2001).

The values of the item parameters are a property of the item, not of the group that responded to the item. Under classical test theory, this is not the case and the item parameters depend on the group of examinees. Even though in IRT the item parameters are group invariant, this does not mean that the numerical values of the item parameter estimates yielded by the maximum likelihood estimation procedure for two groups of examinees taking the same items will always be identical. The obtained numerical values will be subject to variation due to sample size, how well-structured the data are, and the goodness-of-fit of the curve to the data. Even though the underlying item parameter values are the same for two samples, the obtained item parameter estimates will vary from sample to sample. The result is that in an actual testing situation, the group-invariance principle holds but will not be apparent in the several values of the item parameter estimates obtained for the same items. In addition, the item must be used to measure the same latent trait for both groups. An item's parameters do not retain group invariance when taken out of context, i.e., when used to measure a different latent trait or with examinees from a population for which the test is inappropriate (Baker 2001).

Rupp and Zumbo (2004) state, parameter invariance is not guaranteed by the mere fact that an IRT model is fit to the data. The goal of sample-invariant calibration of items as stated by Engelhard (1994, 78) is "to estimate the location of items on a latent variable of interest that will remain unchanged across subgroups of individuals and also across various subgroups of items. (…) If the goal of sample-invariant calibration is achieved, then the item scale values will not be a function of subgroup characteristics, such as ability level, gender, race, or social class".

However, the accuracy of estimating two different trait levels from test data differs between item sets. If an item set is easy, a low trait level will be more accurately estimated than a high trait level. Similarly, if the calibration sample has relatively low trait levels, the difficulty of easy items will be more accurately estimated than hard items (Embretson and Reise 2000). Because the latent scale in IRT is arbitrary, the model parameters are invariant only up to a set of linear transformations. Rupp and Zumbo (2006) also point out that for investigating parameter invariance we need at least two examinee populations or two measurement conditions so that the parameter comparisons are meaningful. It is important to note that the invariance property of item parameters can only be investigated by administering the same items to different samples and then comparing the item parameter estimates obtained across samples.

In IRT models, different parameters are obtained to describe the item characteristic curve. According to the model used, different item parameters (such as difficulty, discrimination, guessing etc.) and person (examinee) parameters are present, which are usually represented by the levels (intensity) of the trait. Since the trait level scores (the output from IRT scaling) are arbitrary (sometimes referred to in the literature as linearly indeterminate), they are not directly comparable across groups of items or examinees. The indeterminacy of the latent trait scale is usually resolved by setting the mean and standard deviation of the latent indicator θ (typically to be distributed with a mean of 0 and a standard deviation of 1).

The invariance property of the IRT item statistics also obviates the need of equating tests; instead, (linear) scaling, rather than equating, is necessary within the framework of IRT. In case of random samples from the same population, the random samples should be comparable with each other within the limits of statistical sampling error.

Furthermore, the research in score equating, differential item functioning and item parameter drift also deals with the lack of invariance and its effects on parameter estimates (Rupp and Zumbo 2006). Differential item functioning is present when items function differently between groups that are defined on differences in examinees' individual characteristics such as gender, ethnic group, or country. Item parameter drift occurs when items function differently across examinee groups associated with separate test administrations or time points. In the following section, some of the research on invariance is presented.

1.6.2 Research on invariance in IRT

Invariance is usually examined by comparing estimated values of the parameters across different populations or conditions. The main interest is then to determine the type of relationship that exists between them in order to assess whether the same IRT model is likely to hold across the examined conditions. Rupp and Zumbo (2004) argue that parameter invariance means equality of parameters and not equality of parameter estimates. In practice it is impossible to observe parameter invariance, especially because of the arbitrary latent scale in IRT models and the inability to achieve a perfect model fit. Instead, as they further explain, the goal of studies on parameter invariance is to quantify likely degrees of lack of invariance as a continuum and not a contrasting categorical state.

Investigating invariance in IRT models is quite frequent in the literature, but in general these studies differ in focus. Some studies deal with the investigation of parameter invariance of IRT parameters in comparison to classical test theory (Fan 1998; Macdonald and Paunonen 2002; Adedoyin et al. 2008; Progar and Sočan 2008), some focus on parameter invariance within item response theory models (Galdin and Laurencelle 2010) and others focus on other procedures and content that could have an effect on parameter estimates in IRT models (Klieme and Baumert 2001; Wells et al. 2002; Michaelides and Haertel 2004; Monseur and Brezner 2007; Monseur et al. 2008; Hencke et al. 2009; Adedoyin 2010).

For the purposes of comparing estimates in investigations of invariance, a measure of linear association is usually used (e.g. Pearson's Product-Moment Correlation Coefficient). Rupp and Zumbo (2004) argue that a correlation coefficient of a large magnitude is a necessary but not sufficient condition for the parameter invariance to hold, particularly because it only measures the strength and direction of a linear relationship and, therefore, fails to detect non-linear relationships. For example it fails to capture additive shifts in parameter estimates that separate one examinee population from another at the test level. They furthermore suggest differential item functioning analyses at the scale and item level, or simulation studies that simulate likely effects for a given scenario to quantify the magnitude of introduced differences in response probabilities and test scores with the use of bias coefficients. Klieme and Baumert (2001) report (as LSA do not aim to evaluate the performance of individuals) that not all cases of differential item functioning have to be interpreted as item bias that will have an effect on the fairness of the test. In addition,

another study reports that moderate amounts of item discrimination, item difficulty parameter and joint discrimination and difficulty parameter drift have relatively minimal effect on examinees' ability estimates (Wells et al. 2002).

The empirical research on invariance seems to have been increasing over the past two decades. Research that draws on comparisons of classical test theory and IRT usually includes a subsection on invariance in IRT parameters. Many of these studies report that the difficulty parameter is more invariant than the discrimination parameter across subpopulations (Fan 1998; Macdonald and Paunonen 2002; Adedoyin et al. 2008).

Adedoyin et al. (2008) reported that estimates of item difficulty parameter based on IRT are invariant across different independent groups (gender, random samples from the population, educational regions, and ability groups). In addition, they report that such estimates are also invariant across varying sample sizes in the aforementioned groups. Macdonald and Paunonen (2002) observed that invariance in discrimination was higher when the true item discrimination values were generated from the wider distribution (0.5 to 2.5) as compared to the narrower distribution (1.0 to 2.0).

Galdin and Laurencelle (2010) state that IRT's estimate of ability is not invariant across a change in the estimation context (shift in the ability level of co-examinees or in the general difficulty level of items). The indeterminacy of θ (because it is arbitrarily centered to 0) results in the fact that estimated ability distributions are generally biased.

Cook et al. (1988) conclude from their study that the attributes that were measured by the test (in their case knowledge of biology) depend on the group to whom the test is administered. They specifically call for caution when achievement tests are administered over several points in time (during the school year). Because students who take the test may be at different stages in their coursework and the same set of items may measure different underlying concepts or dimensions, curriculum-related achievement tests have differential validity, depending on when during a student's course of study he or she chooses to take the test.

In repeating tests across time, it is a usual practice to release test items from previous tests that are no longer used in the assessments. Taylor and Lee (2010) see the releasing of items as potentially dangerous because it might lead to altering the nature of the scale. Releasing items can lead to practicing items that are similar to those on the test. Based on the results,

they concluded that the item parameters for polytomous items are less stable than for dichotomous items. However, the elimination of the unstable polytomous items does not have a serious effect on the resulting scale. Based on their study, they suggest that large-scale tests should be rescaled after several years of implementation. Furthermore, Sykes and Fitzpatrick (1992) find that the increasing difficulty of the difficulty parameter estimate in the Rasch model over time is not attributable to item position but to changes in curriculum emphasis.

Some researchers (Babcock and Albano 2012) point out that there is a possibility of item parameter drift over subsequent administrations. They investigated the stability of Rasch scales over time in certification (licensure, job related) testing. Their findings clearly indicate that the stability of the Rasch scale can maintain near baseline recovery properties if the changes in the latent trait over time are small. They strongly suggest that Rasch IRT scales eventually need to be recalibrated despite the fact that the comparability and consistency in decision making and score reporting common scales can be used.

Klieme and Baumert (2001) assumed that the Rasch scaling model holds only approximately across subpopulations of large-scale studies. They found different national profiles in learning outcomes in six countries that participated in TIMSS 1995, which can be interpreted as differential effects of cultural backgrounds and educational traditions. LSAs capture complex proficiency syndromes, which include various interacting psychological abilities and heterogeneous content components. They state that unidimensional IRT models never show a perfect fit in large samples. This misfit is generally regarded as a negligible specification error or an error variance. As LSAs do not aim to evaluate the performance of individuals, many cases of differential item functioning cannot be interpreted as item bias that lead to an unfair testing situation among countries.

Another reanalysis of TIMSS 1995 found that, in principle, the tests for advanced mathematics can be appropriately described as unidimensional (Klieme 2000). Progar and Sočan (2008) investigated the unidimensionality of IRT models in TIMSS 1995 for a small selection of mathematics and science items (a subsample of items from the item pool). They report that the assumption of unidimensionality holds to a reasonable extent in the subsample of math items but is violated in science items.

When tests are provided at different points in time with the purpose of measuring change in a latent trait, a subset of the items is usually repeated in both assessments (as already

mentioned in chapter 1.2.5 Test equating and linking procedures). These items are referred to as linking items (linking items are the same in previous and current assessments) and are used to construct the link between the previous and the current assessment. That would mean that the item properties (e.g. difficulty and discrimination) have to be comparable across different examinee groups and also across time. Moreover, it is extremely important that these items show sufficient characteristics to obtain reliable estimates. In this context, research on items is also very important.

Some authors (Hencke et al. 2009) investigated whether the selection of items in TIMSS 2003 has an effect on average student performance estimates in various countries. For estimating parameters, they included only items that a country reported were covered with their curriculum (and excluded non-covered items). They then used only these item parameters to obtain plausible values in all participating countries. Their conclusion was that relative positions of countries changes remarkably little when including only covered items (high-performing countries remained high-achieving for any item choice; low-performing countries remained low-achieving; countries in the middle remained in the middle of the achievement distribution). However, they determined that in five countries the mean score was significantly higher when including only covered items in comparison to the mean score when all items were included. Four countries would have increased their relative rank position by one position and one country even by six positions. Nevertheless, the differences in achievement scores of countries that changed the positions are exceptionally small and not significant.

Other researchers (Monseur and Brezner 2007; Monseur et al. 2008) were investigating linking errors in trend estimation for international surveys in education. Under IRT assumptions, the same linking function should be obtained regardless of selection of common items. They determined that the linking error increases as the number of trend items decreases (the uncertainty regarding trend indicators is inversely proportional to the number of link items). According to Monseur et al. (2008), tests with fewer items yield higher linking errors. There is more uncertainty at the extreme scores of the ability distribution due to the variability of the equating transformation than at the center (Michaelides and Haertel 2004). Furthermore, this leads to outcomes of lower variability for countries with low trend estimates (trends around 0) in comparison to countries with high trend estimates.

In addition to item and person invariance, comparisons across IRT models were also in the focus of past research, especially because of the practical advantages of 1PL model to other IRT models but also because choosing a model is sometimes an extremely difficult task. Invariant item comparisons in 2PL and 3PL fail to meet the same quality of invariance as expected in the Rasch model (Embretson and Reise 2000). Embretson and Reise (2000) state that only the Rasch (1PL) model can be justified by conjoint additivity and other fundamental measurement properties. Furthermore, they report that many psychometricians have the opinion that other IRT models do not provide objective measurement. However, proponents of the more complex models often point out that the Rasch model fails to fit important psychological data.

Furthermore, Rudner (1977) provided empirical evidence for a strong linear relationship between difficulty and discrimination parameter values in the 2PL and 3PL models. Brown et al. (2005) report a change of proficiency score distribution in TIMSS 1995 when the 1PL or 3PL model is used. The ranking of countries remains almost the same, and the correlation between proficiency scores obtained from different models is high. Furthermore, greater differences are observed in lower achieving countries.

Fan (1998) in his study reports that the IRT 1PL model difficulty estimates appear to be slightly more invariant across samples than the 2PL and 3PL model item difficulty estimates. The item discrimination indexes of IRT were less invariant across participant samples than the item difficulty indexes were. However, the invariance of item discrimination indexes from IRT decreased with the increase of dissimilarity between samples. The item discrimination indexes were the most invariant across random samples; they were less invariant across female-male samples (i.e. the female-male sample pair was more dissimilar than the random sample pair); they were the least invariant across high-low ability samples (i.e. the high-low ability sample pair was the most dissimilar among the three sampling conditions). However, the qualities of a theoretical model should finally always be validated through strict empirical investigations.

2 Research problem

Item response theory models are used in studies (also in LSA studies, e.g. PIRLS, TIMSS, PISA etc.) because of many advantages over classical test theory. An important feature of IRT models is that trait level estimates with invariant meaning may be obtained from any set of items and item parameters do not depend on sample characteristics (if the IRT model fits the data). This feature enables for example equating of different test forms, IRT models can be used to identify differential item functioning etc.

For example, it does not matter whether we apply a hard or an easy test on the same subject; the student with the highest trait level should have the highest expected score. The distribution of students' trait level should stay the same regardless of the difficulty of the items administered. "When a given IRT model fits the data of interest, several desirable features are obtained. Examinee ability estimates are not test dependent, and item indices are not group dependent. Ability estimates obtained from different sets of items will be the same (except for measurement error), and item parameter estimates obtained in different groups of examinees will be the same (except for the measurement error)" (Hambleton et al. 1991, 8).

Theoretically, one country in LSA studies would give sufficient information to estimate item parameters in international assessments. This is true provided the population covers the range of abilities, yet the uncertainty will differ because we would have more or less information at different points of the distribution.

The model fit of items in TIMSS and PIRLS is checked by comparing the item response functions generated using the item parameters estimated from the data with the empirical item response functions calculated from the latent abilities estimated for each student that responded to the item. In this sense, all items are checked and the items that are reported to be used in scaling are those that show good model fit (a good correspondence of empirical and theoretical data). Other than two graphs (one showing empirical and fitted curves for the polytomous item and the other for the dichotomous item) there is no information about the model fit in the technical reports. Since a model never perfectly fits the data, the question about the invariance of item parameters and proficiency scores remains (especially because the country participation is increasing and countries with different

characteristics are participating in subsequent cycles in LSA studies). We did not directly investigate how the model fit is related to parameter estimates but examined parameter estimates in different conditions under the assumption of a sufficient model fit.

The focus of the dissertation is to observe whether any changes in item parameter estimates and achievement scores are present when different countries are included in the item parameter estimations (in PIRLS and TIMSS). The focus therefore is on the calibration sample. The same question is not raised in PISA, since the countries used for item parameters estimation do not differ substantially between cycles (OECD 2005, OECD 2009). In addition, PISA uses only OECD member countries for estimating item parameters and trends. However, this question is important especially in studies conducted by the IEA since the countries differ from cycle to cycle. In addition of examining the effect of the characteristics of the sample, the invariance is compared also across different content domains and across different IRT models.

2.1 Research questions

The main purpose of the dissertation is to observe invariance in item and person parameter estimates based on the composition of the calibration sample, including that occurring in different content domains and when different IRT models are used in calibration. The effect of the calibration sample and some other calibration characteristics is observed in real data using the data from PIRLS 2006 and TIMSS 2007.

Four different research questions were investigated: firstly, the sample size of the calibration sample (with one country as "unit" in the sample); secondly, the ability of the calibration sample; thirdly, the model used in the calibrations; and finally, the content assessed. Based on the focus, four research questions were investigated.

Research question 1: Are there any differences in item parameters and proficiency scores when we include a different number of countries in the item parameter estimation?

The exclusion of a few countries in the item parameter estimation most probably does not have an effect on the achievement scores since the calibration sample is large enough and probably includes all possible ranges of abilities. Looking at the participation of countries in PIRLS, 29 were common to 2001 and 2006 and 35 were common to 2006 and 2011.

Common countries in the previous cycle represent more than half of participating countries in the current cycle. There is a high probability of having a solid and representative calibration sample and that achievement scores are expected to be reliable. If we observe the participation of countries in TIMSS for 4th and 8th grade students, we can witness the same as in PIRLS with even more countries participating. The least common participants were in 1995 and 2003 for the 4th grade population, namely 18. In all other cycles the common participation exceeded 23. This research question is of great relevance particularly because in TIMSS Advanced 2008 there were only four common countries to the previous cycle in 1995. Also, in the recent part of PIRLS (prePIRLS 2011), only three countries participated. Based on the literature review no significant differences should be observed in item parameter estimates and proficiency scores when a different number of countries (which represents the calibration sample size) is used in the item parameter estimation.

Research question 2: Does the average achievement of the included countries make a difference in terms of item parameters and proficiency scores?

The property of invariant comparison does not mean that the estimates from test data will have identical properties over either items or persons. For example, the accuracy of estimating two different trait levels from test data differs between item sets. If an item set is easy, a low trait level will be more accurately estimated than a high trait level. Similarly, if the calibration sample has relatively low trait levels, the difficulty of easy items will be more accurately estimated than hard items (Embretson and Reise 2000). Fan (1998) found that the item discrimination indexes of both classical test theory and IRT were most invariant across random samples, they were less invariant across female-male samples (i.e., the female-male sample pair was more dissimilar than the random sample pair), and they were least invariant across high-low ability samples (i.e., the high-low ability sample pair was the most dissimilar among the three sampling conditions). Based on the literature review we expect that average achievement in the calibration sample is not correlated with the magnitude of item and person parameter estimates, but it is correlated with the accuracy of the estimates.

Research question 3: Is the same invariance of item parameters and proficiency scores achieved with different IRT models?

The choice of the model usually depends on the data and on model assumptions. Estimation of the more complex models tends to be less practical and also requires larger sample sizes. The model in TIMSS changed from 1995 to 1999 from a Rasch model to a 3PL model because the later model could be expected to fit the data better than the Rasch model (also because multiple choice items were administrated).

As Rupp and Zumbo (2003) point out, the choice of appropriate model can also be a matter of tradition. In PISA, a Rasch model is used and in TIMSS and PIRLS a 3PL model is used. Wu (2010) reports that there are no clearly documented findings which would give sufficient information about comparisons of these two models. From a theoretical point of view, the 3PL model additionally takes into account the discrimination as well as the guessing parameter in multiple choice items and in this manner, provides more information about the items.

The assumption of same discrimination of items across countries in LSAs does not seem very reasonable. If test items are different in terms of discrimination power, it is possible that the two scaling methods could produce different spreads of the student ability distributions (Wu 2010). Nevertheless, the ranking of countries is likely to remain unchanged provided that items do not exhibit differential item functioning across countries. Klieme and Baumert (2001) assumed that the Rasch scaling model holds only approximately across subpopulations of large-scale studies. They found different national profiles in learning outcomes in six countries that participated in TIMSS 1995 which can be interpreted as differential effects of cultural backgrounds and educational traditions.

The advantage of simpler models is that fewer parameters have to be estimated, and very large samples are not required for stable parameter estimation (although sample size in international LSAs is not an issue since the sample size is usually very large). It is easier to interpret them meaningfully from a substantive theoretical viewpoint, and this may be what is desired. Arguments can be made for either model, however, Beaton and Robitaille (2002) state that the trend points towards using the 3PL model (complex models were not widely used before because they are more computationally consuming which, with the modern technology, does not present a burden anymore). The focus of this research question is whether there are differences in achievement scores depending on whether a Rasch model or a 3PL model is used in calibration of items. The model fit is not investigated separately because the technical documentation of the data supports a

sufficient model fit for all included items, and because the Rasch model is expected to express an even better model fit than the 3PL model (because of less included parameters). Based on the research reported in the literature review and the assumption of a sufficient model fit, we do not expect a difference in proficiency scores for countries when different models are used.

Research question 4: Is the same invariance of item parameters and proficiency scores achieved in different content domains (knowledge of mathematics and reading literacy)?

The cognitive items when testing a student's literacy (PIRLS) or knowledge of mathematics (TIMSS) are different. From a student's perspective, both assessments consist of two parts that are presented in one booklet for each student. In PIRLS, one booklet consists of two reading passages (regardless of the type or purpose of text) and in TIMSS, one part of the booklet is usually devoted to mathematics and the other to science items. Furthermore, all questions in one part of PIRLS are related to the same passage or text.

PIRLS assessment material, as well as that of other international assessments of reading literacy such as PISA, is hierarchically structured, which means that several items relate to a single context (Monseur et al. 2011). One of the assumptions in IRT models is that of local item independence. As Monseur et al. (2011) warn, because multiple items are connected together to a common passage, items within a unit are not likely to be conditionally independent, so the assumption might be violated. The violation of the local item independence assumption can have substantial consequences on test parameter estimates and on proficiency estimates. In their research, Monseur et al. (2011) found that the consequence of local item independence violation in PISA is that the relative variability of low-performing countries is overestimated while the relative variability of high-performing countries is underestimated.

The same questions about local independence of items do not occur in TIMSS since mathematics and science assessment includes only one item per stimulus. The structure of presentation of cognitive items in TIMSS and PIRLS differs. For the purpose of comparing the differences across content domains we chose reading and mathematics. As reading and mathematics are regarded to be more comparable across countries than science (science content in TIMSS assessment includes content from biology, chemistry, earth science and physics; these subjects are taught in some countries as an integral subject and in some countries as separate subject), only items from reading and mathematics content were

observed. Based on previous research, we expect that mathematics content shows greater invariance compared to the reading domain.

3 Methods

3.1 Data sets

In order to address the research questions, we used data from PIRLS 2006 and TIMSS 2007, since these were the most current cycles of studies at the time of starting the simulations (the data from the most current cycles of TIMSS and PIRLS 2011 were released in December 2012). The databases were downloaded from the respective study webpages (http://timssandpirls.bc.edu/pirls2006/user_guide.html, http://timssandpirls.bc.edu/TIMSS2007/idb_ug.html).

Multiple choice item responses were recoded in both studies such that zero represented an incorrect answer and one represented a correct answer (the chosen stimulus is presented in the original database). Furthermore, the original (and recoded multiple choice) values for the items were recoded separately for item parameter estimation and for drawing plausible values. In the item parameter estimation, items that were not reached were assigned the code 6, and those that were not administered were assigned the code 8. Not reached items are items that were not reached by examinees due to time reasons (in IEA the first three sequential omitted responses are coded as missing and if more values are missing after that they are coded as not reached). Typically these items occur at the end of an instrument. If not reached items are known, these items may be ignored for each examinee when making statistical inferences about item parameters (Lord 1980). These items do not contain any quantifiable information about the examinee's proficiency. Therefore only observed responses are used. In contrast, omitted items are items that the examinees read and decided not to answer for whatever reason. Omitted items (code 9) and other missing values were treated as incorrect. In drawing plausible values, only items that were not administered were assigned code 8. Not reached items, omitted and other missing items were treated as incorrect.

In PIRLS 2007, 40 countries and five Canadian provinces participated. The provinces were treated as separate educational systems, so there were 45 educational systems in total. Separate educational systems and countries will be referred to hereafter as countries. From

TIMSS 2007 we took a subsample of countries since we wanted to compare the results to PIRLS. For this purpose we selected only countries that were in common with PIRLS 2006. Thus, the data set for TIMSS 2007 includes only 29 countries (including Canadian provinces) even though more countries participated.

To facilitate the tracking of countries, all original country codes were recoded to three digit numbers (Canadian provinces had four digit codes in original files). Four country codes were changed (9132 to 912, 9133 to 913, 9134 to 914, 9135 to 915, and 9136 to 916).

In PIRLS 2006, there were a total of 125 items included in the item parameter estimations. One administered item was excluded due to poor characteristics across countries (according to the PIRLS 2006 technical report the scaling did not converge - item discrimination was too low for many countries). Furthermore, there were 62 constructed response items (of which 28 had a maximum of 2 points, six had a maximum of 3 points and 28 had a maximum of 1 point) and 63 multiple choice items (with 4 options: A, B, C or D). For all items we estimated the difficulty and discrimination parameters. Additionally, the guessing parameter was estimated for 63 items, the step parameters d_1 and d_2 were estimated for 34 items, and the step parameter d_3 was estimated for six items.

In TIMSS 2007, a total of 179 mathematics items were administrated. Two multiple choice items were excluded from the mathematics item pool due to poor item characteristics (according to TIMSS 2007 technical report because of faulty distracters) across countries (the items included in scaling are reported on the study's website). Overall in mathematics, 94 multiple choice items (with 4 options: A, B, C or D), 11 constructed response items with a maximum of 2 points, and 72 with a maximum of 1 point were included in the item parameter estimation. For all (177) items we estimated the difficulty and discrimination parameters. Additionally, the guessing parameter was estimated for 94 multiple choice items, the step parameter d_1 was estimated for 83 constructed response items with a maximum of 1 and 2 points, and the step parameter d_2 was estimated for items with a maximum of two points.

In general we used a 3PL model for multiple choice items, a 2PL for constructed-response items with just two scoring options, and a generalized partial credit model for polytomous constructed response items (more than two response options). The only exception was when investigating different models (the models used in this part are described in chapter 3.6)

In addition to the items (the procedure in both studies was the same) two background variables were included for comparison of subgroup score estimates. These were students' gender, and the number of books at home (there was no specific reason to choose these two variables in particular); both variables were taken from the student background questionnaire. In the gender variable, code 1 represents a girl and code 2 represents a boy. Since there were very few missing values for gender, the missing values were replaced by a random number between 1 and 2 and then the number was rounded to the nearest whole number (either 1 or 2). In PIRLS 2006, there were 118 missing values for gender (from 215 137 students); 60 values were assigned to girls and 58 to boys. In TIMSS 2007, there were 24 missing values for gender (from 100 885 students) and the same procedure was used to replace the missing values as in PIRLS. As a result, 16 values were assigned to girls and 8 to boys.

In number of books at home, the five answer options were 0-10 books, 11-25 books, 26-100 books, 101-200 books and more than 200 books. For conditioning the variable number of books was recoded into several new variables. Because of the missing values in every country (PIRLS: from 1% to 41%; TIMSS: from 0.4% to 19%) in the variable, five dummy variables were created for the number of books and the variable gender was recoded as 0 for girls and 1 for boys.

3.2 Reference scores

Reference score estimates were obtained from the full data set when all (45 in PIRLS and 29 in PIRLS and TIMSS) countries were included in the item parameter estimation in such a way that each country contributed to the parameter estimation equally (in all estimations the senate weight (SENWGT) was used; the weights in each country added up to 500). Firstly we estimated the item parameters. Then we used these item parameters to obtain five plausible values for each student in each country. In order to make the plausible values comparable, each plausible value variable for all countries together was standardized to have a mean of 500 and a standard deviation of 100 (the same procedure was also followed in all other conditions). Estimates of average were obtained from these values for each country separately and were used as reference (baseline) scores for countries to which other results were compared to (every time by country).

The reference achievement scores for countries (for all reference scores that were used: PIRLS with 45 countries, PIRLS with 29 countries and TIMSS with 29 countries) are presented in Table 3.1 together with the original sample size for every country. Since we have five achievement scores (plausible values) for every student, the country achievement is presented as the arithmetic mean of the statistic of interest (in our case, statistics of interest were arithmetic mean achievement, different percentiles, arithmetic mean achievement by gender and arithmetic mean achievement for each category of number of books at home within each country). This procedure follows the recommendations of von Davier et al. (2009) that summarizing results using the plausible values requires calculating the statistic of interest using each of the plausible values, and then finally averaging the results.

Table 3.1: Country's average achievement for all reference conditions with corresponding number of students for PIRLS and TIMSS

Country ID	*N*	Reference condition		
		PIRLS (45)	PIRLS (29)	TIMSS (29)
40	5067	531	523	510
100	3863	540		
158	4589	528	520	575
208	4001	540	532	527
250	4404	514		
268	4402	465	455	443
276	7899	541	533	529
344	4712	556	549	604
348	4068	543	535	514
352	3673	505		
360	4774	399		
364	5411	419	408	412
376	3908	507		
380	3581	544	536	511
414	3958	337	325	328
428	4162	533	525	539
440	4701	529	521	532
442	5101	550		
498	4036	493		
504	3249	326	313	354
528	4156	539	532	538
554	6256	525	517	498
578	3837	494	485	480
616	4854	513		
634	6680	350	336	310

Country ID	*N*	Reference condition		
		PIRLS (45)	PIRLS (29)	TIMSS (29)
642	4273	485		
643	4720	557	550	545
702	6390	550	543	598
703	5380	524	516	501
705	5337	514	505	507
710	14657	303		
724	4094	506		
752	4394	543	536	508
780	3951	431		
807	4002	437		
840	5190	533	525	532
912	3988	548	540	516
913	3748	526	518	523
914	4243	553	545	510
915	4150	551	543	510
916	4436	535		
926	4036	532	524	544
927	3775	521	513	500
956	4479	540		
957	4552	493		

Note. Countries are ordered according to their country code (lowest to highest). *N* = number of students included in the database in countries.

3.3 Variation in reference condition for PIRLS (45)

Most of the simulations were done using PIRLS data that included a pool of 45 countries. To investigate the variation of this reference condition, the item parameter estimation for the reference condition was repeated, each time excluding one of the countries (a jack-knife procedure). In this respect we obtained 45 replications and the variance or the standard deviation of the estimates could be observed. Item parameter estimates were compared, together with countries' average achievement, percentiles and subgroup estimates. This was done to obtain a better sense of the effect that one country has within all countries. The result could also be interpreted as whether 45 countries represent a large enough sample for further analyses (in the case the variation is small) or not (in the case that the variation is large).

3.4 Procedures for including a different number of countries

The strategy we followed was to define five conditions that differed in the number of countries that were included in the item parameter estimations. The number of included countries varied from two to ten of the participating countries (2, 3, 4, 6, and 10 countries). Because of the matrix booklet design there was a large amount of missing data for items in every country. In general, each item had around 80% missing values, which were missing by design. Moreover, this means that in combination with the weight, each item had a total of 200 answers in the two country condition, 300 in three, 400 in four, 600 in six and 1000 in the ten country condition.

These conditions were selected because of the sample guidelines in the literature. Estimation of more model parameters requires larger samples. Du Toit (2003) reports that sample sizes of 250 examinees are marginally acceptable in research applications and that 500 to 1000 should be suitable in operational use for any model. Sample sizes beyond 1000 are not necessary since the additional precision may not justify the additional computational time or data collecting costs.

For each condition, countries were selected at random and the estimation of item parameters and proficiency scores was repeated 100 times to provide information about the variation in the results. In all item parameter calculations, SENWGTs were included. Based on the item parameters estimated, we obtained five plausible values for each student in each of the 45 countries for each replication.

In the first step, we compared item parameters. Comparing average item parameters would not give meaningful information since item parameter estimates are relative/adjusted to the input information (in other words they are not on the same scale). Therefore, we observed the correlation between reference item parameters and new item parameters in each condition. For every condition the arithmetic mean of correlations between new item parameters and reference item parameters is reported. The reported correlation represents the correlation between the same type of item parameters (difficulty, discrimination, guessing and step parameters) averaged across 100 replications.

In the second step, we compared each country's achievement scores. We compared newly obtained overall achievement scores (gained in different conditions) against their reference

estimates for every country. The initial scale is arbitrary; however, for the purpose of comparable scores the plausible values were standardized to have a mean of 500 and a standard deviation of 100. Plausible values are not meant for individual student comparison and they were standardized in every condition. For this reason it is not sensible to compare the distributions or individual achievement scores. Hence, we compared the mean and percentiles of the achievement score distribution of each country (5^{th}, 10^{th}, 50^{th}, 90^{th} and 95^{th}). In addition, two background variables were included in conditioning. This enables comparisons of achievement scores also across the categories of these two variables (gender and the number of books). This is why also achievement scores for different subgroups were calculated.

The estimate of one country (within one condition) represents the arithmetic mean of the estimate of five plausible values. The estimate can be the average achievement score, percentile or average achievement score for two categories in gender or five in number of books variable. As the final result, the mean root squared difference (MRSD) is reported, which represents the arithmetic mean of the root squared differences (between the new and reference scores) across the 100 replications within a condition for every country. Since we were not interested in individual scores across countries (and the design is made to obtain reliable group estimates) we compared the aggregated scores. MRSD is the difference between the average country achievement that was newly calculated and the average country achievement in the reference condition averaged across 100 repetitions.

3.5 High and low achieving countries

In the next step, estimation of item parameters was based on the inclusion of countries regarding their average achievement. Countries were first sorted by their average achievement as determined in the reference condition. Based on this result, countries were divided in three groups with an equal number of countries in each group. The groups were represented by the upper third of countries (15 countries), middle third of countries (15 countries), and lower third of countries (15 countries). Only the upper third (high achieving) and lower third (lower achieving) of countries groups were selected for further comparisons. In this respect we chose two different conditions. Within each condition we

randomly sampled 10 out of 15 countries. The procedure was repeated 100 times for each condition.

The general procedure to obtain and compare achievement scores was the same as already described for "a different number of countries". Furthermore, item parameter estimates from the two conditions were compared with the reference ones and the MRSD of country average achievement scores and percentiles was calculated. Comparisons of subgroup estimated scores were also made (for gender and the number of books) for each of the two selected conditions.

3.6 Different models

Differences between IRT models were also observed. Two conditions were defined. In the first step, items were calibrated using the 1PL or Rasch model (multiple choice items and constructed-response items with just two scoring options) and partial credit model (polytomous constructed response items with more than two response options). In the second step, items were calibrated using the 2PL (for constructed-response items with just two scoring options), 3PL (for multiple choice items) and generalized partial credit model (for polytomous constructed response items with more than two response options). For each condition, ten countries (out of 45) were selected at random and the procedure was repeated 100 times. The general procedure to obtain achievement scores is the same as already described for "a different number of countries".

In this step there are only some common item parameters, namely difficulty (because a 1PL model assumes that discrimination of items is 1, and does not include the guessing parameter), and step parameters from the (generalized) partial credit model. This is why only these four item parameters were compared. The MRSD of country average achievement scores and percentiles were also calculated. Furthermore, the subgroup score estimates were compared (for gender and the number of books) for each of the two presented conditions.

3.7 Content domains

For the purpose of comparing content domains, only common countries that participated in PIRLS 2006 and TIMSS 2007 were selected. In total, 29 countries participated in both studies. Only the sample of 4th grade students in TIMSS was chosen since 4th grade students also participated in PIRLS. Two conditions are represented by two content domains, reading (PIRLS) and mathematics (TIMSS). For each condition we independently and randomly selected ten countries which were included in the item parameter estimations. For item calibration, 2PL, 3PL and generalized partial credit models were used. Within conditions, countries for the calibration sample were sampled at random and the procedure was repeated 100 times for each condition.

The general procedure to obtain and compare achievement scores was the same as already described for "a different number of countries". The new achievement scores were compared to the reference condition (for PIRLS a new reference condition was determined based on only 29 common countries with TIMSS) for each study and each country; MRSD values were calculated for country average achievement scores, percentiles and subgroups. The MRSDs were compared across countries and content domains since PIRLS gives information about reading and TIMSS informs mathematics knowledge.

3.8 Procedures

The procedures in different conditions were repeated for a certain number of times, and the countries were always selected at random (a certain number of countries were selected from all countries, from different achievement groups or from different content domains). In this manner, in each repetition, item parameter and achievement score estimates (which are represented by the arithmetic mean of the statistic of interest across five plausible values; in our case, the average achievement score for countries and subgroups) were obtained. The final result represents the arithmetic mean of the estimates across 100 repetitions with the standard deviation within a condition. Although in international LSAs the standard error consists of the sampling error and imputation error, in our case we only took the sampling error into account. The model fit of items was not checked, since this

was already done by the international study centers. The items used in scaling (as reported in each dataset website) showed sufficient model fit to be included in further procedures.

For scaling data and drawing plausible values (which represent achievement score estimates), we used the same procedures that are used by the international study center. For item parameter estimations, PARSCALE 4.1 (Scientific Software International 2003) was used, and for generating proficiency scores, we used DESI (Direct Estimation Software Interactive 2009). From the IRT models, we used the 1PL, 2PL, 3PL and the (generalized) partial credit models. The parameters of the response models in PARSCALE were estimated using marginal maximum likelihood. In the solution of the likelihood equations, the expectation maximization (EM) algorithm was used.

The logistic models were used with the constant of 1.7. The number of expectation and maximization cycles was set to 500. This was done to increase the likelihood that the convergence criterion was met and not stopped prior to obtaining a stable solution. The precision level for the estimation convergence was set to 0.001. For all models used, we set the number of quadrature points to 30. In the Rasch version of the models, the slope parameters were additionally fixed at a mean of 1 and standard deviation of 0.00001 (because the standard deviation of the slope parameter distribution on normal metric must be positive).

The person parameter θ (achievement score) was estimated in DESI rather than PARSCALE. Achievement scores are five random draws from the conditional (posterior) distribution of scale proficiencies, given the students' item responses, background variables, and model parameters for items. Conditioning assumes that the item parameters are fixed at the estimates found in scaling and fit the latent regression model to the data, i.e. estimates Γ and Σ in its first part. Maximum likelihood estimates of Γ and Σ under unknown θ_is are obtained using an EM algorithm. To estimate the E step (in the EM algorithm), there are two options for integral approximation (approximation of the posterior mean and variance) in DESI: Laplace approximation and numerical quadrature integration. When the dimension of θ_i is greater than two, a Laplace approximation is used; when the dimension is equal or less than two, numerical quadrature integration is used (von Davier et al. 2007). In our case, quadrature integration was used. In all estimations, the quadrature points were set to 41 (from -5 to 5). In the second part of the conditioning stage, plausible values (multiple imputations) for all examinees were obtained. In

conditioning, dichotomous background variables were used (one for gender and five for the number of books) together with the item parameters from PARSCALE. Instead of directly estimating a student's θ, a probability distribution for a student's θ is estimated (the posterior distribution for a student). Plausible values are random draws from this (estimated) distribution for a student's θ.

Item parameters were compared using Pearson's product moment correlation. In the significance tests, a Fisher's z transformation was applied. The comparisons of MRSDs were made with one-way analysis of variance (ANOVA) following standard procedures. Firstly, we checked whether the variances among groups differed. The test of homogeneity of variance was carried out using Levene's test. Based on the result, we either conducted one-way ANOVA (if variances did not differ significantly) or calculated Welch's F statistic (if variances were significantly different). If ANOVA or Welch's F statistic resulted in a significant difference among groups, further post-hoc tests were performed (in the case of one-way ANOVA, Tukey's honestly significant difference (HSD) test and in the case of Welch's F statistic, Games-Howell's test). In the case of only two groups, one-way ANOVA was replaced by the t-test statistic. For all analyses, we used IBM SPSS Statistics for Windows, version 20 (IBM Corp. 2011) and R (R Development Core Team 2010). Unless stated otherwise, all significance tests performed were two sided and based on α level of 0.05.

In addition to the results of correlation coefficients, ANOVA and t-tests, an effect size is reported for all analyses. In case of the correlation coefficient, an effect size of q, in case of ANOVA, eta square (η^2) and in case of a t-test, Cohen's d (d) are respectively reported.

Sensitivity Power Analysis was performed with G*Power (Faul et al. 2007). We used the program to calculate the critical population effect size as a function of α, 1-β, and N for the correlation coefficient, t-test and one-way ANOVA. In case of an α=0.05 and 1-β=0.80, the critical values for effect sizes are:

- Correlation coefficient for comparisons of item parameters within PIRLS: N_1=125, N_2=125, q=0.32 (difficulty and discrimination parameters); N_1=63, N_2=63, q=0.45 (guessing parameter); N_1=34, N_2=34, q=0.63 ($step_1$ and $step_2$ parameters); N_1=6, N_2=6 , q=2.03 ($step_3$ parameter);

- Correlation coefficient for comparisons of item parameters between PIRLS and TIMSS: N_1=125, N_2=177, q=0.29 (difficulty and discrimination parameters); N_1=63, N_2=94, q=0.41 (guessing parameter); N_1=34, N_2=83, q=0.52 ($step_1$ parameter); N_1=34, N_2=11, q=0.99 ($step_2$ parameter);

- ANOVA: $F(4,220)=2.41$, $f=0.23$ (for the case of five groups), $F(2,132)=3.06$, $f=0.27$ (for the case of three groups),

- T-test: $t(88)=1.66$, $d=0.53$ (for the case of PIRLS with 45 countries), $t(56)=1.67$, $d=0.66$ (for the case of PIRLS and TIMSS with 29 countries).

In the social sciences, specifying the effect size is the most difficult aspect of power analysis. This is at least partly due to the relatively scarce empirical evidence regarding magnitudes of effect sizes in the disciplines. Since there is not enough evidence in the investigated field operational definitions of "small", "medium", and "large" values of each effect size index were used as recommended by Cohen (1992) to provide the reader with some sense of the magnitude of effect sizes.

For the correlation (q), the small, medium, and large effect sizes used were represented by the values 0.10, 0.30, and 0.50 respectively. To test if the two population means are equal (d), the cut points for effect sizes were 0.20, 0.50, and 0.80 and for the analysis of variance test (η^2) 0.01, 0.06, and 0.14 respectively (Cohen 1992).

In the last part of the results, we observed the relationship of achievement and MRSD and standard deviation of MRSD. For this purpose, linear and quadratic regression models were fitted to the data. The regression coefficients are presented together with model estimates. The model fits were compared with ANOVA.

3.9 Simulation results

In general, during the item parameter estimation phase using PARSCALE we encountered three problems, because of which item parameter estimations could not be completed:

1. The iterative procedure did not converge after 500 cycles;

2. Estimation sometimes failed during the EM algorithm when the information matrix could not be inverted;

3. Estimation sometimes failed during the EM algorithm because of other reasons.

We excluded these repetitions from our analyses and performed additional repetitions. The number of non-converged, non-invertible and other unsuccessful cases across all used conditions is presented in the Table 3.2 (within a condition first 100 successful repetitions were taken).

Table 3.2: Unsuccessful runs in item parameter estimation

Condition	Non-converged	Non-invertible	Other	Total
Different number of countries				
2	32	54	96	182
3	38	37	74	149
4	21	24	33	78
6	5	12	12	29
10	3	3	10	16
Countries by achievement				
Low	5	36	114	155
High	3	18	1	22
Different models				
Rasch PCM	10	1	3	14
3PL 2PL GPCM[a]	3	3	10	16
Different content domains				
Reading	1	7	16	24
Mathematics	0	4	3	7
Variation – reference[b]				
Reading	0	3	5	8

Notes: [a] for the 3PL 2PL GPCM model the results from the condition of different number of countries (with ten included countries were used)

[b] In this condition the procedure was repeated only 45 times since there were 45 countries.

The most unsuccessful runs can be observed in the condition of two countries. A lot of runs were also not successful in the condition of lower achieving countries. When the number of included countries was increased, the number of unsuccessful runs decreased. In the conditions of six countries, ten countries, high achieving countries, reading, and mathematics there were less than 30 unsuccessful runs. We did not further explore the unsuccessful runs although the information is important and should be used in the interpretation of the results.

4 Results

Firstly some descriptive data on the used items and variables is presented for both included studies, PIRLS and TIMSS. Furthermore the variation in the reference condition in PIRLS is presented, since this reference condition was used in most of the research question (in three out of four). The structure of this section follows the sequence of the research questions.

4.1 Data description – PIRLS 2006

The data from the PIRLS 2006 study was used. In PIRLS 2006 there were a total of 125 items included in the item parameter estimations (one administered item was excluded because of poor characteristics across countries). There were 62 constructed response items. From the constructed items, 28 had the maximum of 1 point, 28 a maximum of 2 points and six a maximum of 3 points. 63 of the included items were multiple choice items. Average correct response rates according to the item type can be seen in Table 4.1.

Table 4.1: Average percentage of responses in categories of items across countries for PIRLS 2006 (45 countries)

Category	MC	CR_1	CR_2	CR_3
0	6.7	8.6	7.4	5.4
1	12.9	11.0	5.3	5.2
2	/	/	6.7	3.6
3	/	/	/	4.9
Not reached (6)	0.3	0.3	0.6	0.8
Not administered (8)	80.1	80.2	80.1	80.1

Notes: MC=multiple choice items, CR_1=constructed response items with maximum of 1 point, CR_2=constructed response items with maximum of 2 points and CR_3=constructed response items with maximum of 3 points.

Since not all students respond to all items, there is a different number of responses for every item. On average there were 80% missing responses on items regardless of the item type. Furthermore, we observed an average of 7% of incorrect responses across items. The percentage of (partially) correct responses varies from 11 to 13% across different types of

items. Very few students (less than 1%) failed to reach all items and therefore their answers are categorized as “not reached”.

Average percent of correct responses across countries is presented in the International report (Mullis et al. 2007, 311). Countries percent correct was ranging from 21 to 69% with the international average of 54%. Only nine countries were showing less than 50% of correct answers on average.

In Table 4.2, the percentages across categories for the included background variables are presented. The values are presented for each country separately.

Table 4.2: Percentages across categories for gender and number of books in countries for PIRLS 2006

Country ID	Gender		Number of books at home					
	Girl	Boy	Missing	0-10	11-25	26-100	101-200	More than 200
40	49.5	50.5	2.8	11.3	23.7	36.6	13.9	11.8
100	49.4	50.6	3.5	24.0	18.1	23.5	14.5	16.5
158	47.7	52.3	2.7	17.8	22.0	29.3	14.0	14.2
208	51.6	48.4	2.0	7.3	17.2	34.8	22.3	16.5
250	48.5	51.5	6.1	8.9	16.5	29.7	18.8	20.0
268	48.0	52.0	10.6	12.3	19.9	26.1	13.4	17.7
276	49.0	51.0	9.1	6.4	19.9	31.7	16.1	16.8
344	48.7	51.3	2.8	18.3	21.4	29.5	15.6	12.3
348	50.3	49.7	2.4	9.1	21.5	31.9	17.8	17.4
352	49.9	50.1	4.1	4.2	15.6	36.0	21.5	18.6
360	49.1	50.9	20.3	41.0	25.1	8.4	2.9	2.3
364	46.4	53.6	14.3	50.4	20.0	8.3	3.5	3.5
376	48.2	51.8	11.8	9.9	18.7	30.0	14.5	15.1
380	48.4	51.6	1.9	13.5	27.1	28.3	13.9	15.3
414	49.8	50.2	41.1	15.2	16.6	12.8	6.7	7.7
428	48.1	51.9	2.7	7.8	16.3	37.5	19.5	16.3
440	48.7	51.3	2.3	11.3	27.3	33.5	15.1	10.5
442	49.1	50.9	0.9	8.0	16.5	28.0	21.0	25.6
498	49.6	50.4	1.8	34.4	31.0	20.6	7.0	5.3
504	47.3	52.7	13.8	53.0	17.1	9.4	3.5	3.3
528	50.9	49.1	2.0	10.0	25.1	36.2	15.9	10.8
554	49.2	50.8	5.5	8.8	15.0	30.3	20.5	19.9
578	49.4	50.6	9.6	5.9	13.9	30.5	19.3	20.9
616	51.4	48.6	4.1	10.9	27.3	32.5	13.3	11.9
634	49.6	50.4	21.8	16.2	13.1	15.4	10.9	22.6
642	48.1	51.9	6.0	30.0	25.8	23.0	8.5	6.7
643	50.8	49.2	1.1	8.4	21.5	36.1	15.9	17.0

Country ID	Gender		Number of books at home					
	Girl	Boy	Missing	0-10	11-25	26-100	101-200	More than 200
702	48.2	51.8	1.9	8.2	17.9	37.5	20.0	14.6
703	48.9	51.1	1.8	9.6	18.4	37.1	20.2	12.9
705	48.1	51.9	1.9	8.3	22.5	38.1	17.3	12.0
710	51.5	48.5	22.5	37.0	16.3	11.8	6.1	6.4
724	49.3	50.7	5.4	11.7	22.0	30.8	15.0	15.2
752	47.8	52.2	3.3	4.3	13.9	34.6	22.6	21.4
780	49.4	50.6	6.6	17.5	23.5	27.1	11.8	13.5
807	48.6	51.4	14.2	16.2	26.0	26.7	9.0	7.9
840	50.6	49.4	4.7	11.7	19.9	31.5	16.6	15.6
912	48.8	51.3	3.5	7.9	14.5	31.5	20.7	21.8
913	49.4	50.6	6.3	10.3	20.4	32.9	17.8	12.2
914	48.4	51.6	3.8	7.3	13.8	31.4	21.5	22.1
915	50.2	49.8	4.3	6.2	15.5	32.3	22.3	19.5
916	49.2	50.8	5.2	6.8	13.2	30.8	21.1	22.8
926	49.6	50.4	2.2	9.5	15.7	29.1	20.9	22.6
927	50.7	49.4	2.6	11.1	17.9	29.4	20.0	19.2
956	49.9	50.1	1.8	8.0	19.5	36.3	19.8	14.6
957	49.8	50.3	7.6	10.0	17.7	29.6	18.2	16.8
Total	49.2	50.8	6.8	14.6	19.6	28.6	15.6	14.8

The distribution for gender in all countries was between 48% and 50% for girls and between 50% and 52% for boys. On average there were 7% missing responses across countries in the number of books at home variable. In one country, 41% of students did not respond to this question. On average, students most frequently reported that they had between 26 and 100 books at home but the distribution of responses varied across countries.

4.2 Data description – TIMSS 2007

In the investigation of invariance across content domains, we additionally used the data from TIMSS 2007. However, we included only countries that participated in both studies (PIRLS 2006 and TIMSS 2007). Thus, 29 countries that participated in TIMSS 2007 were included and the investigation was carried out on mathematics items only. Overall, 177 items assessing different mathematics contents were used. Of these, 94 were multiple choice items and the rest were constructed responses with either 1 or 2 score points.

Table 4.3: Average percentage of responses in categories of items across countries for TIMSS 2007 (mathematics items)

Category	MC	CR_1	CR_2
0	6.0	7.1	6.2
1	8.0	6.8	3.1
2	/	/	4.6
Not reached (6)	0.3	0.4	0.3
Not administered (8)	85.8	85.8	85.8

Notes: MC=multiple choice items, CR_1=constructed response items with maximum of 1 point, CR_2=constructed response items with maximum of 2 points.

Items used in TIMSS were also assembled in blocks. The sampling frame remained similar as in PIRLS. Thus, there were more missing values for items. On average the data showed 86% missing responses for every item. Less than 0.5% of students failed to answer all items in the time available and their responses were categorized as "not reached". The percentage of incorrect answers varied between 6% and 7% across different item types and the percentage correct varied between 7% and 8%.

According to the TIMSS 2007 international mathematics report (Mullis et al. 2008, 405), if we only take into account the 29 selected countries, the average percent correct of countries ranges from 18 to 77%. Average percent correct across all countries is 51%. Seven countries showed an average percent of correct responses that is less than 50%. All other countries achieved at least 50% correct on average or higher.

Table 4.4 shows the background characteristics for two variables, gender and the number of books at home, for every country that was included in TIMSS 2007.

Table 4.4: Percentages across categories for gender and number of books in selected countries for TIMSS 2007

Country ID	Gender		Number of books at home					
	Girl	Boy	Missing	0-10	11-25	26-100	101-200	More than 200
40	48.2	51.8	1.6	11.2	28.0	34.0	13.2	12.0
158	48.4	51.6	1.2	15.8	24.4	31.3	13.2	14.0
208	51.2	48.8	4.2	8.2	22.4	36.5	16.8	11.8
268	47.0	53.0	5.2	16.6	22.8	27.1	12.2	16.2
276	49.0	51.0	14.2	7.2	21.6	30.4	14.4	12.2
344	48.6	51.4	1.6	16.2	22.0	33.3	14.8	12.2
348	50.6	49.4	2.2	10.2	24.0	31.7	16.2	15.6
364	49.0	51.0	2.0	51.5	24.8	12.0	5.2	4.4
380	48.8	51.2	1.2	14.2	30.5	30.3	11.8	12.0
414	51.6	48.4	19.0	17.8	24.2	19.6	7.8	11.6
428	47.8	52.2	2.2	7.6	21.0	40.5	16.0	12.6
440	48.8	51.2	1.2	15.2	35.3	33.3	8.6	6.4
504	49.2	50.8	13.8	45.7	20.0	11.2	4.6	4.8
528	48.0	52.0	4.2	8.4	23.6	38.8	14.2	10.8
554	49.6	50.4	1.4	9.6	17.8	33.4	21.2	16.6
578	49.8	50.2	2.4	7.2	22.4	36.6	18.8	12.6
634	51.2	48.8	13.6	16.6	16.8	21.6	12.4	19.2
643	50.0	50.0	0.4	10.0	25.9	38.5	14.0	11.2
702	48.6	51.4	1.2	10.2	21.2	37.0	18.0	12.4
703	48.8	51.2	2.0	11.2	31.8	35.4	11.6	8.0
705	49.4	50.6	2.2	8.8	29.1	37.1	13.0	9.8
752	50.2	49.8	2.4	6.4	20.2	34.4	20.2	16.4
840	51.0	49.0	1.8	13.2	20.6	33.4	15.8	15.2
912	48.2	51.8	1.8	6.0	18.8	33.3	22.6	17.4
913	51.2	48.8	1.4	10.8	23.2	38.9	14.8	10.8
914	48.0	52.0	1.6	5.6	17.2	35.7	22.2	17.6
915	49.0	51.0	2.6	5.6	17.4	36.0	20.2	18.2
926	48.8	51.2	1.2	9.0	17.0	33.2	21.2	18.4
927	50.6	49.4	1.0	11.8	19.4	32.6	18.4	16.8
Total	49.3	50.7	3.8	13.4	22.9	32.0	14.9	13.0

As can be seen from the Table 4.4, boys and girls were approximately equally represented in every included country. On average, in TIMSS 2007 countries, students most frequently reported that they had between 26 and 100 books at home. The least frequent category on average was the one with the least number of books (0-10), although in one country almost half of the students reported to have between zero and ten books at home. The average missing response rate across countries was 4% and varied between 0.4% and 19%.

4.3 Variation in reference condition – reading (45 countries)

In order to gain a better insight into the data, we conducted a jack-knife procedure to obtain an estimate of the variability in the reference condition. We repeated the scaling procedure 45 times (with 44 countries), each time taking one country out of the calibration sample. The result can also be used to gain information if the number (45) of countries is sufficient in size to represent a good starting point to sample different countries. In eight of the runs, item parameters could not be estimated and therefore results are presented for 37 repetitions. The item parameter correlations were very stable for all item parameter estimates. It was evident that the estimates showed very little variability in every item parameter. The correlations are almost perfect in all parameters (slope, location, $step_1$, $step_2$ and $step_3$: *Minimum*=1.00, Q_1=1.00, *Median*=1.00, *Mean*=1.00, Q_3=1.00, *Maximum*=1.00, *SD*=1.00; asymptote: *Minimum*=0.99, Q_1=1.00, *Median*=1.00, *Mean*=1.00, Q_3=1.00, *Maximum*=1.00, *SD*=1.00).

In Table 4.5 descriptives for average MRSD and standard deviation of MRSD in the reference condition are presented. The descriptives of statistics of interest were calculated based on average values (for 45 countries in one repetition) obtained from 37 repetitions.

Table 4.5: Descriptives of MRSD when investigating variation in the reference condition

Statistic of interest	*M*	*Min*	*Max*
Mean by country			
M	0.05	0.01	0.48
SD	0.06	0.01	0.62
Percentiles of countries distributions			
M(5pct)	0.21	0.10	1.88
M(10pct)	0.17	0.07	1.54
M(50pct)	0.06	0.03	0.23
M(90pct)	0.09	0.06	0.15
M(95pct)	0.11	0.07	0.18
SD(5pct)	0.25	0.10	2.42
SD(10pct)	0.20	0.06	1.97
SD(50pct)	0.06	0.02	0.28
SD(90pct)	0.10	0.04	0.16
SD(95pct)	0.11	0.05	0.16
Mean by country by gender (category)			
M(girl)	0.04	0.01	0.39
M(boy)	0.05	0.02	0.57
SD(girl)	0.05	0.01	0.50

Statistic of interest	*M*	*Min*	*Max*
SD(boy)	0.07	0.01	0.74
Mean by country by number of books (category)			
M(missing)	0.10	0.03	0.83
M(0–10 books)	0.08	0.03	0.50
M(11–25 books)	0.05	0.02	0.28
M(26–100 books)	0.04	0.01	0.22
M(101–200 books)	0.05	0.01	0.31
M(more than 200 books)	0.04	0.01	0.25
SD(missing)	0.12	0.03	1.06
SD(0–10 books)	0.09	0.03	0.65
SD(11–25 books)	0.05	0.02	0.36
SD(26–100 books)	0.04	0.01	0.28
SD(101–200 books)	0.05	0.01	0.42
SD(more than 200 books)	0.05	0.01	0.33

Note. Each entry represents the average of a statistic of interest among 37 replications. *M*=mean, *Min*=minimum, *Max*=maximum, *SD*=standard deviation.

The average MRSD across countries was 0.05 (*Minimum*: 0.01, *Maximum*: 0.48) with an average standard deviation of 0.06 (*Minimum*: 0.01, *Maximum*: 0.62). The absolute differences to the reference (when all countries were included in the calibration sample) were very small and on average did not exceed 0.25 points (on a 500 point scale). Overall the observed differences were very small. This result showed that with 45 countries we can obtain a very stable solution or invariant results of item and person parameter estimates across the limited number of included countries in PIRLS 2006.

4.4 A different number of countries

The first research question was dealing with the sample size of the calibration sample. Table 4.6 shows descriptive statistics for the correlation between item parameters for different conditions and reference item parameters.

Table 4.6: Correlation characteristics of item parameters across conditions

Nr. of countries[a]	Parameter	*Min*	Q_1	*Me*	*M*	Q_3	*Max*	*SD*
2	slope	0.60	0.76	0.80	0.79	0.84	0.90	0.06
	location	0.62	0.85	0.87	0.87	0.90	0.95	0.05
	asymptote	0.26	0.45	0.51	0.52	0.59	0.75	0.11
	$step_1$	0.41	0.92	0.94	0.91	0.96	0.98	0.10
	$step_2$	0.38	0.92	0.94	0.91	0.95	0.98	0.10
	$step_3$	-0.19	0.63	0.75	0.68	0.84	0.98	0.26
3	slope	0.27	0.81	0.85	0.83	0.87	0.92	0.10
	location	0.54	0.88	0.91	0.89	0.93	0.96	0.07
	asymptote	0.30	0.50	0.57	0.58	0.67	0.80	0.11
	$step_1$	0.81	0.94	0.95	0.95	0.97	0.99	0.03
	$step_2$	0.81	0.93	0.95	0.94	0.96	0.98	0.03
	$step_3$	-0.13	0.67	0.79	0.75	0.88	0.99	0.20
4	slope	-0.05	0.85	0.89	0.87	0.90	0.93	0.10
	location	0.78	0.91	0.94	0.93	0.95	0.98	0.03
	asymptote	0.29	0.57	0.66	0.64	0.71	0.81	0.11
	$step_1$	0.74	0.96	0.97	0.96	0.98	0.99	0.03
	$step_2$	0.74	0.96	0.96	0.96	0.98	0.99	0.03
	$step_3$	0.05	0.74	0.84	0.80	0.90	0.99	0.17
6	slope	0.82	0.90	0.92	0.91	0.94	0.96	0.03
	location	0.85	0.94	0.96	0.95	0.96	0.98	0.03
	asymptote	0.40	0.65	0.73	0.70	0.76	0.89	0.09
	$step_1$	0.92	0.97	0.98	0.98	0.98	0.99	0.01
	$step_2$	0.92	0.97	0.98	0.97	0.98	0.99	0.01
	$step_3$	0.35	0.76	0.88	0.83	0.92	0.99	0.13
10	slope	0.64	0.94	0.95	0.95	0.96	0.98	0.03
	location	0.76	0.97	0.97	0.97	0.98	0.99	0.03
	asymptote	0.51	0.76	0.81	0.81	0.87	0.95	0.08
	$step_1$	0.97	0.98	0.99	0.99	0.99	1.00	0.01
	$step_2$	0.97	0.98	0.99	0.99	0.99	0.99	0.01
	$step_3$	0.63	0.86	0.93	0.90	0.95	1.00	0.07

Note. Each entry represents the statistic of interest among 100 replications. *Min*=minimum, Q_1=first quartile, *Me*=median, *M*=mean, Q_3=third quartile, *Max*=maximum, *SD*=standard deviation of the correlations among 100 repetitions. [a]Represents the number of countries that were included in the calibration sample.

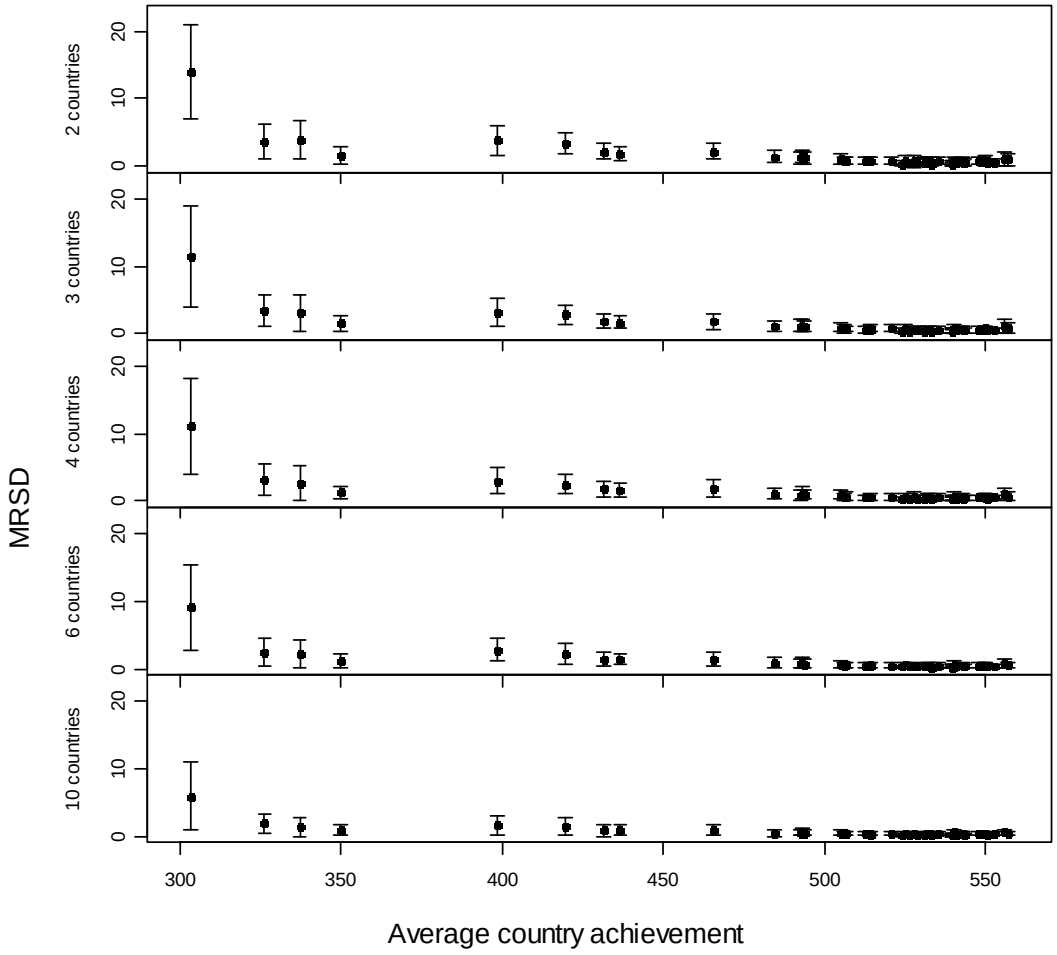
The lowest correlation coefficients were observed in the guessing (asymptote) parameter estimates. Moreover, the correlation increased when more countries were included in the item parameter estimation. After including ten countries the correlation was 0.81 and was significantly higher than in the condition with four countries (z=2.02, p=.043, q=0.37), three countries (z=2.54, p=.011, q=0.46) and two countries (z=3.02, p=.003, q=0.55). Other conditions did not differ significantly from the condition with ten included countries.

All correlations within the discrimination (slope) parameter were above 0.79. The correlation in the last condition (with ten countries) was significantly larger than in all the previous conditions (z>2.38, p<.017, q>0.30). The same held true for the difficulty (location) parameters where the correlations in all conditions were large (above 0.87), and again the correlation in the last condition was significantly larger than in all the previous conditions (z>2.03, p<.042, q>0.26). The correlations in $step_1$ and $step_2$ parameters were generally large (above 0.91) and differences across correlations in the $step_1$ parameter between the last condition compared with other conditions were statistically significant (z>2.76, p<.006, q>0.70, except the condition with six countries: z=1.37, p=0.17, q=0.35). In the $step_2$ parameter, the correlation in the condition with ten countries was significantly larger than in all other conditions (z>2.18, p<.029, q>0.55). We did not observe any significant differences in correlations in $step_3$ parameter, which was most probably due to the small number of items (six) with this parameter (0.28<q>0.64). However, in general, the correlations were moderate (above 0.68) and the effect sizes were medium and large.

To summarize the results, in general the correlations between new and baseline item parameters were large, although the largest correlations were found for the difficulty item parameter, moderately smaller for the discrimination parameter, and smallest for the pseudo guessing item parameter. We also observed that in general, the correlation increased whenever more countries were included in the item parameter estimation. Moreover, when ten countries were included in the calibration sample, all of the average correlations between item parameters (except for guessing) were above 0.90. From this, we can conclude that item parameters are relatively invariant under the investigated conditions and are almost the same as the baseline parameters when ten countries are included in the item parameter estimation.

Based on the estimated item parameters, we obtained achievement scores for each student in each country following standard procedures for international studies. In Figure 4.1 we present the reference achievement scores of countries and the MRSDs between reference achievement scores and the new calculated scores for different conditions and for each county. The result for every condition represents the MRSD averaged over 100 replications and the standard deviation of these estimates across replications.

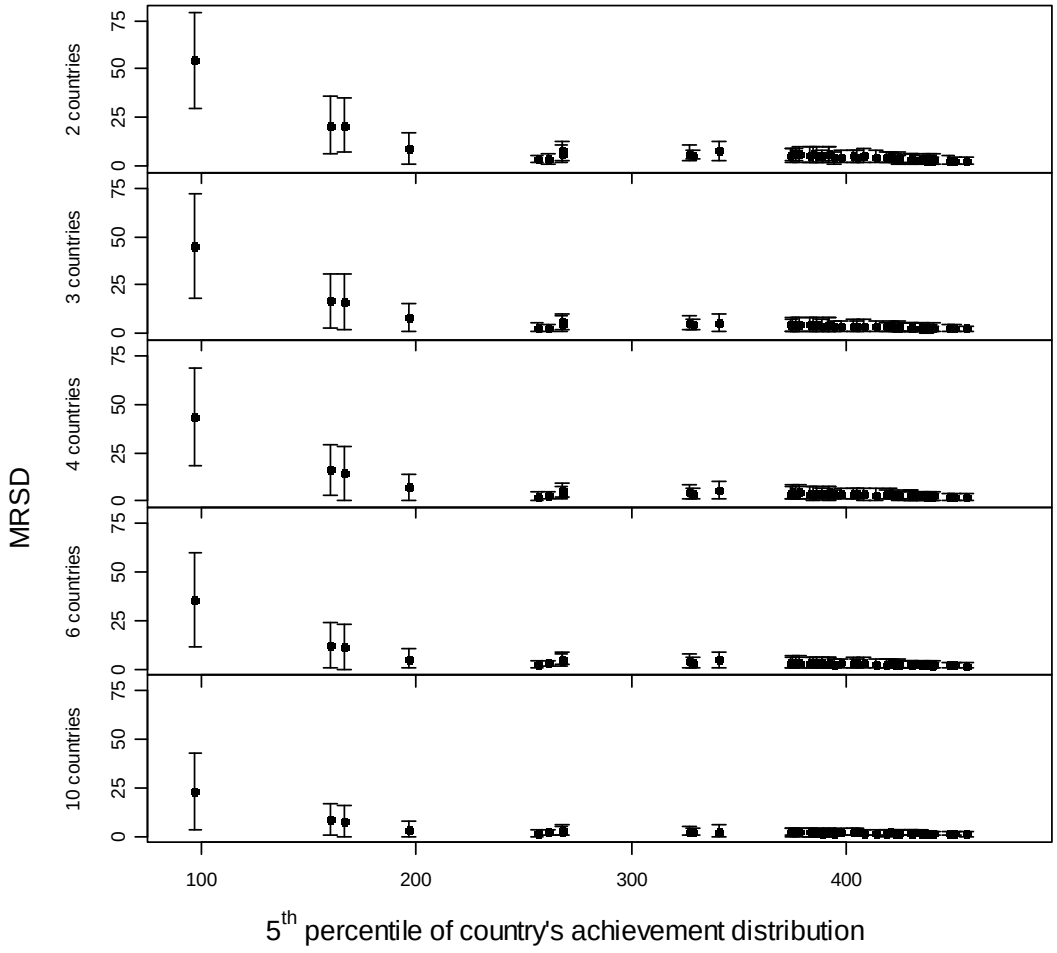
Figure 4.1: Reference achievement scores of countries with the corresponding MRSD and standard deviation of MRSD in different conditions



The average achievement scores of countries showed less variability and smaller values of MRSD when more countries were included in the item parameter estimation. From Figure 4.1 we can observe a clear pattern that the lower achieving countries showed greater MRSD values and greater variability in MRSD values (with one exception, the country with average achievement of 350 points).

The variances of mean MRSD values do not differ significantly between conditions (F=1.213, p=0.306). The results of one-way ANOVA show that the differences between conditions are not statistically significant and the effect size is small (F(4,220)=1.48; p=.209, η^2=0.026).

Figure 4.2: Reference 5th percentile of countries with the corresponding MRSD and standard deviation of MRSD in different conditions

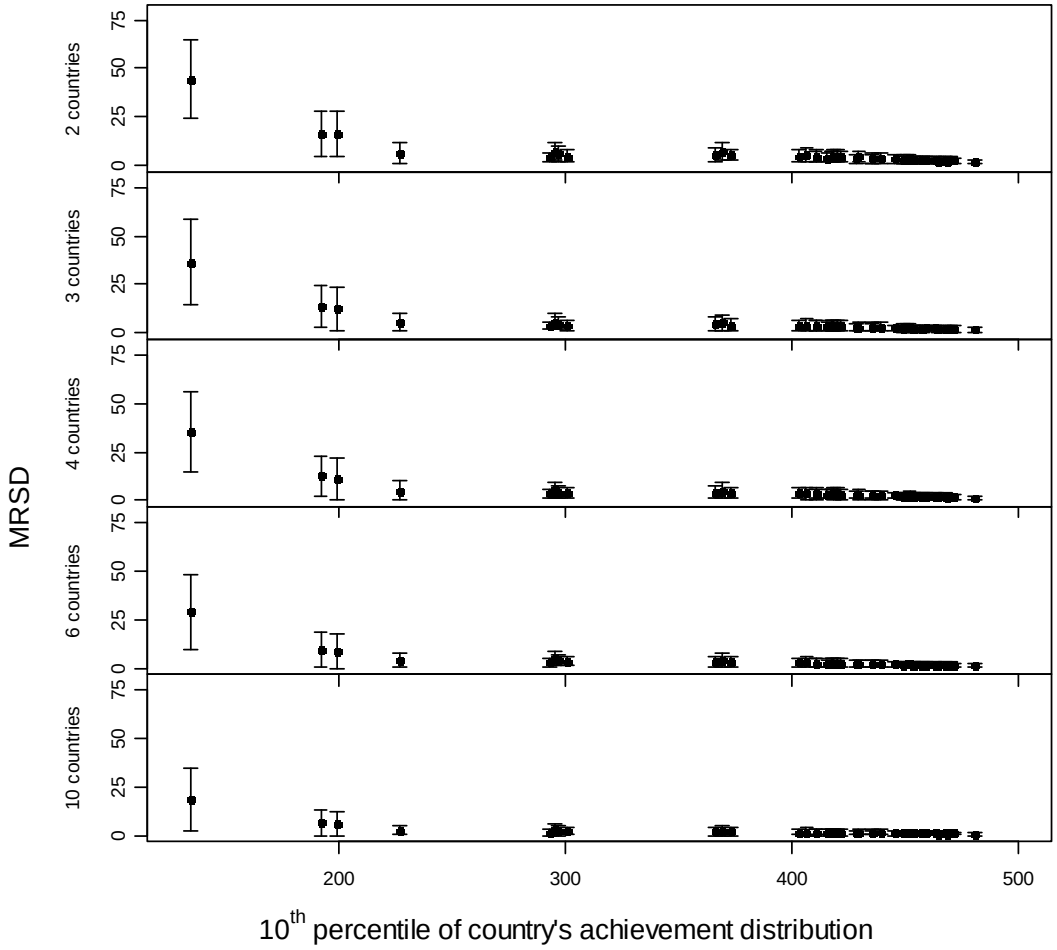


The results in Figure 4.2 show the same pattern as for the average achievement scores. We can observe that the MRSD values decreased when more countries were included in the calibration sample. Countries with the 5th percentile below 200 showed greater variability of the estimate and the variability decreased in all countries in conditions with more countries.

The variances of the mean MRSD values did not differ significantly between conditions (F=0.941, p=0.441). The differences in means between conditions were found to be significant (F(4, 220)=2.549, p=0.040, η^2=0.044). When observing the post hoc test, only the condition of two countries differed significantly from the condition with ten countries

(*Mean difference*=3.88, *p*=0.026, *d*=0.62). The MRSD of countries was significantly lower in the condition of ten countries than in the condition of two countries. No other pair of conditions showed significant differences.

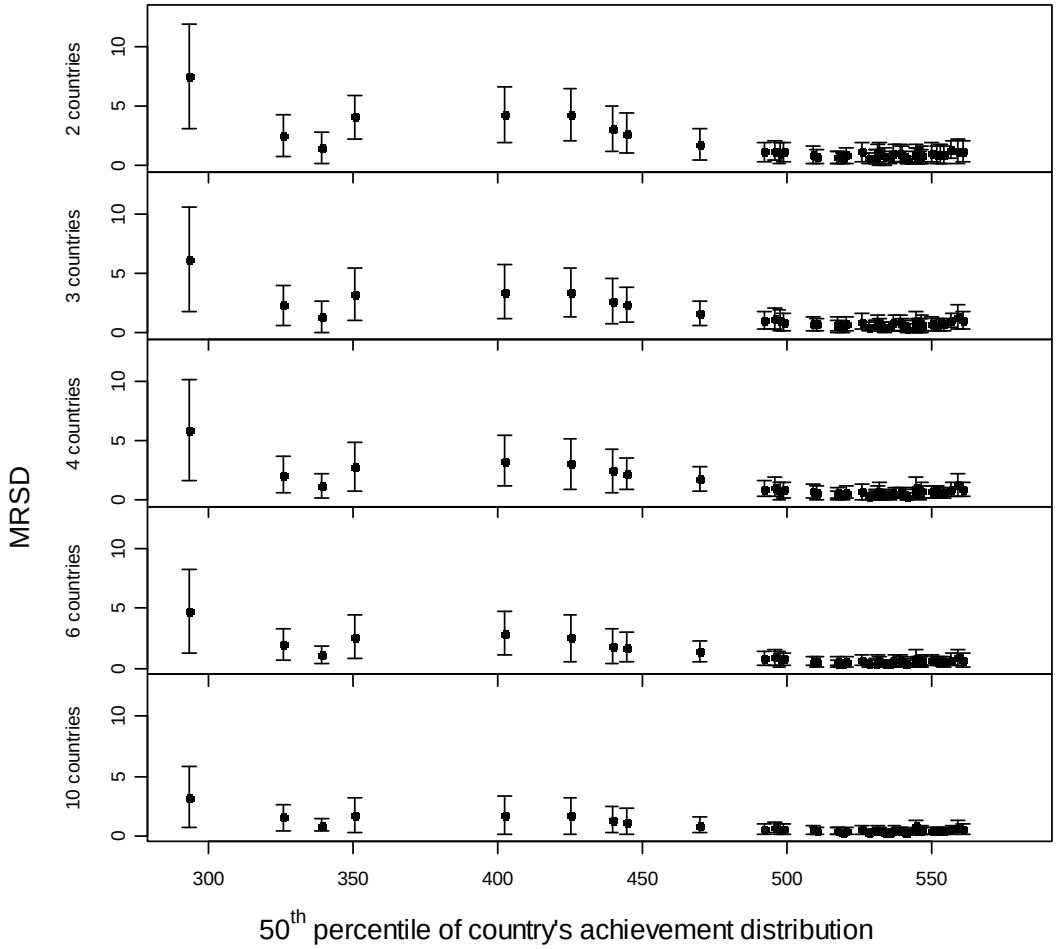
Figure 4.3: Reference 10th percentile of countries with the corresponding MRSD and standard deviation of MRSD in different conditions



Countries with the 10th percentile below 200 showed greater variability in the estimate and also greater values of MRSD. As can be seen from Figure 4.3 in general the MRSD of countries decreased across conditions with more countries and the same held true for variability.

The variances of mean MRSD did not differ significantly between conditions (F=0.984, p=0.417). The mean values between conditions were not significantly different (F(4, 220)=2.396, p=0.051, η^2=0.042).

Figure 4.4: Reference 50th percentile of countries with the corresponding MRSD and standard deviation of MRSD in different conditions

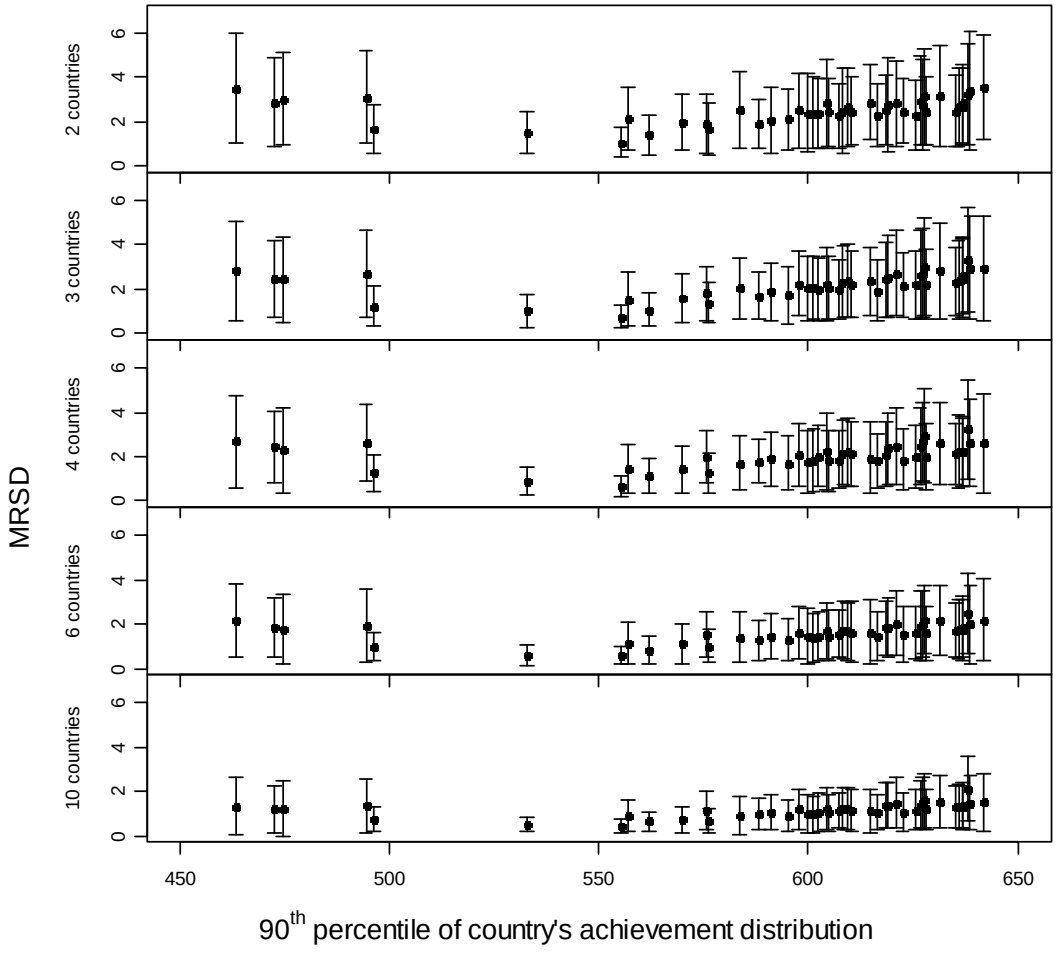


In the 50th percentile (Figure 4.4), all of the countries' MRSD values were below 8. Again the MRSD of countries decreased across conditions with more countries and the same was observed for variability.

The variances of means did not differ significantly between conditions (F=2.294, p=0.060). The mean values were found to be significantly different between conditions

($F(4,220)=3.671$, $p=0.006$, $\eta^2=0.063$). After considering the results from post hoc test, only the difference between the two- and ten-country conditions was found to be significant (*Mean difference*=0.75, $p=0.004$, $d=0.75$). The MRSD of countries in the condition of ten countries was significantly lower than in the condition of two countries. No other pair of conditions showed significant differences.

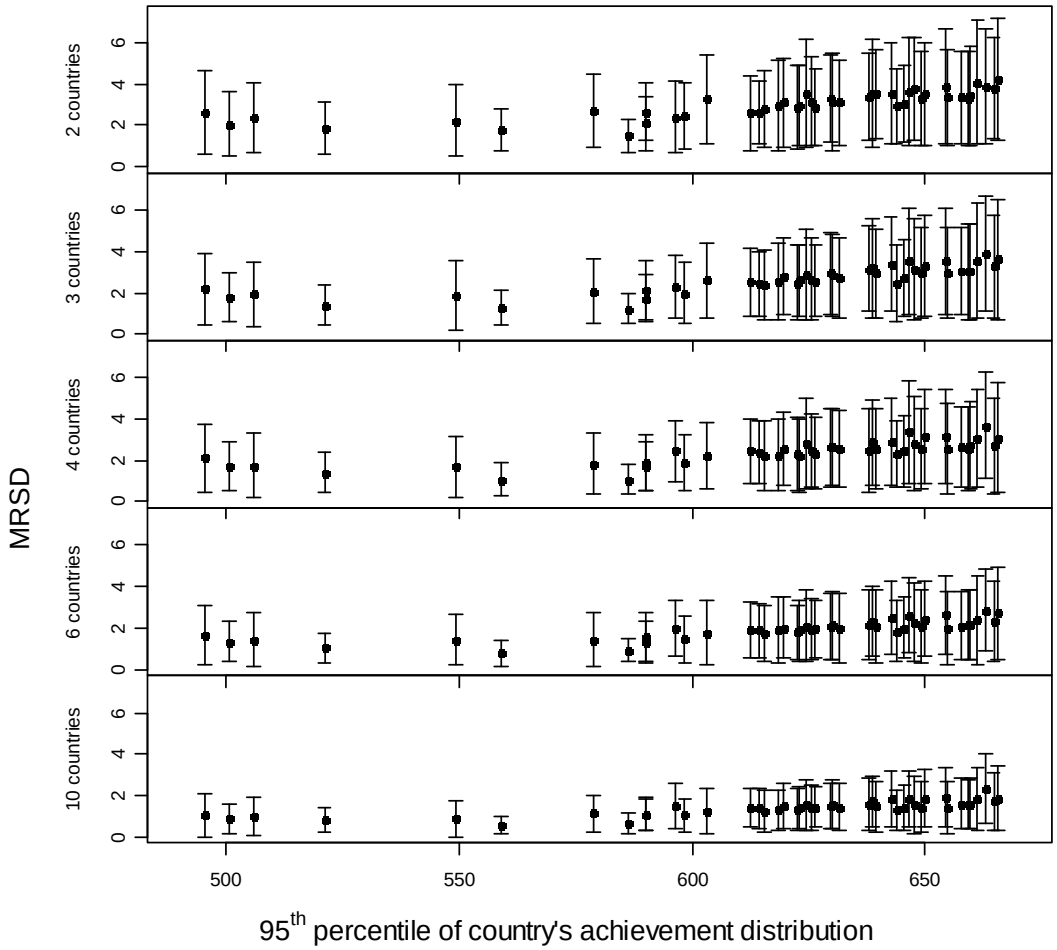
Figure 4.5: Reference 90th percentile of countries with the corresponding MRSD and standard deviation of MRSD in different conditions



The variances of means differed significantly between conditions ($F=3.339$, $p=0.011$). The differences in mean values between conditions were found to be significant (*Welch's* $F(4,108)=67.855$, $p<0.001$). Results of Games-Howell post hoc test revealed only one non-significant difference. The condition of three countries did not significantly differ from the

condition of four countries, the effect size was small (*Mean difference*=0.15, *p*=0.677, *d*=0.30). All the other pairs of conditions showed significant differences (0.304<*Mean difference*<1.32, *p*<0.001, 0.55<*d*<2.95). In the ten country condition significantly lower values of MRSD were observed than in other conditions.

Figure 4.6: Reference 95th percentile of countries with the corresponding MRSD and standard deviation of MRSD in different conditions



The variances of MRSD of 95th percentile differed significantly between conditions (*F*=4.349, *p*=0.002). The differences in mean values between conditions were found to be significant (*Welch's F*(4,108)=78.803, *p*<0.001). Results of Games-Howell post hoc test revealed only one non-significant difference among conditions. The condition of three countries did not significantly differ from the condition of four countries, the effect size

was small (*Mean difference*=0.24, p=0.300, d=0.41). All other pairs of conditions showed significant differences (0.37<*Mean difference*<1.63, p<0.046, 0.59<d<3.18). In the ten-country condition, significantly lower values of MRSD were observed than in all other conditions.

Furthermore, we were also interested in differences in MRSD in the background variables of gender and number of books at home. The results are presented in Table 4.7 and Table 4.8.

Table 4.7: Descriptives of MRSD by gender across conditions

Category	Nr. of countries	*M*	*Me*	*SD*	*Min*	*Max*
girls	2	1.3	0.8	1.7	0.4	11.4
	3	1.1	0.7	1.4	0.3	9.3
	4	1.0	0.6	1.4	0.3	9.2
	6	0.9	0.5	1.1	0.3	7.4
	10	0.6	0.4	0.7	0.2	4.9
boys	2	1.7	0.9	2.6	0.4	16.8
	3	1.4	0.8	2.1	0.4	13.7
	4	1.3	0.7	2.0	0.3	13.2
	6	1.1	0.6	1.7	0.3	10.8
	10	0.8	0.5	1.1	0.2	7.1

Note: M=mean, *Me*=median, *SD*=standard deviation, *Min*=minimum, *Max*=maximum.

The variances between conditions did not differ significantly for either the category of girls (F=1.006, p=0.405) or the category of boys (F=1.320, p=0.263). There were no significant differences in MRSD scores in any of the investigated conditions. The differences in MRSD scores were also found to be insignificant for both girls (F(4,220)=1.732, p<0.144, η^2=0.031) and boys (F(4,220)=1.411, p=0.231, η^2=0.025). However, in both cases the effect size was small in magnitude.

Table 4.8: Descriptives for MRSD for number of books across conditions

Category	Nr. of countries	*M*	*Me*	*SD*	*Min*	*Max*
missing	2	3.2	2.6	3.5	0.9	24.1
	3	2.7	2.2	2.8	0.8	19.6
	4	2.6	2.1	2.7	0.9	18.9
	6	2.0	1.7	2.2	0.6	15.5
	10	1.4	1.2	1.4	0.5	10.1
0 – 10 books	2	2.4	2.0	2.1	0.8	14.9
	3	2.1	1.7	1.7	0.7	12.1
	4	2.0	1.6	1.6	0.7	11.8

Category	Nr. of countries	*M*	*Me*	*SD*	*Min*	*Max*
	6	1.6	1.2	1.3	0.5	9.6
	10	1.1	0.9	0.9	0.4	6.3
11 – 25 books	2	1.4	1.1	1.3	0.4	8.3
	3	1.2	0.9	1.0	0.4	6.8
	4	1.2	0.9	1.0	0.4	6.7
	6	0.9	0.7	0.8	0.3	5.4
	10	0.7	0.5	0.5	0.2	3.6
26 – 100 books	2	1.1	0.8	1.0	0.4	6.1
	3	0.9	0.6	0.8	0.4	5.1
	4	0.9	0.6	0.8	0.3	5.1
	6	0.7	0.5	0.7	0.3	4.2
	10	0.5	0.4	0.4	0.2	2.7
101 – 200 books	2	1.2	0.9	1.4	0.5	9.3
	3	1.1	0.8	1.2	0.4	7.8
	4	1.0	0.7	1.1	0.3	7.5
	6	0.9	0.6	0.9	0.3	6.2
	10	0.6	0.4	0.6	0.2	4.0
more than 200 books	2	1.2	0.9	1.1	0.4	7.3
	3	1.1	0.8	0.9	0.3	6.1
	4	1.0	0.8	0.9	0.3	5.9
	6	0.8	0.6	0.8	0.2	4.8
	10	0.6	0.5	0.5	0.2	3.2

Note: M=mean, *Me*=median, *SD*=standard deviation, *Min*=minimum, *Max*=maximum.

The variances did not differ significantly across conditions for any category ($0.726<F<2.059$, $0.087<p<0.575$). The differences in mean MRSD values between conditions were found to be significant in all but one category and small or medium in size (missing: $F(4,220)=3.013$, $p=0.019$, $\eta^2=0.052$; 0-10 books: $F(4,220)=4.699$, $p=0.001$, $\eta^2=0.079$; 11-25 books: $F(4,220)=4.216$, $p=0.003$, $\eta^2=0.071$; 26-100 books: $F(4,220)=3.652$, $p=0.007$, $\eta^2=0.062$; more than 200 books: $F(4,220)=3.599$, $p=0.007$, $\eta^2=0.061$). The difference in 101-200 books category was not significant across conditions and the effect size was small ($F(4,220)=2.310$, $p=0.059$, $\eta^2=0.040$).

The only significant difference in the category "missing" was between the condition with ten countries and the condition with two countries (*Mean difference*=1.77, $p=0.014$, $d=0.66$). The MRSD in the ten country condition was significantly lower than in the two-country condition the effect was of medium size.

The only significant differences in the category of "0-25 books" were between the condition with ten countries and the condition with two and three countries (*Mean difference*=1.32, $p=0.001$, $d=0.83$ and *Mean difference*=0.99, $p=0.027$, $d=0.74$). The

MRSD in ten-country condition was significantly lower than in the two- and three-country conditions. The size of the effect was medium.

The only significant differences in category of "26-100 books" were between the condition with ten countries and the conditions with two and three countries (*Mean difference*=0.76, *p*=0.002, *d*=.79 and *Mean difference*=0.57, *p*=0.039, *d*=0.70). The MRSD in the ten-country condition was significantly lower than in the two and three country conditions. This effect can be considered medium in size.

The only significant difference in the category of "101-200 books" was between condition with ten and the condition with two countries (*Mean difference*=0.63, *p*=0.045, *d*=0.59). The MRSD in the ten-country condition was significantly lower than in the two-country condition. The same was true for the category of "more than 200 books". The MRSD in ten-country condition was significantly lower than in the two-country condition (*Mean difference*=0.63, *p*=0.005, *d*=0.74). Both effect sizes were medium in magnitude.

4.5 Low and high achieving countries

The next research question was about countries' average achievement in the calibration sample. We investigated the item and person parameters under two conditions: when only lower achieving countries were included in the calibration sample, and when only higher achieving countries were included in the calibration sample. The countries were sorted by their achievement and three equal groups of countries were formed (15 countries in each group). Only the group with "high" achieving and "low" achieving countries were used in the calibration phase. The range of average achievement scores in the "low" achieving group was between 303 and 506 points and in "high" achieving countries between 539 and 557 points. The results are presented in more detail on the following pages.

Table 4.9: Correlation characteristics of item parameters across conditions

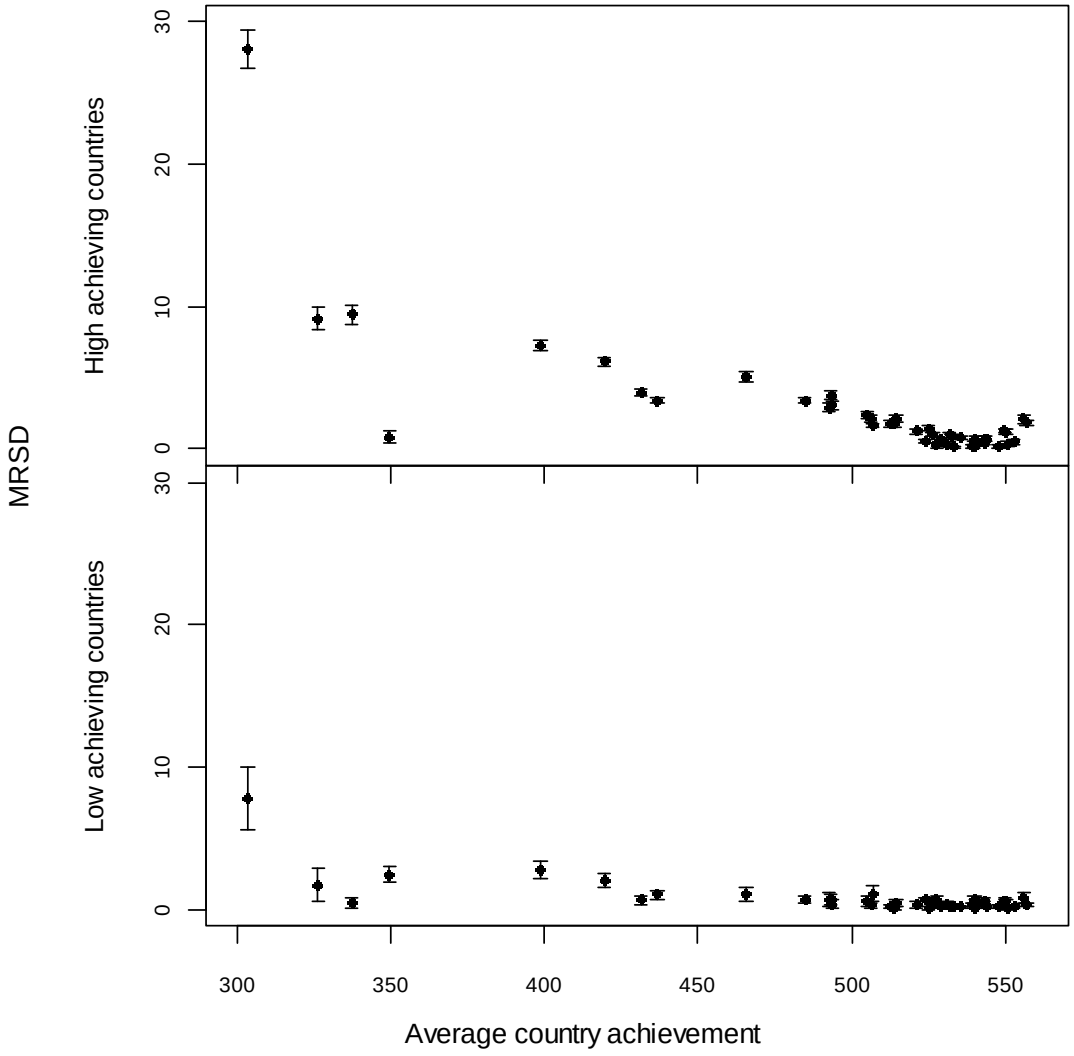
Condition[a]	Parameter	*Min*	Q_1	*Me*	*M*	Q_3	*Max*	*SD*
High								
	slope	0.92	0.93	0.93	0.93	0.94	0.95	0.01
	location	0.89	0.93	0.94	0.94	0.95	0.96	0.01
	asymptote	0.60	0.64	0.67	0.67	0.70	0.75	0.03
	$step_1$	0.97	0.98	0.98	0.98	0.99	0.99	0.00

Condition[a]	Parameter	*Min*	Q_1	*Me*	*M*	Q_3	*Max*	*SD*
	$step_2$	0.97	0.98	0.98	0.98	0.99	0.99	0.00
	$step_3$	0.86	0.92	0.95	0.94	0.97	0.99	0.03
Low								
	slope	0.82	0.87	0.88	0.88	0.90	0.93	0.02
	location	0.86	0.95	0.96	0.96	0.97	0.98	0.02
	asymptote	0.57	0.69	0.77	0.75	0.81	0.89	0.08
	$step_1$	0.87	0.95	0.96	0.96	0.97	0.99	0.02
	$step_2$	0.86	0.94	0.96	0.95	0.97	0.98	0.02
	$step_3$	0.28	0.71	0.86	0.80	0.92	0.97	0.16

Note. Each entry represents the statistic of interest among 100 replications. *Min*=minimum, *Me*=median, *Max*=maximum, *M*=mean, *SD*=standard deviation of the correlations among 100 repetitions. [a]Represents the achievement category of countries that were included in the item parameter estimations.

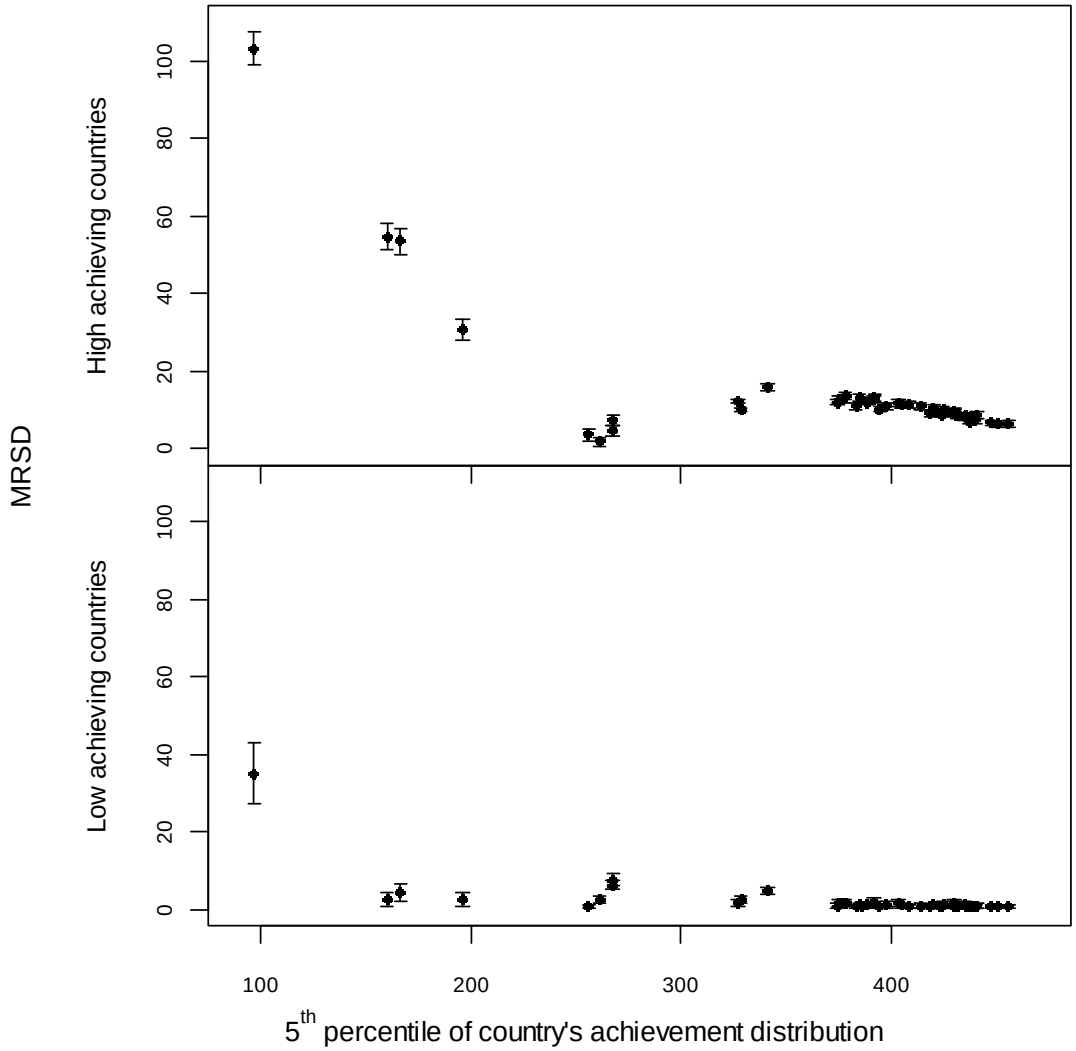
The correlations between most parameters did not differ significantly between the two conditions although the detected effect sizes were of all magnitudes, small, medium and large for the $step_3$ parameter ($step_1$ parameter: z=1.38, p=0.168, q=0.35; $step_2$ parameter: z=1.83, p=0.067, q=0.47; $step_3$ parameter: z=0.78, p=0.435, q=0.64; location: z=1.62, p=0.105, q=0.21; and asymptote: z=0.89, p=0.374, q=0.16). The only significant difference was found for the slope parameter and we could categorize it as small in size (z=2.21, p=0.027, q=0.28). The correlation in slope parameter was significantly higher when higher achieving countries were included in the item parameter estimation in contrast to lower achieving countries. Furthermore, we observed differences in achievement scores based on "high" and "low" achieving countries. The results are presented on the following pages.

Figure 4.7: Reference achievement scores of countries with the corresponding MRSD and standard deviation of MRSD in different conditions



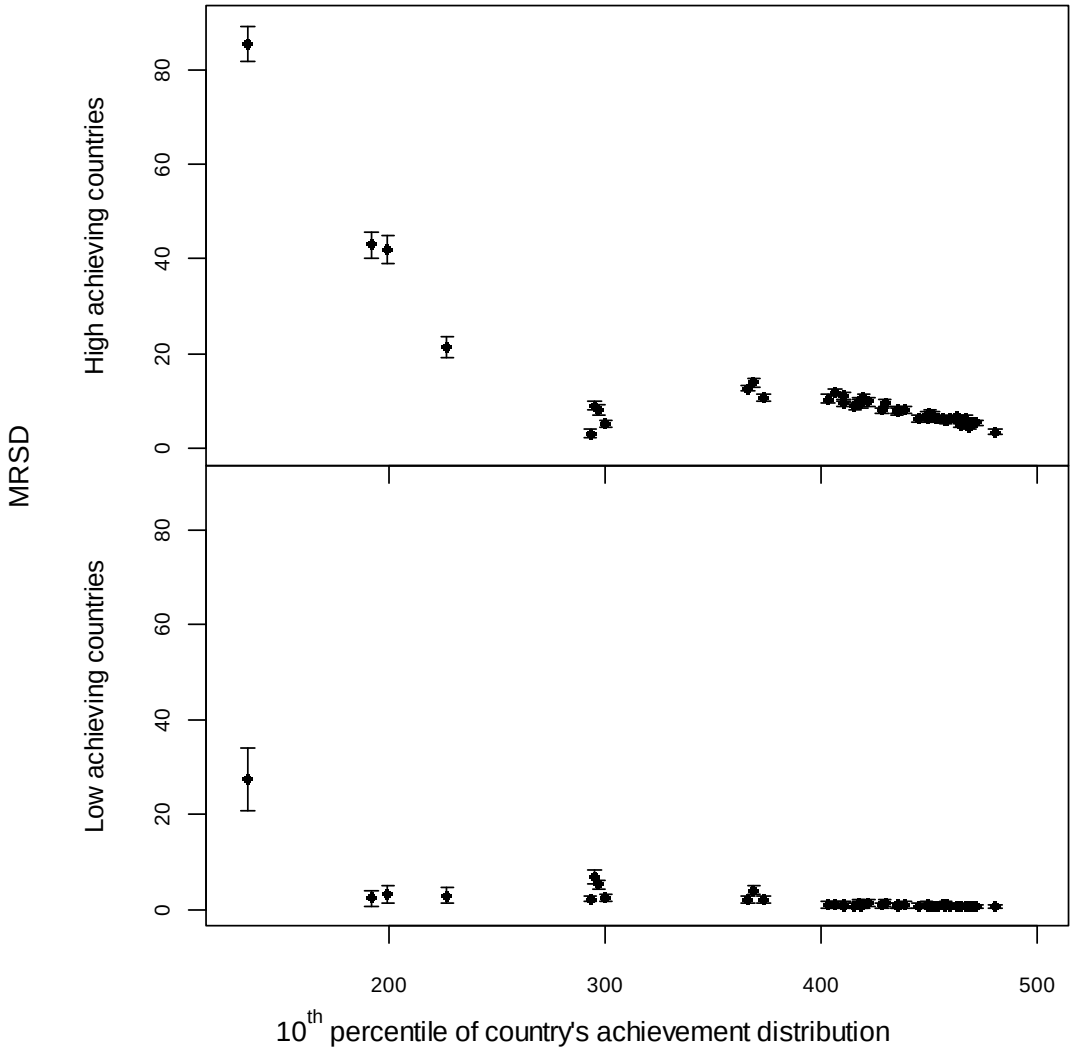
In Figure 4.7 the MRSDs for countries are shown. The variances between the groups of higher and lower achieving countries are significantly different (F=9.432, p=0.003). Also, the MRSDs of these groups are significantly different (t(51)=2.621, p<0.001, d=0.55). The achievement scores based on high achieving countries compared to the reference was higher than the difference in achievement scores based on lower achieving countries compared to the reference (M_{high}=2.60, M_{low}=0.77). The size of the effect can be considered as medium. Furthermore, we observed the difference in certain percentiles of countries distributions.

Figure 4.8: Reference 5th percentile of countries with the corresponding MRSD and standard deviation of MRSD in different conditions



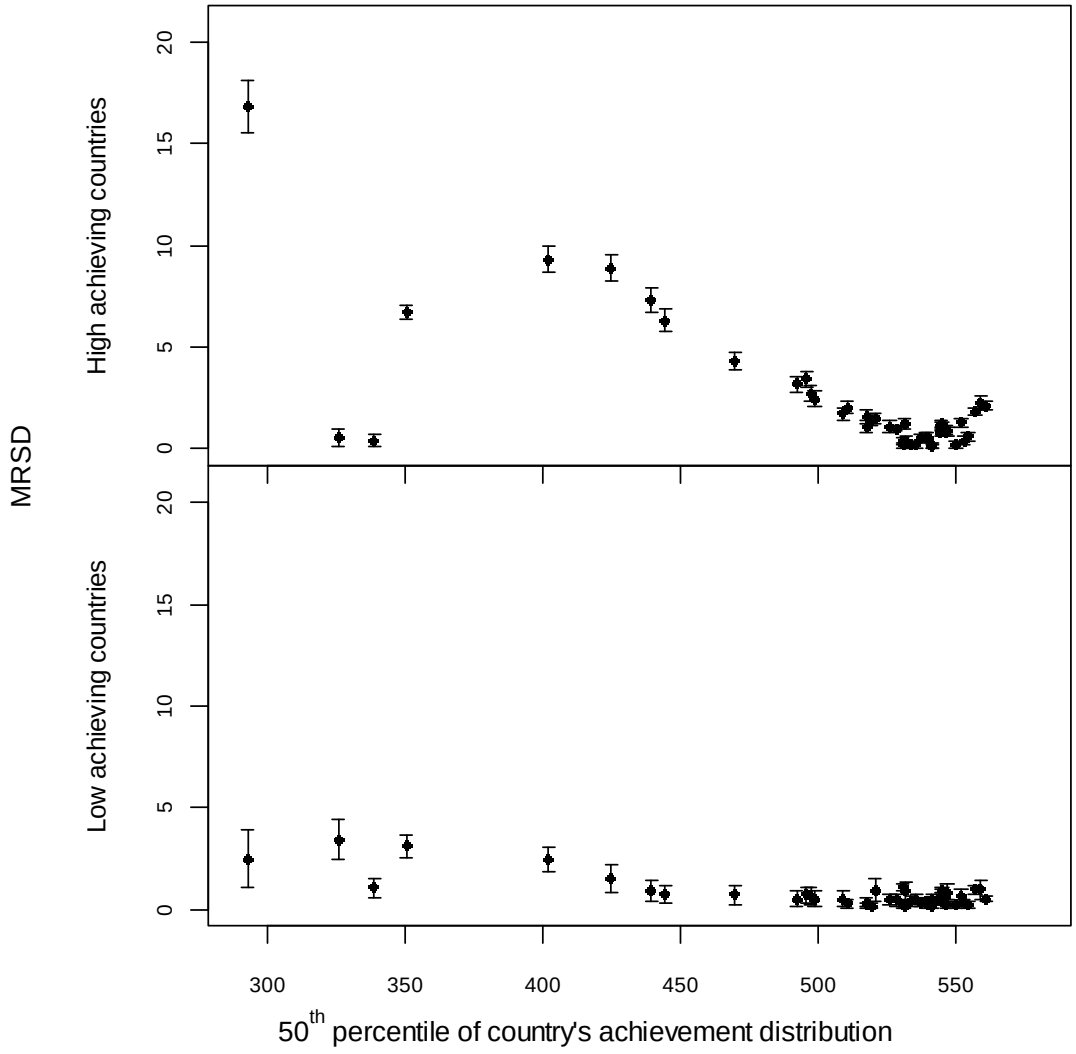
We rejected the hypothesis of equal variances across high and low achieving groups of countries (F=7.286, p=0.008). The differences in MRSD in the 5th percentile of high and low achieving countries was found to be significant with large effect size (t(52)=4.491, p<0.001, d=0.95). The average 5th percentile in high achieving countries showed a greater difference from the reference than the average difference of low achieving countries (M_{high}=14.18, M_{low}=2.36).

Figure 4.9: Reference 10th percentile of countries with the corresponding MRSD and standard deviation of MRSD in different conditions



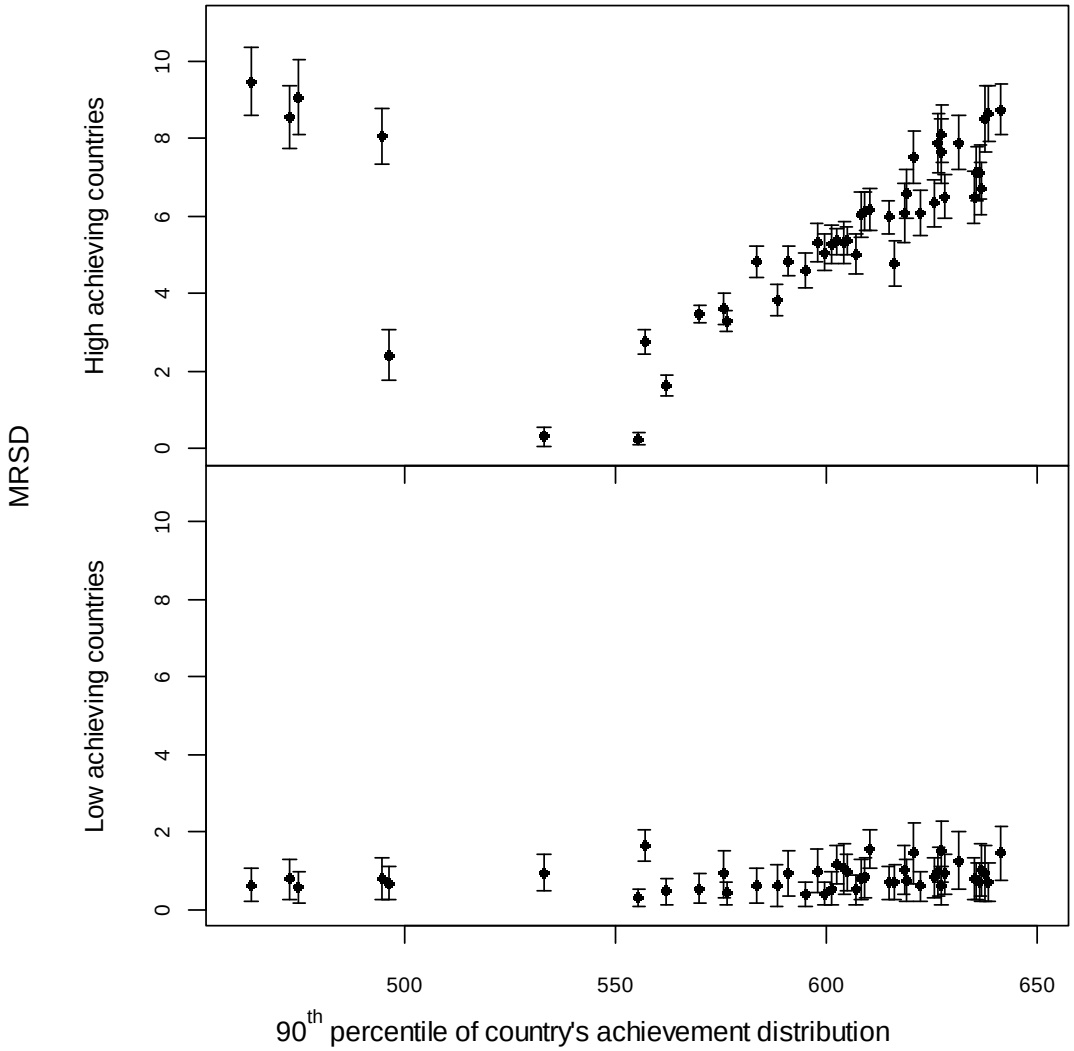
We rejected the hypothesis of equal variances across high and low achieving groups of countries (F=7.076, p=0.009). The differences in the MRSD in the 10th percentile of high and low achieving countries were found to be significant with large effect size (t(52)=4.396, p<0.001, d=0.93). The average 10th percentile in high achieving countries showed a greater difference from the reference than the average difference of low achieving countries (M_{high}=11.36, M_{low}=1.99).

Figure 4.10: Reference 50th percentile of countries with the corresponding MRSD and standard deviation of MRSD in different conditions



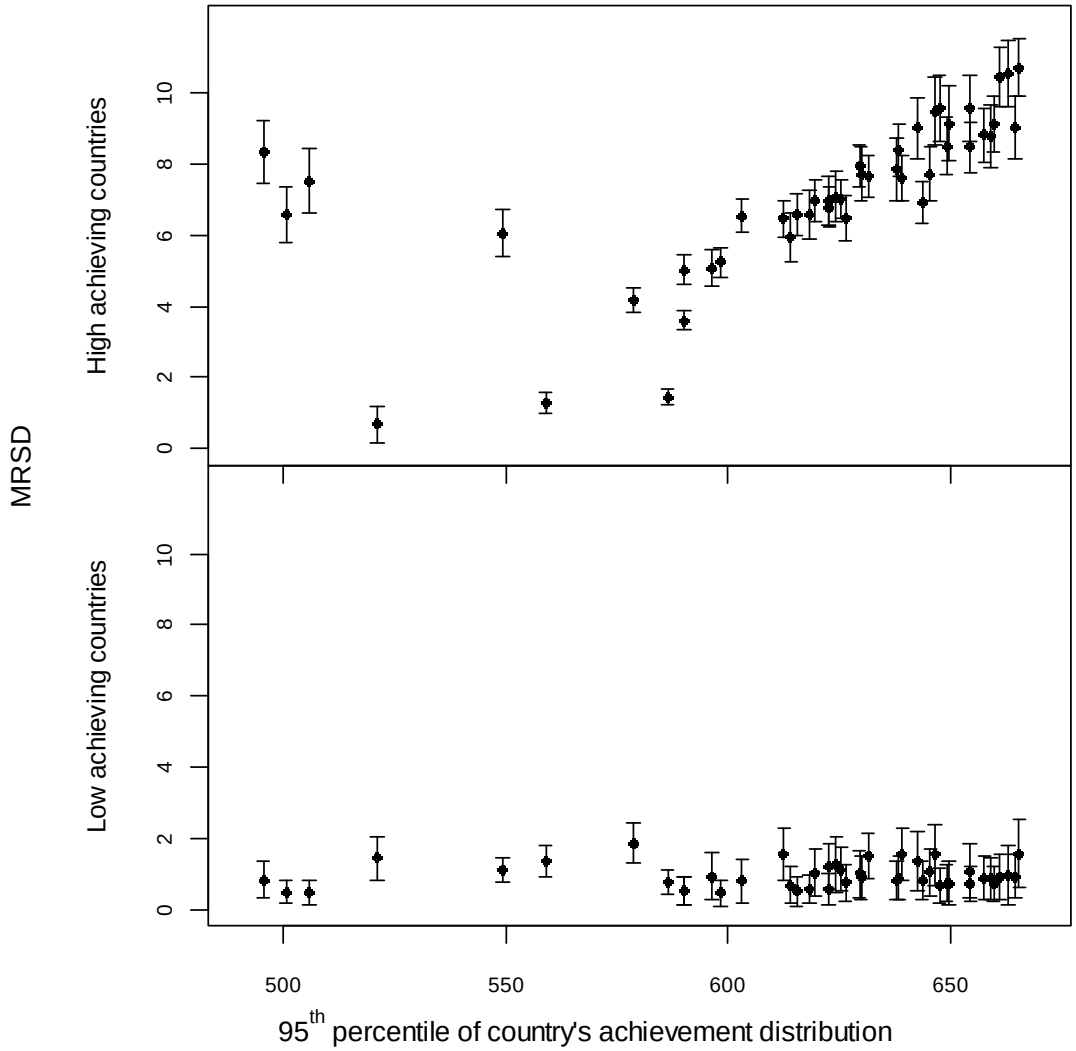
We rejected the hypothesis of equal variances across high and low achieving groups of countries (F=18.195, p<0.001). The differences in the MRSD in the 50th percentile of high and low achieving countries were found to be significant with medium effect size (t(49)=3.073, p<0.001, d=0.65). The average 50th percentile in high achieving countries showed a greater difference from the reference than the average difference of low achieving countries (M_{high}=2.28, M_{low}=0.78).

Figure 4.11: Reference 90th percentile of countries with the corresponding MRSD and standard deviation of MRSD in different conditions



We rejected the hypothesis of equal variances across high and low achieving groups of countries (F=50.912, p<0.001). The differences in the MRSD in the 90th percentile of high and low achieving countries were found to be significant with large effect size (t(46)=14.896, p<0.001, d=3.14). The effect size for this analysis was found to exceed Cohen's convention for a large effect. The average 90th percentile in high achieving countries showed a greater difference from the reference than the average difference of low achieving countries (M_{high}=5.79, M_{low}=0.84).

Figure 4.12: Reference 95th percentile of countries with the corresponding MRSD and standard deviation of MRSD under different conditions



We rejected the hypothesis of equal variances across high and low achieving groups of countries (F=36.495, p<0.001). The differences in the MRSD in the 95th percentile of high and low achieving countries were found to be significant and the effect size was large (t(46)=17.967, p<0.001, d=3.79). The average 95th percentile in high achieving countries showed a greater difference from the reference than the average difference of low achieving countries (M_{high}=7.14, M_{low}=0.96).

Table 4.10: Descriptives for MRSD across conditions by gender

Category	Condition	*M*	*Me*	*SD*	*Min*	*Max*
girls	higher	2.2	1.1	3.6	0.1	22.9
	lower	0.7	0.5	1.0	0.1	6.3
boys	higher	3.2	1.6	5.6	0.1	33.6
	lower	0.8	0.4	1.5	0.1	9.4

Note: M=mean, *Me*=median, *SD*=standard deviation, *Min*=minimum, *Max*=maximum.

Variances between the two groups of countries in girls were found to be significantly different (F=7.832, p=0.006). The MRSD in higher achieving countries was significantly higher than in lower achieving countries, the effect size was medium (t(51)=2.615, p=0.012, d=0.55).

Variances between the two groups of countries in boys were significantly different (F=9.093, p=0.003). The MRSD in higher achieving countries was significantly higher than for lower achieving countries, the effect size was medium (t(50)=2.750, p=0.008, d=0.58).

Table 4.11: Descriptives of MRSD across conditions for number of books

Category	Condition	*M*	*Me*	*SD*	*Min*	*Max*
missing	higher	6.7	5.6	7.3	0.9	47.7
	lower	1.3	0.7	2.1	0.2	13.9
0 – 10 books	higher	5.2	4.4	4.3	0.4	30.1
	lower	1.0	0.5	1.2	0.3	8.0
11 – 25 books	higher	2.8	2.3	2.6	0.2	16.5
	lower	0.7	0.4	0.8	0.2	4.4
26 – 100 books	higher	1.8	1.2	2.1	0.1	12.3
	lower	0.6	0.4	0.7	0.1	3.6
101 – 200 books	higher	2.1	1.6	2.8	0.1	18.7
	lower	0.7	0.6	0.9	0.2	5.1
more than 200 books	higher	2.1	1.7	2.4	0.1	14.9
	lower	0.7	0.5	0.7	0.2	4.0

Note: M=mean, *Me*=median, *SD*=standard deviation, *Min*=minimum, *Max*=maximum.

Variances within all categories in the number of books at home differed significantly between higher and lower achieving country conditions (missing: F=6.165, p=0.015; 0-10 books: F=8.048, p=0.006; 11-25 books: F=14.506, p<0.001; 26-100 books: F=13.469, p<0.001; 101-200 books: F=5.714, p=0.019; more than 200 books: F=10.405, p=0.002). The MRSD in higher achieving country condition was significantly higher than in lower achieving country condition in all categories of the variable number of books at home and

the effect sizes can be considered medium and large (books0: $t(51)=4.759$, $p<0.001$, $d=1.00$; books1: $t(51)=6.329$, $p<0.001$, $d=1.33$; books2: $t(52)=5.256$, $p<0.001$, $d=1.11$; books3: $t(53)=3.588$, $p<0.001$, $d=0.76$; books4: $t(52)=3.111$, $p=0.003$, $d=0.66$; books5: $t(53)=3.643$, $p=0.001$, $d=0.77$).

4.6 Different models

For model comparisons we selected three different conditions. The condition 3PL 2PL GPCM – 3PL 2PL GPCM is comparing the 3PL 2PL GPCM models with ten countries included in the calibration sample (in this case we used the same results as obtained for a different number of countries when ten countries were included in the item parameter estimation) to the 3PL 2PL GPCM reference (when all countries were included in the calibration sample). In this condition 3PL and 2PL models were used in addition to the generalized partial credit model. The selection of ten countries was repeated and compared to the reference condition (when all 45 countries were included in the item parameter estimation). In the next step, item parameters were obtained including all countries and with the use of Rasch model and the partial credit model. This was the reference for the "Rasch" condition. Then item parameters based on the Rasch and partial credit models were obtained in repeated procedure, randomly selecting ten countries into the calibration sample. We also compared the repeated Rasch model with the Rasch reference (Rasch – Rasch) as well as the Rasch model with the 3PL, 2PL and GPCM reference (Rasch – 3PL 2PL GPCM). The correlations under different conditions are presented in Table 4.12.

Table 4.12: Correlation characteristics of item parameters across conditions

Condition[a]	Item parameter	*Min*	Q_1	*Me*	*M*	Q_3	*Max*	*SD*
Rasch – Rasch								
	location	0.98	0.99	0.99	0.99	0.99	1.00	0.00
	$step_1$	0.97	0.99	0.99	0.99	0.99	1.00	0.00
	$step_2$	0.97	0.99	0.99	0.99	0.99	0.99	0.00
	$step_3$	0.73	0.91	0.95	0.93	0.97	1.00	0.05
3PL 2PL GPCM – 3PL 2PL GPCM								
	location	0.76	0.97	0.97	0.97	0.98	0.99	0.03
	$step_1$	0.97	0.98	0.99	0.99	0.99	1.00	0.01
	$step_2$	0.97	0.98	0.99	0.99	0.99	0.99	0.01
	$step_3$	0.63	0.86	0.93	0.90	0.95	1.00	0.07

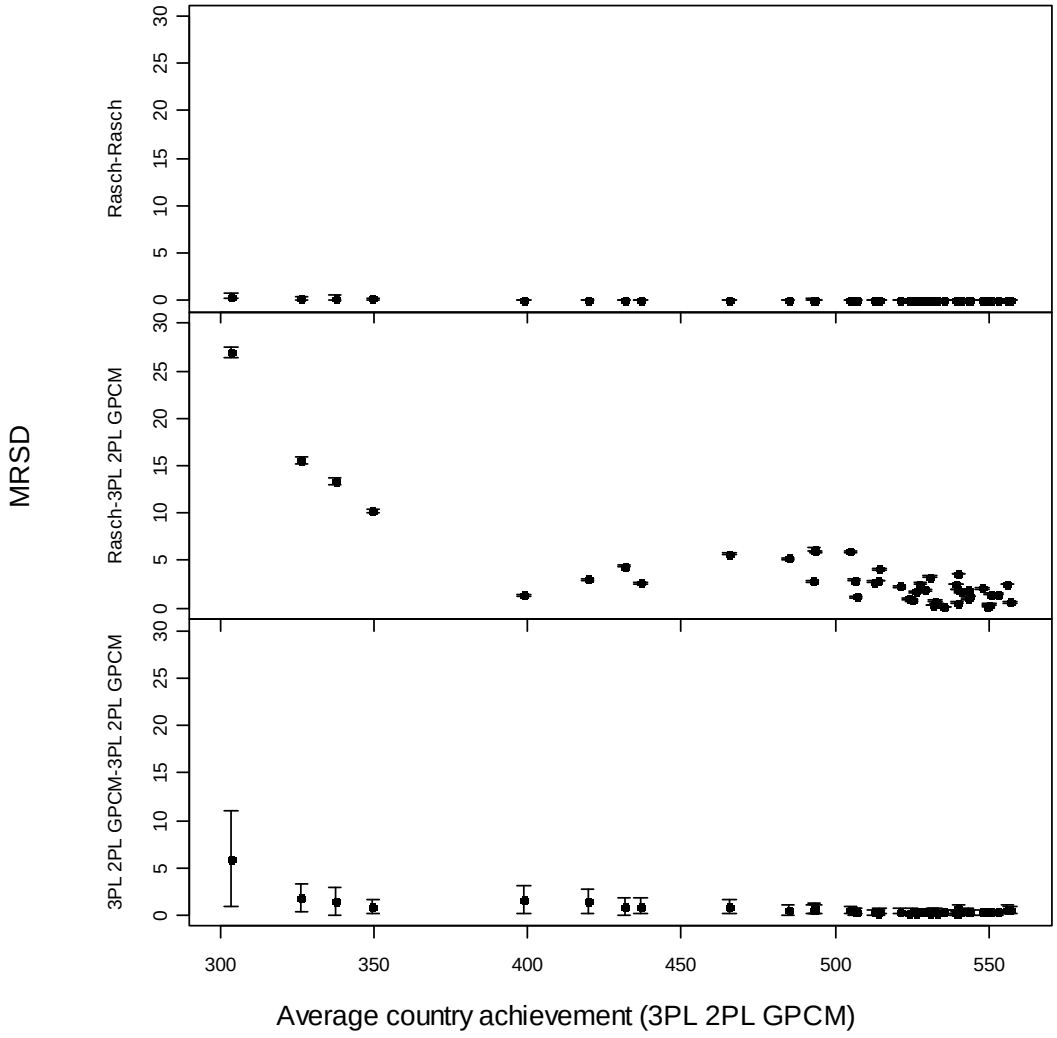
Condition[a]	Item parameter	*Min*	Q_1	*Me*	*M*	Q_3	*Max*	*SD*
Rasch – 3PL 2PL GPCM								
	location	0.86	0.89	0.89	0.89	0.90	0.91	0.01
	$step_1$	0.95	0.97	0.97	0.97	0.98	0.99	0.01
	$step_2$	0.93	0.95	0.96	0.96	0.97	0.98	0.01
	$step_3$	0.67	0.88	0.93	0.92	0.96	1.00	0.06

Note. Each entry represents the statistic of interest among 100 replications. *Min*=minimum, *Me*=median, *Max*=maximum, *M*=mean, *SD*=standard deviation of the correlations among 100 repetitions. [a]Represents the model used in the item parameter estimations and the reference model to which the repetitions were compared to.

All presented correlations were very high (above 0.89). There were no significant differences in $step_1$, $step_2$ and $step_3$ parameters between the Rasch - Rasch and 3PL 2PL GPCM - 3PL 2PL GPCM conditions but these two conditions differed significantly in the mean correlation of the location parameter (z=4.33, p<0.001, q=0.55). The correlation for the location parameter in the Rasch – 3PL 2PL GPCM condition was significantly lower when compared to both other conditions (Rasch – Rasch: z=9.57, p<0.001, q=1.22; 3PL 2PL GPCM – 3PL 2PL GPCM: z=5.24, p<0.001, q=0.67). All reported effect sizes for this test are considered large in magnitude.

In addition, the correlation between same parameters in the reference conditions (all countries with the 3PL 2PL GPCM model and all countries with the Rasch model) were very high (location= .90, $step_1$= .98, $step_2$= .97, $step_3$= .98). Also correlation of the two reference achievement scores (Rasch and 3PL 2PL GPCM) was very high (0.997, p<0.001).

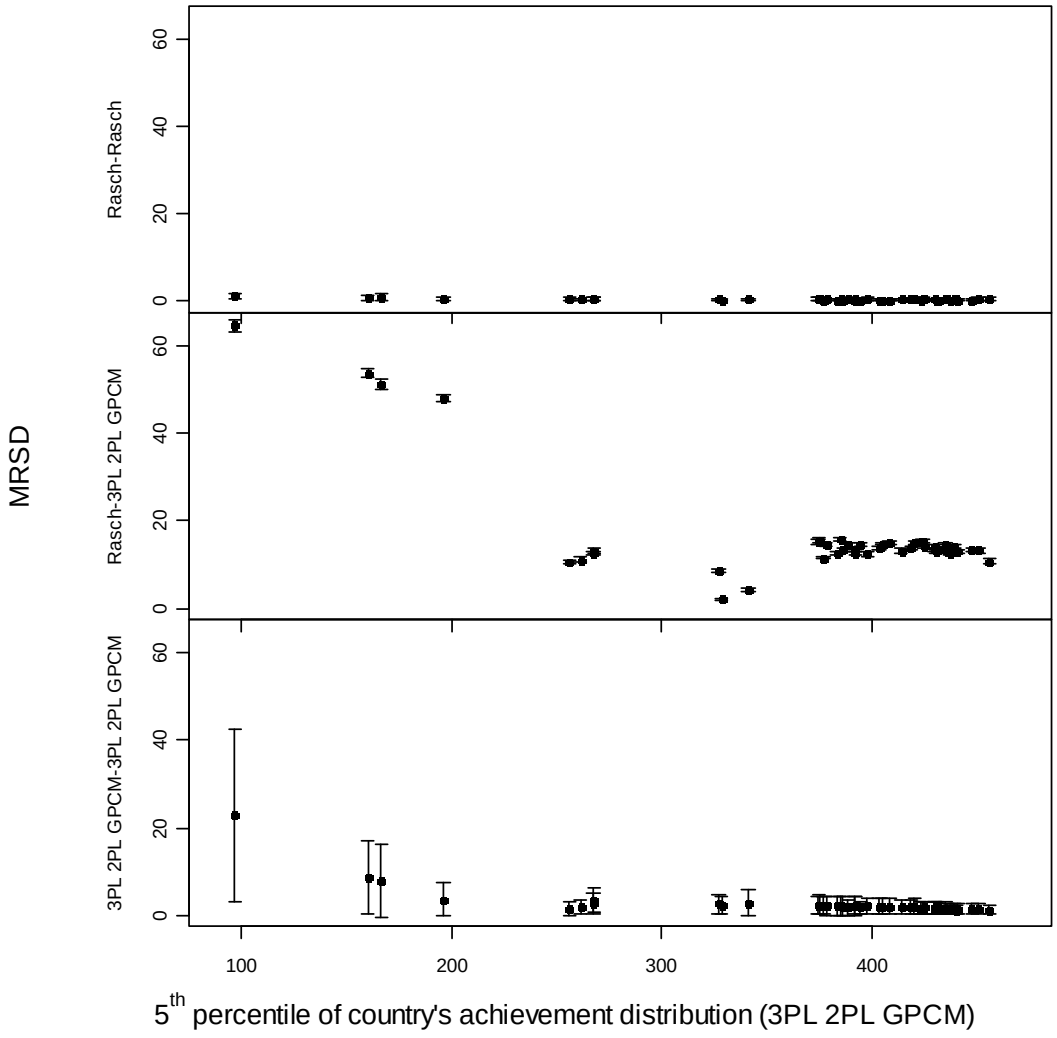
Figure 4.13: Reference achievement scores of countries with the corresponding MRSD and standard deviation of MRSD under different conditions



The variances across conditions using different models were found to be significantly different (F=20.53, p<0.001). The MRSD for countries differed significantly across models (*Welch's* F(2,59)=21.29, p<0.001). The highest average MRSD was found in the Rasch - 3PL 2PL GPCM comparison (M=3.58), and the smallest differences were found in the Rasch - Rasch comparison, which showed an average MRSD of 0.08 across 100 repetitions (the average MRSD in 3PL 2PL GPCM - 3PL 2PL GPCM condition was 0.66). As can be seen in Figure 4.13, the average differences when the results were based on the Rasch model were very stable across countries. In other words, the variation between repetitions within the Rasch – Rasch condition was very small. This contrasted with the

3PL 2PL GPCM - 3PL 2PL GPCM comparison, where the variation of scores in lower achieving countries (average achievement below 450) was higher than for other countries. According to post-hoc comparisons all pairs of conditions differed significantly and showed large effect sizes (0.58<*Mean difference*<3.50, p<0.001, 0.89<d<1.04).

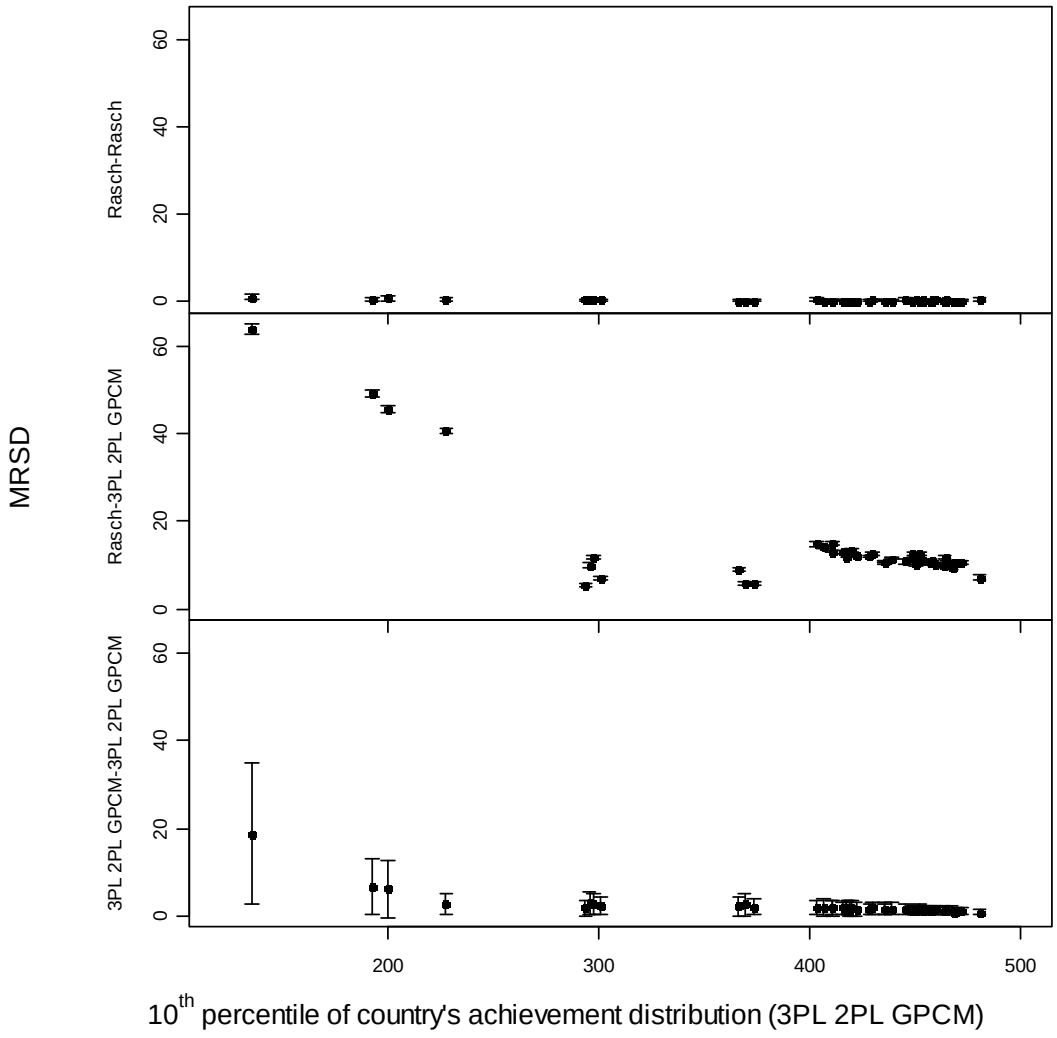
Figure 4.14: Reference 5th percentile of countries (3PL 2PL GPCM) with the corresponding MRSD and standard deviation of MRSD under different conditions



In the 5th percentile, the variances of groups were not found to be homogenous (F=14.20, p<0.001). The differences between the conditions were statistically significant (*Welch's* F(2,59)=49.57, p<0.001). The biggest difference was found when comparing the Rasch model with the 3PL 2PL GPCM model as a reference model (M=16.63), although all pairs

of conditions differed significantly and the effect sizes were large (2.40<*Mean difference*<16.22, p<0.001, 1.00<d<1.86).

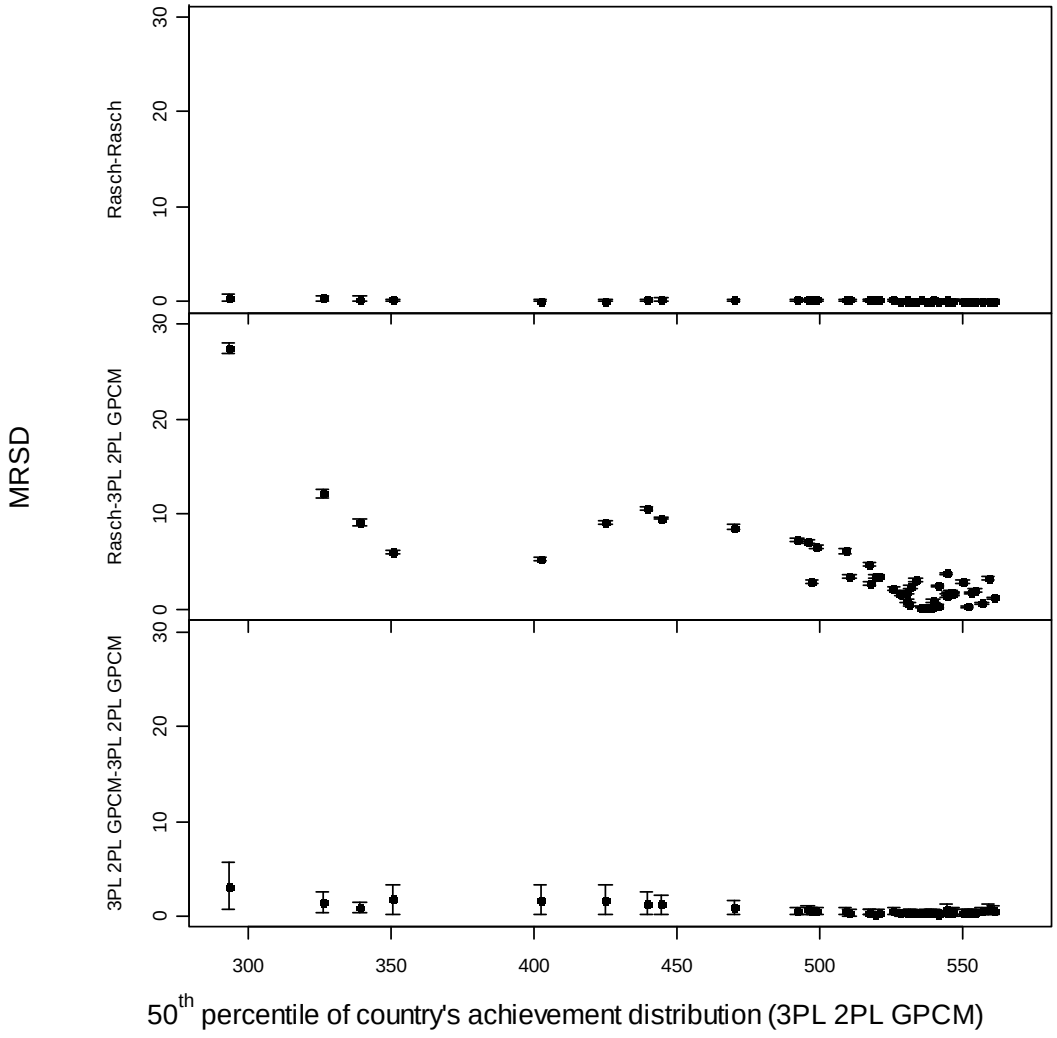
Figure 4.15: Reference 10th percentile of countries (3PL 2PL GPCM) with the corresponding MRSD and standard deviation of MRSD under different conditions



When observing the 10th percentile, the variances of groups were not found to be homogenous (F=14.64, p<0.001). The differences between the conditions were statistically significant (*Welch's* F(2,59)=43.29, p<0.001). The biggest difference was found when comparing the Rasch model with the 3PL 2PL GPCM model as a reference model

(*M*=14.39), although all pairs of conditions differed significantly and the effect sizes were large (1.95<*Mean difference*<14.06, $p<0.001$, 1.00<*d*<1.70).

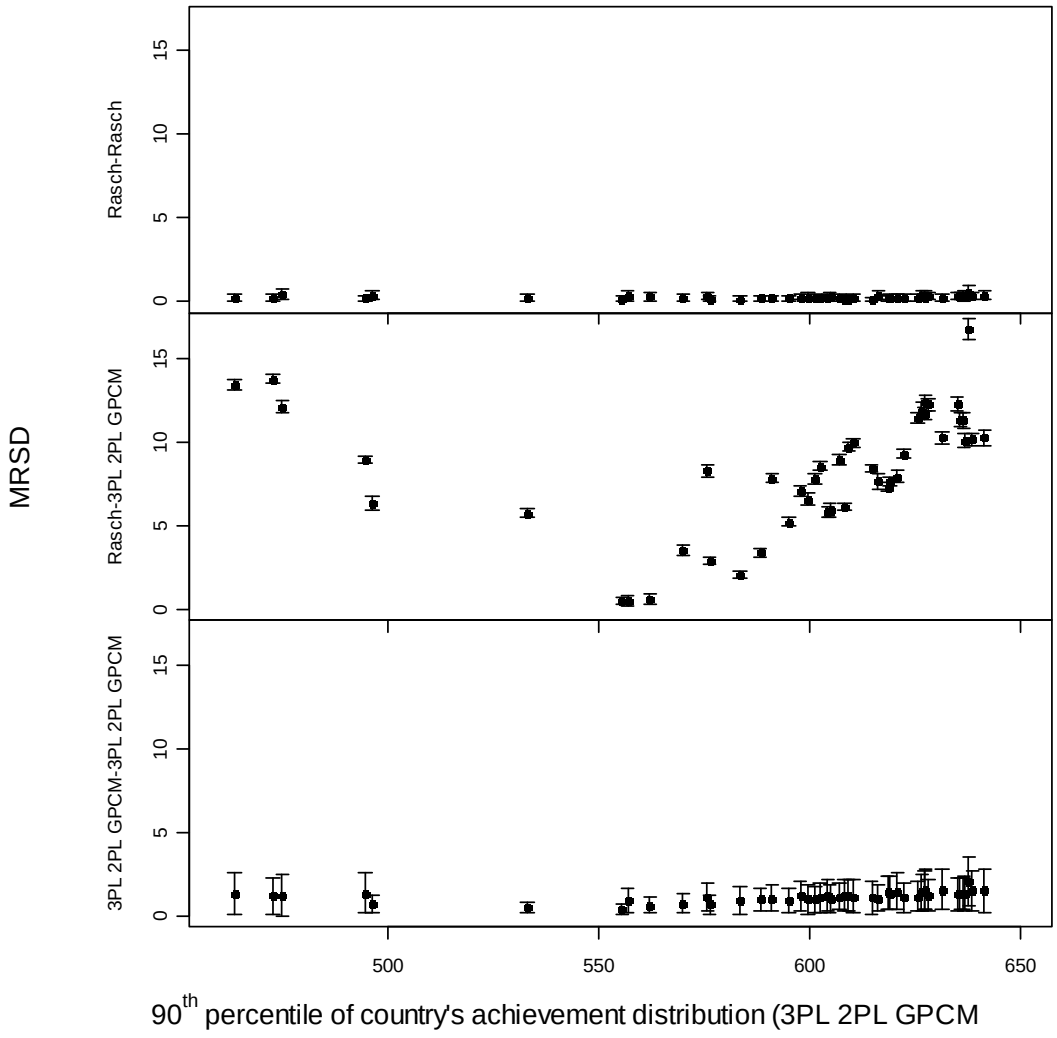
Figure 4.16: Reference 50th percentile of countries (3PL 2PL GPCM) with the corresponding MRSD and standard deviation of MRSD under different conditions



In the 50th percentile, the variances of groups were also not found to be homogenous (*F*=33.14, $p<0.001$). The differences between conditions were statistically significant (*Welch's F*(2,59)=33.13, $p<0.001$). The biggest difference was found when comparing the Rasch model with the 3PL 2PL GPCM model as a reference model (*M*=4.11), although all

pairs of conditions differed significantly and the effect sizes were found to be large (0.51<*Mean difference*<3.95, p<0.001, 1.18<d<1.26).

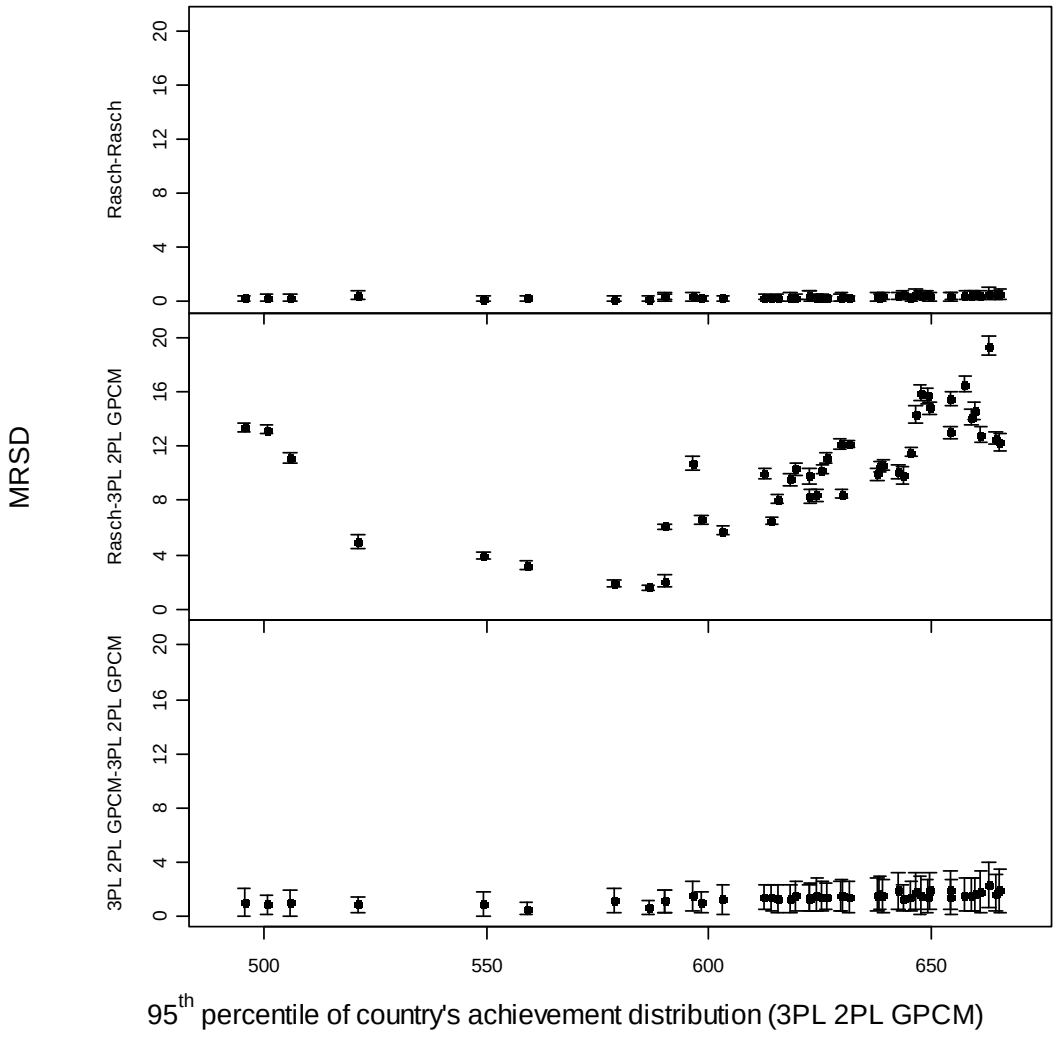
Figure 4.17: Reference 90th percentile of countries (3PL 2PL GPCM) with the corresponding MRSD and standard deviation of MRSD under different conditions



When observing the 90th percentile, the variances of groups were not found to be homogenous (F=65.54, p<0.001). The differences between conditions were statistically significant (*Welch's* F(2,62)=277.85, p<0.001). The biggest difference was found when comparing the Rasch model with the 3PL 2PL GPCM model as a reference model

(*M*=8.24), although all pairs of conditions differed significantly and the effect sizes were large (0.90<*Mean difference*<7.96, *p*<0.001, 2.69<*d*<3.98).

Figure 4.18: Reference 95th percentile of countries (3PL 2PL GPCM) with the corresponding MRSD and standard deviation of MRSD under different conditions



Finally, when observing the 95th percentile, the variances of groups were not found to be homogenous (*F*=59.01, *p*<0.001). The differences between conditions were statistically significant (*Welch's F*(2,59)=303.89, *p*<0.001). The biggest difference was found when comparing the Rasch model with the 3PL 2PL GPCM model as a reference model

(*M*=10.32), although all pairs of conditions differed significantly and the effect sizes were large (1.04<*Mean difference*<9.95, *p*<0.001, 3.04<*d*<3.98).

Table 4.13: Descriptives of MRSD across conditions by gender

Category	Condition	*M*	*Me*	*SD*	*Min*	*Max*
girls	Rasch - Rasch	0.08	0.06	0.07	0.04	0.43
	Rasch - 3PL 2PL GPCM	3.04	1.90	3.73	0.08	22.40
	3PL 2PL GPCM - 3PL 2PL GPCM	0.60	0.37	0.73	0.19	4.87
boys	Rasch - Rasch	0.09	0.07	0.09	0.03	0.50
	Rasch - 3PL 2PL GPCM	4.39	2.67	6.01	0.14	31.62
	3PL 2PL GPCM - 3PL 2PL GPCM	0.76	0.46	1.08	0.19	7.08

Note: M=mean, *Me*=median, *SD*=standard deviation, *Min*=minimum, *Max*=maximum.

Variances between conditions in category of girls differed significantly (*F*=21.09, *p*<0.001). The MRSD differed significantly between conditions (*Welch's F*(2,59)=25.28, *p*<0.001). Based on the Games-Howell post hoc test, the MRSD in the Rasch – 3PL 2PL GPCM condition was significantly higher than in other conditions and the Rasch – Rasch condition showed the smallest differences (0.52<*Mean difference*<2.96, *p*<0.001, 0.91<*d*<1.12). All the effect sizes for this test can be considered as large.

Variances between conditions in the category of boys also differed significantly (*F*=19.58, *p*<0.001). The MRSD differed significantly between conditions (*Welch's F*(2,59)=19.64, *p*<0.001). Based on the Games-Howell post hoc test, the MRSD in the Rasch – 3PL 2PL GPCM condition was significantly higher than in other conditions and the Rasch – Rasch condition showed the smallest differences (0.66<*Mean difference*<4.30, *p*<0.001, 0.84<*d*<1.01). All the effect sizes for this test can be considered as large.

Table 4.14: Descriptives of MRSD across conditions for number of books

Category	Condition	*M*	*Me*	*SD*	*Min*	*Max*
missing	Rasch - Rasch	0.22	0.18	0.12	0.08	0.74
	Rasch - 3PL 2PL GPCM	7.90	6.76	7.22	0.24	37.75
	3PL 2PL GPCM - 3PL 2PL GPCM	1.40	1.22	1.44	0.52	10.07
0 – 10 books	Rasch - Rasch	0.15	0.13	0.08	0.07	0.51
	Rasch - 3PL 2PL GPCM	7.10	6.65	4.98	0.35	32.11
	3PL 2PL GPCM - 3PL 2PL GPCM	1.10	0.88	0.87	0.37	6.31
11 – 25 books	Rasch - Rasch	0.11	0.11	0.05	0.04	0.30
	Rasch - 3PL 2PL GPCM	4.22	3.62	3.05	0.33	18.32
	3PL 2PL GPCM - 3PL 2PL GPCM	0.67	0.51	0.53	0.25	3.57

Category	Condition	*M*	*Me*	*SD*	*Min*	*Max*
26 – 100 books	Rasch - Rasch	0.08	0.07	0.04	0.04	0.21
	Rasch - 3PL 2PL GPCM	2.77	2.10	2.39	0.24	12.34
	3PL 2PL GPCM - 3PL 2PL GPCM	0.52	0.38	0.43	0.20	2.72
101 – 200 books	Rasch - Rasch	0.11	0.10	0.05	0.05	0.31
	Rasch - 3PL 2PL GPCM	2.91	2.37	2.90	0.18	17.14
	3PL 2PL GPCM - 3PL 2PL GPCM	0.61	0.42	0.61	0.20	4.00
more than 200	Rasch - Rasch	0.12	0.10	0.08	0.05	0.46
	Rasch - 3PL 2PL GPCM	2.84	2.17	2.92	0.07	16.11
	3PL 2PL GPCM - 3PL 2PL GPCM	0.59	0.46	0.49	0.18	3.16

Note: M=mean, *Me*=median, *SD*=standard deviation, *Min*=minimum, *Max*=maximum.

Variances in all categories of the variable number of books differed across conditions (missing: F=17.41, p<0.001; 0 – 10 books: F=19.02, p<0.001; 11 – 25 books: F=31.92, p<0.001; 26 – 100 books: F=56.09, p<0.001; 101 – 200 books: F=29.56, p<0.001; more than 200 books: F=28.84, p<0.001). The average MRSD differed significantly across conditions in all categories (missing: $F(2,59)$=40.77, 1.19<*Mean difference*<7.68, p<0.001, 1.16<d<1.50; 0 – 10 books: $F(2,59)$=70.14, 0.96<*Mean difference*<6.95, p<0.001, 1.56<d<1.97; 11 – 25 books: $F(2,59)$=64.20, 0.55<*Mean difference*<4.11, p<0.001, 1.47<d<1.90; 26 – 100 books: $F(2,59)$=50.35, 0.43<*Mean difference*<2.69, p<0.001, 1.39<d<1.59; 101 – 200 books: $F(2,59)$=35.65, 0.50<*Mean difference*<2.80, p<0.001, 1.10<d<1.37; more than 200 books: $F(2,60)$=38.83, 0.47<*Mean difference*<2.72, p<0.001, 1.07<d<1.33). According to the post hoc test, in every category the Rasch - Rasch condition showed significantly lower MRSD values than in the other two conditions (p<0.001) and the Rasch - 3PL 2PL GPCM condition showed the highest MRSD values (significantly higher than in other conditions; p<0.001). All the effect sizes for this test can be considered as large.

4.7 Different content domains

In the final research question, we were interested in the invariance across content domains. The content domains assessed in TIMSS and PIRLS differ (besides in frequency of different item types) in the structure of presenting cognitive items. For investigating the last research question we chose data from TIMSS conducted in 2007 and PIRLS conducted in 2006. In both studies the knowledge of 4th grade students was assessed across countries. Although TIMSS assesses knowledge in mathematics and in science we only investigated

the mathematics achievement. For reasons of comparison we only chose those countries that participated in both cycles of studies. Thus, we identified 29 countries with identical country codes in both studies so that the comparisons between content domains can be meaningful. To sum up we compared two conditions one representing the condition of reading and the other the condition of mathematics. Within the condition of reading we compared parameters of ten randomly sampled countries in PIRLS out of 29 to the reference parameters (all selected 29 countries in PIRLS). The procedure was repeated 100 times following the same logic as in previous research questions. And the same was done within the domain of mathematics. We compared parameters of ten randomly sampled countries in TIMSS out of 29 to the reference parameters (all selected 29 countries in TIMSS).

Once again, first the correlations in item parameters were observed, and then the differences in achievement scores were examined.

Table 4.15: Correlation characteristics of item parameters across conditions

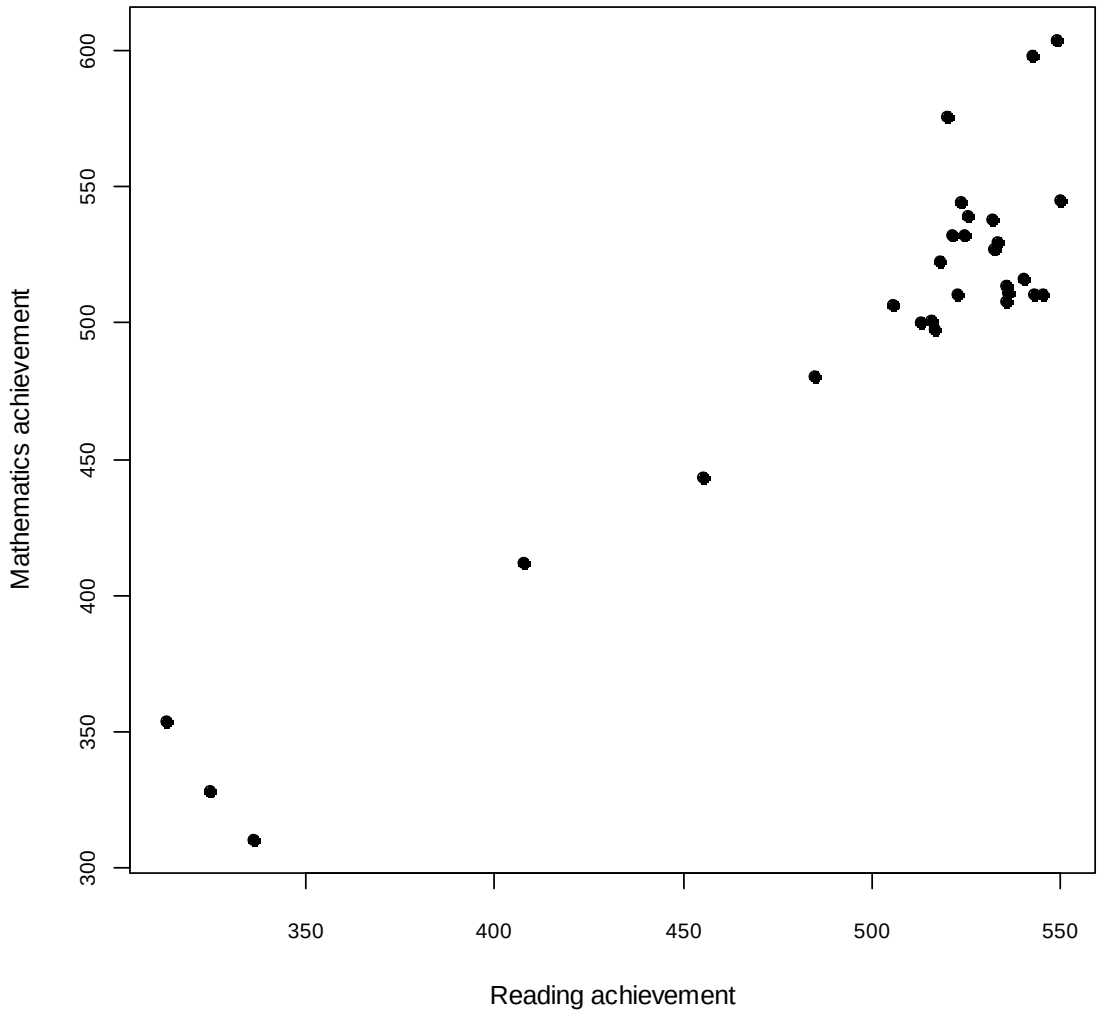
Condition[a]	Item parameter	*Min*	Q_1	*Me*	*M*	Q_3	*Max*	*SD*
Reading								
	slope	0.92	0.95	0.96	0.96	0.97	0.98	0.01
	location	0.92	0.97	0.98	0.98	0.98	0.99	0.01
	asymptote	0.60	0.78	0.84	0.82	0.90	0.93	0.09
	$step_1$	0.98	0.99	0.99	0.99	0.99	1.00	0.00
	$step_2$	0.98	0.99	0.99	0.99	0.99	1.00	0.00
	$step_3$	0.80	0.92	0.95	0.94	0.97	0.99	0.04
Mathematics								
	slope	0.84	0.92	0.93	0.93	0.94	0.96	0.02
	location	0.93	0.97	0.98	0.97	0.98	0.99	0.01
	asymptote	0.54	0.77	0.90	0.85	0.92	0.95	0.10
	$step_1$	0.87	0.97	0.98	0.97	0.99	1.00	0.02
	$step_2$	0.87	0.97	0.98	0.97	0.99	1.00	0.02
	$step_3$[b]	NA	NA	NA	NA	NA	NA	NA

Note. Each entry represents the statistic of interest among 100 replications. *Min*= minimum, *Me*=median, *Max*=maximum, *M*=mean, *SD*=standard deviation of the correlations among 100 repetitions. [a]Represents the content domain. [b] There were no items with more than 2 score points in mathematics.

The lowest correlation coefficients were again observed for the asymptote parameter, all other correlations were very high (above 0.93). The correlation for the slope parameter was significantly higher in the condition of reading than in mathematics the effect size was

small (z=2.25, p=0.024, q=0.29); this was the only significant difference in the correlation coefficients across conditions in the content domains.

Figure 4.19: The correlation between reading and mathematics achievement in countries participating in PIRLS 2006 and TIMSS 2007

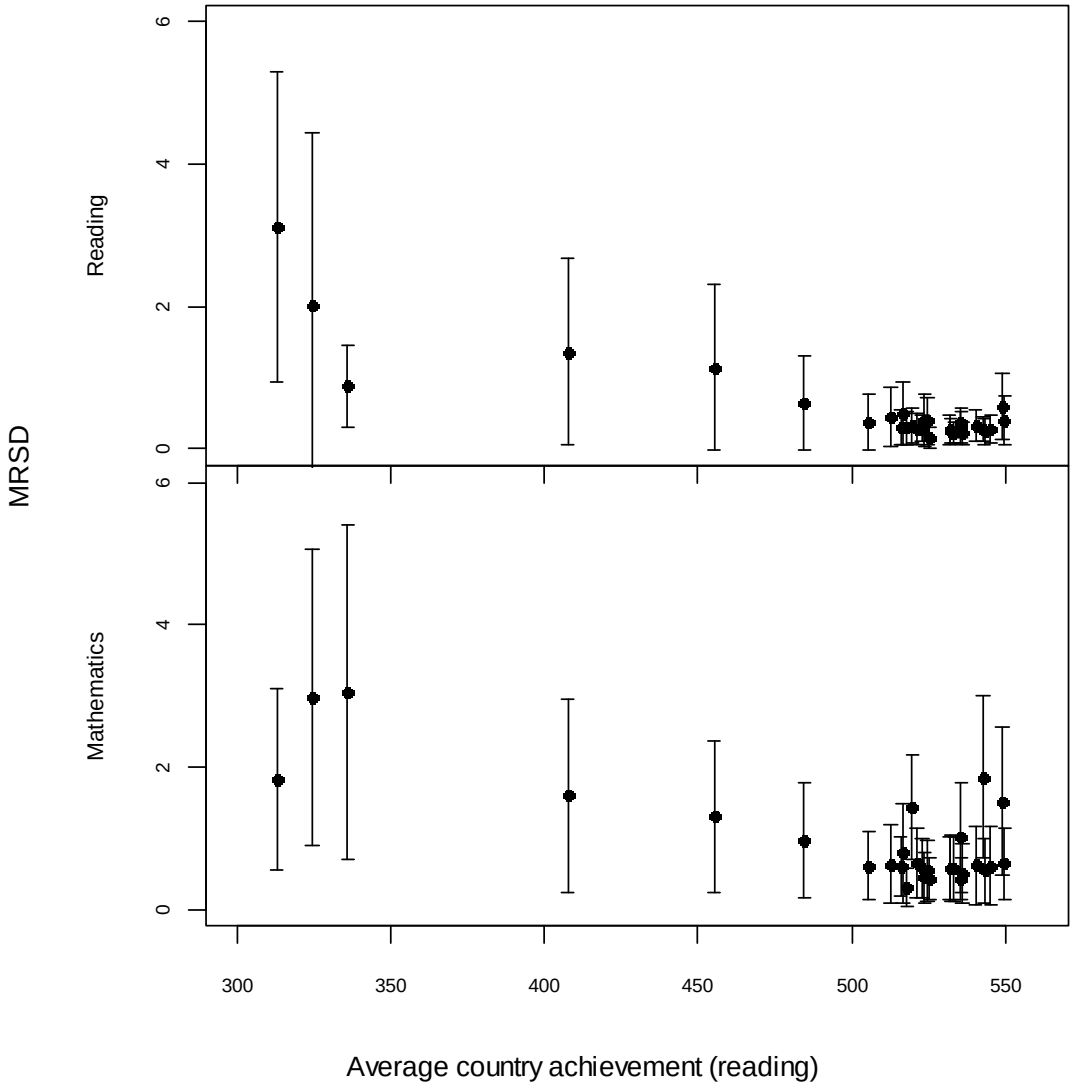


The correlation between the achievement scores in reading and mathematics was very high (r=0.89). The correlation was calculated from the reference reading and mathematics condition (when all 29 countries were included in the item parameter estimations). In general, countries that had a high achievement in mathematics also had a high achievement in reading (or vice versa).

The high correlation between content domains was also found in other studies. For example, Cromley (2008) reported that the average of the correlations between the individual scores in mathematics, sciences and reading for countries is above 0.8. The reason most probably lies in the fact that reading results represent an aspect of general verbal competencies that are also included in solving mathematics and science items (Bulle 2011).

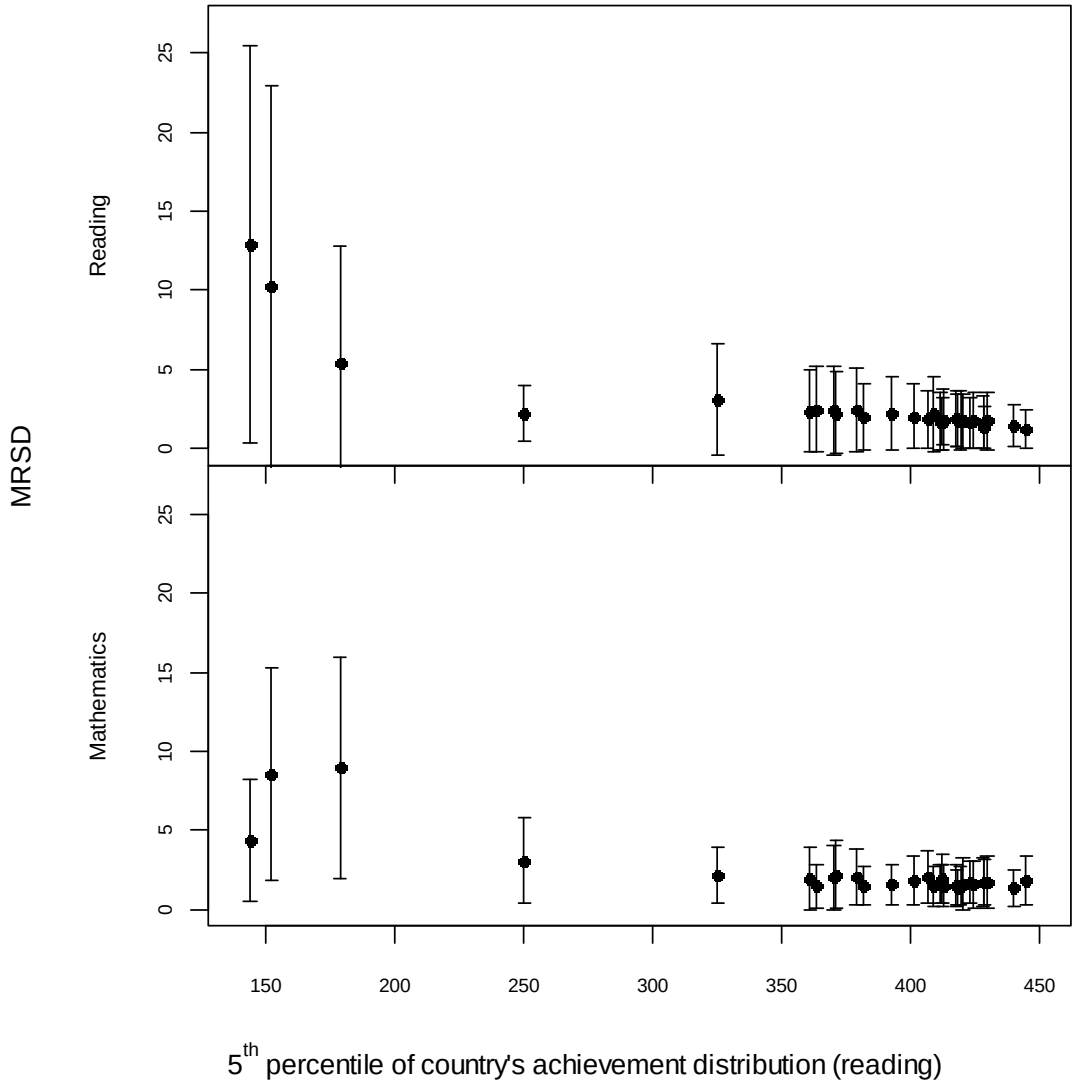
Due to high correlation between average mathematics and reading achievement in countries we present the average reading achievement as the reference in figures in this chapter. This is only for the purpose of presenting the results, so that countries between the content domains are vertically aligned in the figures (as in all other chapters). For example one country had an average achievement in reading of 336 points with MRSD of 0.9 points and an average achievement in mathematics 310 points with MRSD of 3. In Figure 4.20 this country is presented with the achievement score of reading (336, the third one from the left) in both content domains (also mathematics) although the MRSD in mathematics is based on the comparisons to 310 points. This is only to follow the same presentation style as in previous chapters. In all calculations the proper scores were used as the reference – either reading or mathematics.

Figure 4.20: Reference achievement scores in reading for countries with the corresponding MRSD and standard deviation of MRSD under different conditions



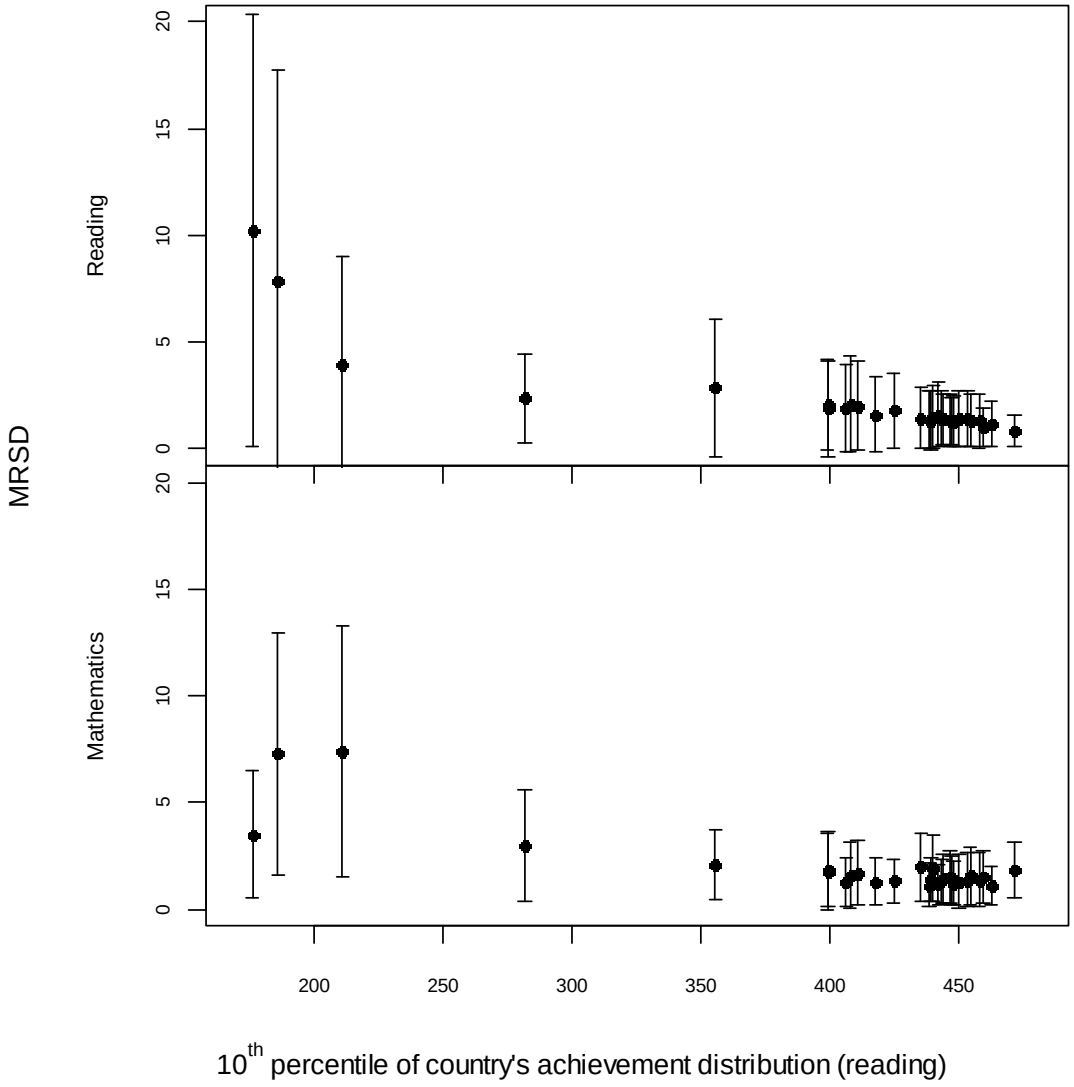
The variances between conditions did not differ significantly (F=1.38, p=0.25). The results of the t-test showed that the differences between content domains were statistically significant and medium in size (t(56)=2.28; p=.027, d=0.60). In the domain of mathematics the differences were greater than in the field of reading ($M_{mathematics}$=0.98, $M_{reading}$=0.57).

Figure 4.21: Reference 5th percentile of countries (in reading) with the corresponding MRSD and standard deviation of MRSD under different conditions



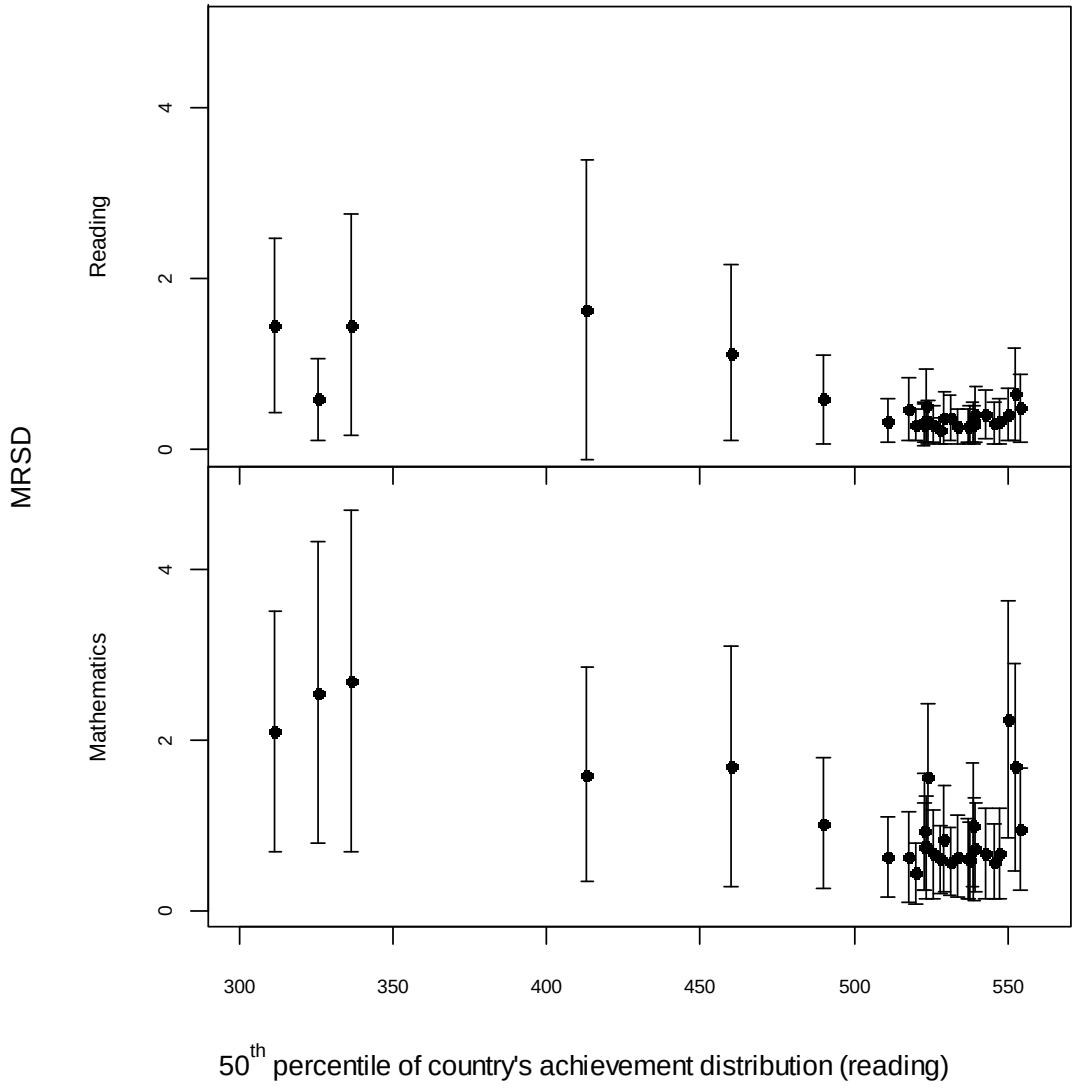
The variances between conditions did not differ significantly (F=0.54, p=0.47). The results of the t-test showed that the differences between content domains were not statistically significant and the effect size was trivial (t(56)=0.66; p=.509, d=0.18).

Figure 4.22: Reference 10th percentile of countries (in reading) with the corresponding MRSD and standard deviation of MRSD under different conditions



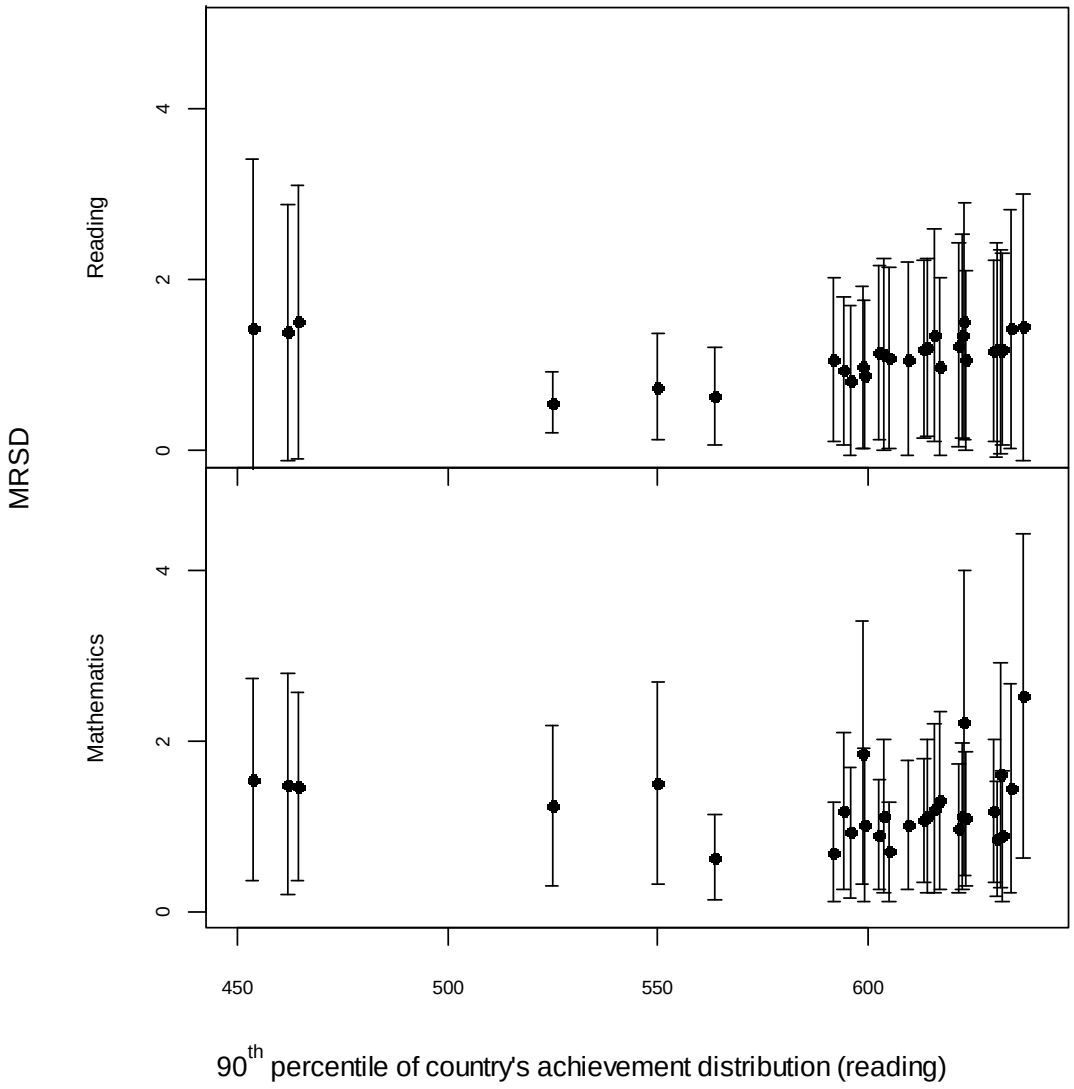
The variances between conditions did not differ significantly (F=0.34, p=0.56). The results of the t-test showed that the differences between content domains were not statistically significant the effect size was trivial (t(56)=0.29; p=.775, d=0.08).

Figure 4.23: Reference 50th percentile of countries (in reading) with the corresponding MRSD and standard deviation of MRSD under different conditions



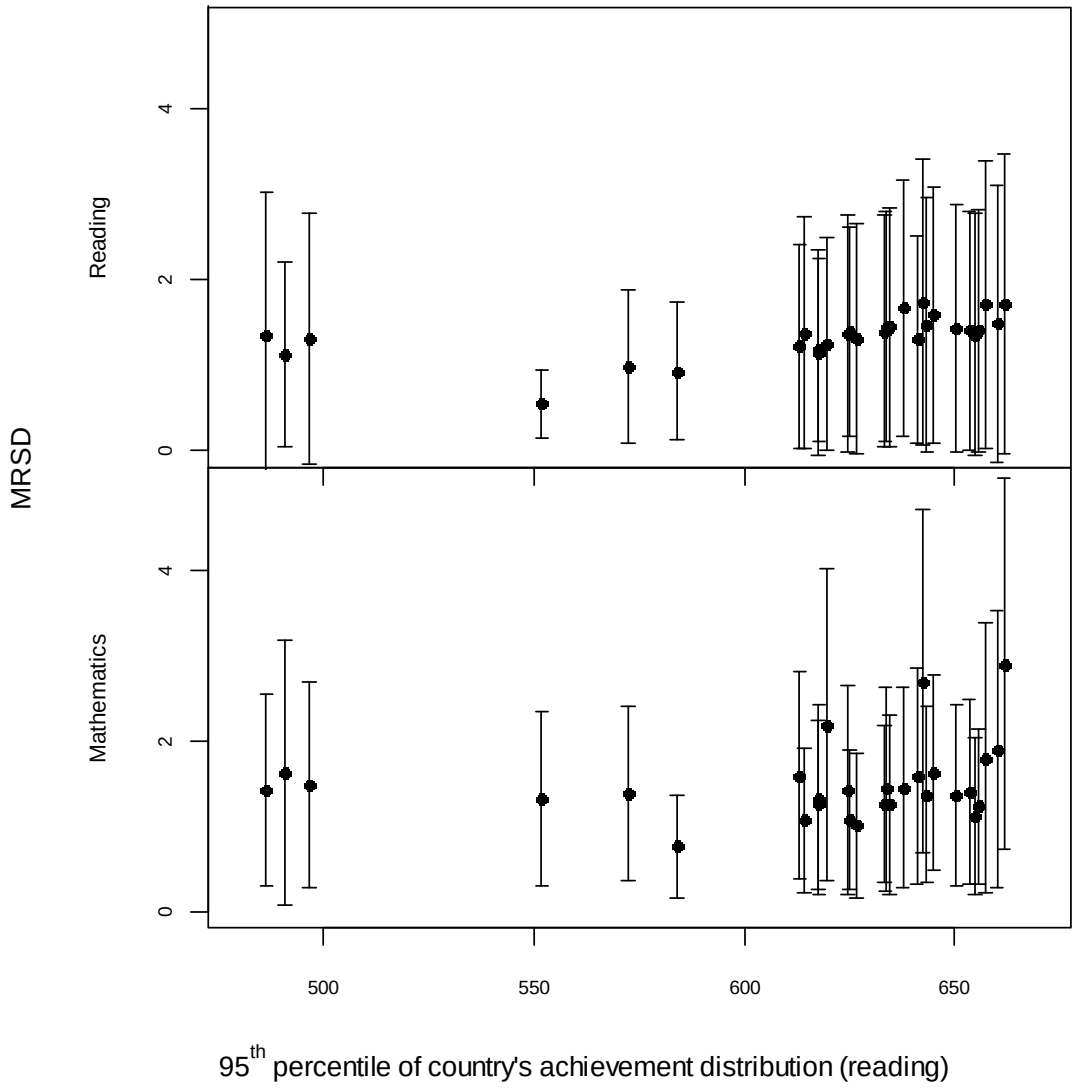
The variances between conditions differed significantly (F=8.70, p=0.005). The results of the t-test showed that the differences between content domains were statistically significant and the effect size was large (t(56)=3.92; p<0.001, d=1.02). In the domain of mathematics, the differences were greater than in the field of reading ($M_{mathematics}$=1.07, $M_{reading}$=0.52).

Figure 4.24: Reference 90th percentile of countries (in reading) with the corresponding MRSD and standard deviation of MRSD under different conditions



The variances between conditions differed significantly (F=4.29, p=0.043). The results of the t-test showed that the differences between content domains were not statistically significant, the effect size was small (t(45)=1.13; p=.265, d=0.30).

Figure 4.25: Reference 95th percentile of countries (in reading) with the corresponding MRSD and standard deviation of MRSD under different conditions



The variances between conditions did not differ significantly (F=3.60, p=0.063). The results of the t-test showed that the differences between content domains were not statistically significant but the effects were found to be small in size (t(45)=1.49; p=.141, d=0.39).

Table 4.16: Descriptives for MRSD across conditions by gender

Category	Condition	*M*	*Me*	*SD*	*Min*	*Max*
girls	reading	0.48	0.33	0.48	0.18	2.54
	mathematics	0.98	0.69	0.60	0.34	2.56
boys	reading	0.70	0.39	0.82	0.22	3.63
	mathematics	1.00	0.58	0.87	0.32	3.76

Note: M=mean, *Me*=median, *SD*=standard deviation, *Min*=minimum, *Max*=maximum.

The variances between conditions in girls differed significantly (F=4.15, p=0.046). The results of the t-test showed that the differences between content domains were statistically significant and the effect was large in size (t(53)=3.50, p=.001, d=0.92). In the domain of mathematics the differences were greater than in the field of reading.

The variances between conditions in boys did not differ significantly (F=0.35, p=0.558). The results of the t-test showed that the differences between content domains were not statistically significant and the effect was small in size (t(56)=1.34; p=.186, d=0.35).

Table 4.17: Descriptives for MRSD across conditions for number of books

Category	Condition	*M*	*Me*	*SD*	*Min*	*Max*
missing	reading	1.40	1.19	1.04	0.50	5.94
	mathematics	1.55	1.30	1.15	0.56	5.96
0 – 10 books	reading	1.05	0.88	0.56	0.46	3.24
	mathematics	1.19	0.98	0.61	0.55	3.08
11 – 25 books	reading	0.63	0.50	0.38	0.23	2.07
	mathematics	1.03	0.77	0.56	0.46	2.66
26 – 100 books	reading	0.45	0.36	0.30	0.17	1.47
	mathematics	0.91	0.63	0.57	0.26	2.29
101 – 200 books	reading	0.49	0.40	0.34	0.19	1.95
	mathematics	0.95	0.65	0.70	0.25	3.18
more than 200 books	reading	0.48	0.39	0.28	0.17	1.36
	mathematics	0.94	0.62	0.71	0.38	3.13

Note: M=mean, *Me*=median, *SD*=standard deviation, *Min*=minimum, *Max*=maximum.

Variances in category “missing” did not differ significantly (F=0.11, p<0.739). The average MRSD in conditions with different content domains also did not differ significantly, the effect size was trivial (t(56)=0.51, p<0.611, d=0.13).

Variances in category “0 – 10 books” did not differ significantly (F=0.34, p<0.564). The average MRSD in conditions with different content domains also did not differ significantly, the effect size was small (t(56)=0.92, p<0.361, d=0.24).

Variances in category "11 – 25 books" differed significantly (F=4.61, p<0.036). The average MRSD in conditions with different content domains also differed significantly, the effect size was large (t(49)=3.15, p=0.003, d=0.84). In the domain of mathematics, the differences were greater (M=1.03) than in the field of reading (M=0.63).

Variances in category "26 – 100 books" differed significantly (F=12.06, p<0.001). The average MRSD in conditions with different content domains also differed significantly, the effect size was large (t(42)=3.88, p<0.001, d=1.04). In the domain of mathematics, the differences were greater (M=0.91) than in the field of reading (M=0.45).

Variances in category "101 – 200 books" differed significantly (F=11.42, p<0.001). The average MRSD in conditions with different content domains also differed significantly, the effect size was large (t(40)=3.17, p=0.002, d=0.85). In the domain of mathematics, the differences were greater (M=0.95) than in the field of reading (M=0.49).

Variances in category "more than 200 books" differed significantly (F=12.56, p=0.001). The average MRSD in conditions with different content domains also differed significantly, the effect size was large (t(36)=3.24, p=0.003, d=0.87). In the domain of mathematics, the differences were greater (M=0.94) than in the field of reading (M=0.48).

4.8 Association of achievement scores with MRSD and standard deviation of MRSD across conditions

In this section, a short summary of the average mean achievement and percentiles across all investigated conditions is presented. In addition to the MRSD and its standard deviation, their association with achievement scores was observed. In Table 4.18, the MRSD of countries was averaged for every condition. The average MRSD is presented together with the standard deviation, minimum and maximum MRSD values across all conditions.

Table 4.18: Descriptives for MRSD across conditions

Statistic of interest	Condition	*M*	*SD*	*Min*	*Max*
Average country achievement					
	2 countries	1.43	2.13	0.36	14.05
	3 countries	1.23	1.74	0.31	11.45
	4 countries	1.15	1.68	0.28	11.12
	6 countries	0.94	1.38	0.25	9.08
	10 countries	0.66	0.89	0.17	5.94
	Low achieving countries	0.77	1.23	0.12	7.80
	High achieving countries	2.59	4.49	0.12	28.09
	Rasch-Rasch	0.08	0.08	0.03	0.47
	Rasch-3PL 2PL GPCM	3.58	4.78	0.13	26.87
	3PL 2PL GPCM -3PL 2PL GPCM	0.66	0.89	0.17	5.94
	Reading	0.57	0.63	0.16	3.11
	Mathematics	0.98	0.72	0.30	3.06
5^{th} percentile					
	2 countries	6.70	8.15	2.84	54.53
	3 countries	5.49	6.67	2.50	45.03
	4 countries	5.16	6.43	2.22	43.64
	6 countries	4.08	5.23	1.74	35.68
	10 countries	2.81	3.39	1.21	23.07
	Low achieving countries	2.36	5.21	0.61	35.10
	High achieving countries	14.18	16.86	1.72	103.13
	Rasch-Rasch	0.41	0.18	0.24	1.17
	Rasch-3PL 2PL GPCM	16.63	12.34	2.19	64.55
	3PL 2PL GPCM -3PL 2PL GPCM	2.81	3.39	1.21	23.07
	Reading	2.75	2.59	1.25	12.92
	Mathematics	2.35	1.87	1.36	8.96
10^{th} percentile					
	2 countries	5.33	6.63	1.82	44.42
	3 countries	4.42	5.42	1.66	36.61
	4 countries	4.15	5.21	1.32	35.47
	6 countries	3.29	4.22	1.22	28.86
	10 countries	2.29	2.75	0.86	18.82
	Low achieving countries	1.99	4.09	0.47	27.42
	High achieving countries	11.36	13.71	3.14	85.37
	Rasch-Rasch	0.33	0.14	0.18	0.94
	Rasch-3PL 2PL GPCM	14.39	11.70	5.38	63.90
	3PL 2PL GPCM -3PL 2PL GPCM	2.29	2.75	0.86	18.82
	Reading	2.17	2.03	0.80	10.23
	Mathematics	2.04	1.56	1.11	7.40
50^{th} percentile					
	2 countries	1.42	1.32	0.48	7.47
	3 countries	1.21	1.09	0.40	6.21

Statistic of interest	Condition	*M*	*SD*	*Min*	*Max*
	4 countries	1.12	1.04	0.35	5.95
	6 countries	0.93	0.85	0.36	4.74
	10 countries	0.66	0.56	0.22	3.20
	Low achieving countries	0.78	0.74	0.17	3.44
	High achieving countries	2.28	3.19	0.12	16.82
	Rasch-Rasch	0.16	0.07	0.09	0.44
	Rasch-3PL 2PL GPCM	4.11	4.75	0.12	27.51
	3PL 2PL GPCM -3PL 2PL GPCM	0.66	0.56	0.22	3.20
	Reading	0.52	0.39	0.22	1.64
	Mathematics	1.07	0.65	0.44	2.68
90^{th} percentile					
	2 countries	2.50	0.55	1.05	3.55
	3 countries	2.19	0.56	0.75	3.36
	4 countries	2.04	0.52	0.63	3.22
	6 countries	1.61	0.41	0.60	2.49
	10 countries	1.18	0.31	0.46	2.12
	Low achieving countries	0.84	0.32	0.29	1.67
	High achieving countries	5.79	2.21	0.24	9.46
	Rasch-Rasch	0.28	0.08	0.15	0.54
	Rasch-3PL 2PL GPCM	8.24	3.70	0.44	16.73
	3PL 2PL GPCM -3PL 2PL GPCM	1.18	0.31	0.46	2.12
	Reading	1.13	0.25	0.56	1.51
	Mathematics	1.24	0.43	0.63	2.52
95^{th} percentile					
	2 countries	3.03	0.63	1.49	4.24
	3 countries	2.66	0.64	1.24	3.91
	4 countries	2.41	0.55	1.09	3.69
	6 countries	1.93	0.46	0.77	2.85
	10 countries	1.41	0.35	0.54	2.32
	Low achieving countries	0.96	0.35	0.46	1.87
	High achieving countries	7.14	2.28	0.69	10.70
	Rasch-Rasch	0.37	0.10	0.21	0.60
	Rasch-3PL 2PL GPCM	10.32	4.13	1.58	19.44
	3PL 2PL GPCM -3PL 2PL GPCM	1.41	0.35	0.54	2.32
	Reading	1.35	0.25	0.55	1.75
	Mathematics	1.49	0.46	0.76	2.90

Note: M=mean, *SD*=standard deviation, *Min*=minimum, *Max*=maximum.

From the results presented in Table 4.18, it can be seen that the largest difference with the reference scores was based on inclusion of higher achieving countries in the calibration sample and when comparing the Rasch model results to the 3PL 2PL and GPCM models. In these two conditions, the highest MRSD values were obtained for all investigated statistics (means and percentiles). In general, extreme percentiles (5^{th}, 10^{th}, 90^{th} and 95^{th})

show larger MRSD values than the averages (mean and 50th percentile). Nevertheless, the greatest MRSD values were observed for lower percentiles (5th and 10th). For these two statistics, the conditions with two, three and four included countries also showed greater MRSD values.

In Table 4.19, the standard deviation of MRSD (that was obtained from 100 repetitions within a condition for each country) of countries was averaged for every condition. The average standard deviation MRSD is presented together with the standard deviation, minimum and maximum MRSD values across all conditions.

Table 4.19: Descriptives for standard deviation of MRSD across conditions

Statistic of interest	Condition	*M(SD)*	*SD(SD)*	*Min(SD)*	*Max(SD)*
SD(Average country achievement)					
	2 countries	0.94	1.07	0.26	6.95
	3 countries	0.88	1.14	0.25	7.50
	4 countries	0.83	1.08	0.23	7.07
	6 countries	0.70	0.98	0.18	6.40
	10 countries	0.54	0.78	0.12	5.06
	Low achieving countries	0.31	0.34	0.09	2.17
	High achieving countries	0.25	0.22	0.08	1.39
	Rasch-Rasch	0.06	0.05	0.02	0.30
	Rasch-3PL 2PL GPCM	0.09	0.10	0.03	0.54
	3PL 2PL GPCM -3PL 2PL GPCM	0.54	0.78	0.12	5.06
	Reading	0.49	0.57	0.14	2.42
	Mathematics	0.73	0.51	0.27	2.35
SD(5th percentile)					
	2 countries	4.14	4.06	2.00	24.67
	3 countries	3.95	4.33	1.71	27.08
	4 countries	3.77	4.08	1.64	25.29
	6 countries	3.28	3.77	1.32	23.94
	10 countries	2.44	3.03	1.00	19.70
	Low achieving countries	0.95	1.12	0.42	7.91
	High achieving countries	1.08	0.75	0.47	4.12
	Rasch-Rasch	0.29	0.12	0.16	0.74
	Rasch-3PL 2PL GPCM	0.48	0.22	0.25	1.38
	3PL 2PL GPCM -3PL 2PL GPCM	2.44	3.03	1.00	19.70
	Reading	2.93	2.91	1.24	12.68
	Mathematics	1.99	1.45	1.11	6.99
SD(10th percentile)					
	2 countries	3.32	3.31	1.21	20.32
	3 countries	3.14	3.54	1.23	22.35
	4 countries	3.00	3.35	0.97	20.94

Statistic of interest	Condition	*M(SD)*	*SD(SD)*	*Min(SD)*	*Max(SD)*
	6 countries	2.58	3.08	0.86	19.66
	10 countries	1.94	2.46	0.64	16.05
	Low achieving countries	0.81	0.93	0.33	6.53
	High achieving countries	0.91	0.63	0.54	3.61
	Rasch-Rasch	0.24	0.10	0.15	0.62
	Rasch-3PL 2PL GPCM	0.40	0.17	0.23	1.12
	3PL 2PL GPCM -3PL 2PL GPCM	1.94	2.46	0.64	16.05
	Reading	2.26	2.29	0.73	10.09
	Mathematics	1.69	1.23	0.90	5.91
SD(50^{th} percentile)					
	2 countries	0.96	0.70	0.33	4.35
	3 countries	0.87	0.72	0.33	4.37
	4 countries	0.80	0.71	0.27	4.25
	6 countries	0.67	0.61	0.26	3.48
	10 countries	0.52	0.48	0.17	2.55
	Low achieving countries	0.34	0.24	0.12	1.40
	High achieving countries	0.30	0.21	0.07	1.31
	Rasch-Rasch	0.12	0.05	0.06	0.31
	Rasch-3PL 2PL GPCM	0.18	0.08	0.10	0.53
	3PL 2PL GPCM -3PL 2PL GPCM	0.52	0.48	0.17	2.55
	Reading	0.43	0.37	0.15	1.75
	Mathematics	0.78	0.44	0.35	2.00
SD(90^{th} percentile)					
	2 countries	1.69	0.42	0.66	2.64
	3 countries	1.56	0.44	0.53	2.36
	4 countries	1.46	0.40	0.47	2.23
	6 countries	1.23	0.33	0.42	1.88
	10 countries	0.91	0.25	0.31	1.46
	Low achieving countries	0.50	0.13	0.22	0.78
	High achieving countries	0.56	0.19	0.16	0.96
	Rasch-Rasch	0.21	0.06	0.12	0.42
	Rasch-3PL 2PL GPCM	0.33	0.09	0.20	0.62
	3PL 2PL GPCM -3PL 2PL GPCM	0.91	0.25	0.31	1.46
	Reading	1.11	0.32	0.36	1.98
	Mathematics	0.98	0.34	0.50	1.89
SD(95^{th} percentile)					
	2 countries	2.06	0.49	0.81	3.03
	3 countries	1.92	0.51	0.68	2.90
	4 countries	1.79	0.44	0.70	2.67
	6 countries	1.51	0.36	0.57	2.24
	10 countries	1.09	0.27	0.43	1.68
	Low achieving countries	0.57	0.15	0.33	0.97
	High achieving countries	0.68	0.20	0.22	1.06
	Rasch-Rasch	0.27	0.07	0.15	0.43

Statistic of interest	Condition	*M(SD)*	*SD(SD)*	*Min(SD)*	*Max(SD)*
	Rasch-3PL 2PL GPCM	0.44	0.13	0.23	0.74
	3PL 2PL GPCM -3PL 2PL GPCM	1.09	0.27	0.43	1.68
	Reading	1.33	0.29	0.39	1.75
	Mathematics	1.19	0.36	0.61	2.17

Note: M(SD)=mean of the standard deviation, *SD(SD)*=standard deviation of the mean standard deviation within a condition, *Min(SD)*=minimum value for standard deviation, *Max*=maximum value for standard deviation.

The highest variability was observed in the conditions with two, three and four countries in the calibration sample. The results were consistent for all investigated statistics. The most stable results were observed in the Rasch-Rasch comparison. Larger variability was again present at the extreme percentiles, the largest in the lower percentiles. This was observed for all statistics across all conditions.

Furthermore, we were interested in the relationship of MRSD and achievement. From the figures presented in previous chapters, some trends could be observed. Because the association of MRSD and achievement does not always appear linear, we fitted a linear and a quadratic regression model for all conditions and statistics of interest. The estimated coefficients of determination of both regression models and their comparison are presented in Table 4.20 (for the association of achievement and MRSD) and Table 4.21 (for the association of achievement and standard deviation of MRSD). The estimated regression coefficients are presented in Appendix A.

Table 4.20: Comparison of linear and quadratic regression models across conditions for MRSD with achievement

Statistic of interest	Condition	R^2_{lin}	$adj\ R^2_{lin}$	R^2_{qua}	$adj\ R^2_{qua}$	F	p	p_{lin}
Average country achievement								
	2 countries	0.57	0.56	0.67	0.65	12.49	0.001	
	3 countries	0.58	0.57	0.69	0.67	13.67	0.001	
	4 countries	0.56	0.55	0.66	0.64	12.29	0.001	
	6 countries	0.57	0.56	0.67	0.65	11.75	0.001	
	10 countries	0.57	0.56	0.69	0.67	15.05	0.000	
	Low achieving countries	0.51	0.50	0.59	0.57	8.46	0.006	
	High achieving countries	0.59	0.58	0.66	0.65	9.48	0.004	
	Rasch-Rasch	0.69	0.68	0.89	0.88	76.98	0.000	
	Rasch-3PL 2PL GPCM	0.68	0.67	0.79	0.78	22.99	0.000	
	3PL 2PL GPCM -3PL 2PL GPCM	0.57	0.56	0.69	0.67	15.05	0.000	
	Reading	0.77	0.76	0.79	0.78	2.36	0.136	0.000

Statistic of interest	Condition	R^2_{lin}	$adj\ R^2_{lin}$	R^2_{qua}	$adj\ R^2_{qua}$	F	p	p_{lin}
	Mathematics	0.44	0.42	0.88	0.87	90.16	0.000	
5th percentile								
	2 countries	0.55	0.54	0.81	0.80	58.85	0.000	
	3 countries	0.54	0.53	0.82	0.81	64.42	0.000	
	4 countries	0.52	0.51	0.81	0.80	61.47	0.000	
	6 countries	0.53	0.51	0.80	0.79	56.84	0.000	
	10 countries	0.53	0.52	0.82	0.81	66.10	0.000	
	Low achieving countries	0.40	0.38	0.60	0.58	21.69	0.000	
	High achieving countries	0.54	0.53	0.84	0.84	82.44	0.000	
	Rasch-Rasch	0.56	0.55	0.81	0.81	56.56	0.000	
	Rasch-3PL 2PL GPCM	0.55	0.54	0.86	0.85	89.12	0.000	
	3PL 2PL GPCM -3PL 2PL GPCM	0.53	0.52	0.82	0.81	66.10	0.000	
	Reading	0.77	0.76	0.87	0.87	22.16	0.000	
	Mathematics	0.70	0.69	0.91	0.91	63.21	0.000	
10th percentile								
	2 countries	0.59	0.58	0.80	0.79	42.57	0.000	
	3 countries	0.58	0.57	0.81	0.80	48.31	0.000	
	4 countries	0.56	0.55	0.79	0.78	44.36	0.000	
	6 countries	0.56	0.55	0.78	0.77	41.84	0.000	
	10 countries	0.56	0.55	0.80	0.79	49.44	0.000	
	Low achieving countries	0.44	0.43	0.60	0.58	16.70	0.000	
	High achieving countries	0.60	0.59	0.85	0.84	66.98	0.000	
	Rasch-Rasch	0.52	0.50	0.82	0.81	70.73	0.000	
	Rasch-3PL 2PL GPCM	0.59	0.58	0.87	0.86	87.57	0.000	
	3PL 2PL GPCM -3PL 2PL GPCM	0.56	0.55	0.80	0.79	49.44	0.000	
	Reading	0.80	0.79	0.87	0.86	13.33	0.001	
	Mathematics	0.68	0.67	0.92	0.91	71.51	0.000	
50th percentile								
	2 countries	0.66	0.65	0.67	0.65	1.57	0.217	0.000
	3 countries	0.68	0.68	0.70	0.68	1.55	0.220	0.000
	4 countries	0.66	0.66	0.68	0.66	1.58	0.215	0.000
	6 countries	0.72	0.71	0.73	0.72	2.05	0.159	0.000
	10 countries	0.74	0.74	0.76	0.75	3.22	0.080	0.000
	Low achieving countries	0.65	0.64	0.68	0.67	4.25	0.045	
	High achieving countries	0.48	0.47	0.49	0.46	0.26	0.612	0.000
	Rasch-Rasch	0.68	0.68	0.74	0.73	9.02	0.004	
	Rasch-3PL 2PL GPCM	0.69	0.68	0.70	0.68	1.02	0.317	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	0.74	0.74	0.76	0.75	3.22	0.080	0.000
	Reading	0.62	0.60	0.65	0.62	2.60	0.119	0.000
	Mathematics	0.33	0.31	0.80	0.78	60.80	0.000	
90th percentile								
	2 countries	0.04	0.01	0.65	0.63	74.24	0.000	
	3 countries	0.11	0.09	0.69	0.68	79.37	0.000	
	4 countries	0.07	0.05	0.60	0.58	55.80	0.000	

Statistic of interest	Condition	R^2_{lin}	$adj\ R^2_{lin}$	R^2_{qua}	$adj\ R^2_{qua}$	F	p	p_{lin}
	6 countries	0.10	0.08	0.67	0.66	73.12	0.000	
	10 countries	0.17	0.15	0.65	0.64	58.31	0.000	
	Low achieving countries	0.07	0.04	0.09	0.04	0.96	0.332	0.089
	High achieving countries	0.03	0.01	0.74	0.73	117.65	0.000	
	Rasch-Rasch	0.02	0.00	0.27	0.23	14.07	0.001	
	Rasch-3PL 2PL GPCM	0.03	0.00	0.74	0.73	115.65	0.000	
	3PL 2PL GPCM -3PL 2PL GPCM	0.17	0.15	0.65	0.64	58.31	0.000	
	Reading	0.00	-0.04	0.70	0.68	60.80	0.000	
	Mathematics	0.05	0.01	0.75	0.73	72.65	0.000	
95^{th} percentile								
	2 countries	0.62	0.61	0.77	0.76	27.26	0.000	
	3 countries	0.66	0.65	0.80	0.79	30.46	0.000	
	4 countries	0.57	0.56	0.69	0.68	16.63	0.000	
	6 countries	0.61	0.60	0.74	0.73	21.53	0.000	
	10 countries	0.63	0.62	0.73	0.72	15.59	0.000	
	Low achieving countries	0.01	-0.01	0.01	-0.03	0.16	0.695	0.498
	High achieving countries	0.38	0.36	0.72	0.70	50.68	0.000	
	Rasch-Rasch	0.26	0.25	0.48	0.45	17.29	0.000	
	Rasch-3PL 2PL GPCM	0.23	0.21	0.71	0.70	71.31	0.000	
	3PL 2PL GPCM -3PL 2PL GPCM	0.63	0.62	0.73	0.72	15.59	0.000	
	Reading	0.29	0.27	0.62	0.59	22.37	0.000	
	Mathematics	0.23	0.20	0.89	0.88	150.30	0.000	

Note: R^2_{lin}= coefficient of determination for a linear regression model, $adj\ R^2_{lin}$= adjusted coefficient of determination for a linear regression model, R^2_{qua}= coefficient of determination for a quadratic regression model, $adj\ R^2_{qua}$ = adjusted coefficient of determination for a quadratic regression model, F = F-test statistic when comparing the linear and quadratic model with ANOVA and the corresponding p-value, p_{lin} = in case of insignificant difference between linear and quadratic model the significance of the linear coefficient is presented.

While observing the average achievement of countries, the association between MRSD and achievement scores seems to follow a quadratic pattern. In all conditions, the inclusion of a quadratic term into the model yielded a significantly better fit than the linear one, except in the content domain of reading (where the difference is not significant). The largest change in the estimated coefficient of determination was for the content domain of mathematics (the R^2 increased from 0.44 to 0.88) and in the Rasch-Rasch condition (from 0.69 to 0.89).

For the 50^{th} percentile in most conditions, the linear model showed the more appropriate fit. In two conditions (low and Rasch-Rasch) the quadratic model fitted significantly better, but the change in the coefficient of the determination was not exceptionally large (0.03 and 0.06). In the content domain of mathematics, the quadratic model fitted significantly better

than the linear and the estimated coefficient of determination increased substantially (from 0.33 to 0.80).

The 5th and the 10th percentiles were best described by a quadratic model. The change in R^2 in all conditions was larger than for the average. In the 90th percentile, no linear association could be observed (R^2 between 0.00 and 0.17). The quadratic association was evident in all conditions (with the smallest in the Rasch-Rasch condition), however only in the condition when low achieving countries were included in the calibration sample. In the last case, we can conclude that the achievement and the MRSD were not associated. Similar results can be found for the 95th percentile, where once again in the case of low achieving countries condition no association could be observed.

In general, the results suggested that for extreme values of countries' distribution, the association of achievement and MRSD was quadratic. This means that the average achievement scores in percentiles showed the least MRSD compared to higher or lower achievement scores, where the MRSD values were higher. This association was especially evident for the 90th percentile.

In Table 4.21, the association of achievement and standard deviation of MRSD is presented, following the same logic as before. A comparison of linear and quadratic regression models is presented for the association of achievement and standard deviation of MRSD.

Table 4.21: Comparison of linear and quadratic regression model across conditions for standard deviation of MRSD with achievement

Statistic of interest	Condition	R^2_{lin}	$R^{2\,adj}_{lin}$	R^2_{qua}	$R^{2\,adj}_{qua}$	F	p	p_{lin}
SD(Average country achievement)								
	2 countries	0.64	0.63	0.77	0.76	24.95	0.000	
	3 countries	0.59	0.58	0.73	0.71	20.80	0.000	
	4 countries	0.60	0.59	0.72	0.71	18.68	0.000	
	6 countries	0.61	0.60	0.72	0.71	16.73	0.000	
	10 countries	0.60	0.59	0.69	0.68	13.25	0.001	
	Low achieving countries	0.56	0.55	0.67	0.65	13.39	0.001	
	High achieving countries	0.68	0.67	0.83	0.82	35.39	0.000	
	Rasch-Rasch	0.71	0.70	0.89	0.89	73.97	0.000	
	Rasch-3PL 2PL GPCM	0.67	0.67	0.86	0.86	59.51	0.000	
	3PL 2PL GPCM -3PL 2PL GPCM	0.60	0.59	0.69	0.68	13.25	0.001	
	Reading	0.76	0.75	0.76	0.74	0.13	0.724	0.000

Statistic of interest	Condition	R^2_{lin}	$adj\ R^2_{lin}$	R^2_{qua}	$adj\ R^2_{qua}$	F	p	p_{lin}
	Mathematics	0.56	0.54	0.87	0.86	59.80	0.000	
SD(5th percentile)								
	2 countries	0.66	0.65	0.89	0.88	83.87	0.000	
	3 countries	0.62	0.62	0.87	0.86	80.32	0.000	
	4 countries	0.62	0.61	0.87	0.86	77.94	0.000	
	6 countries	0.61	0.60	0.86	0.85	72.96	0.000	
	10 countries	0.59	0.58	0.85	0.84	71.54	0.000	
	Low achieving countries	0.50	0.48	0.75	0.73	41.21	0.000	
	High achieving countries	0.77	0.76	0.95	0.95	153.91	0.000	
	Rasch-Rasch	0.58	0.57	0.75	0.74	29.24	0.000	
	Rasch-3PL 2PL GPCM	0.59	0.58	0.84	0.83	63.31	0.000	
	3PL 2PL GPCM -3PL 2PL GPCM	0.59	0.58	0.85	0.84	71.54	0.000	
	Reading	0.81	0.80	0.90	0.89	23.44	0.000	
	Mathematics	0.73	0.72	0.91	0.91	55.16	0.000	
SD(10th percentile)								
	2 countries	0.71	0.71	0.89	0.88	64.11	0.000	
	3 countries	0.66	0.65	0.86	0.85	59.78	0.000	
	4 countries	0.66	0.66	0.86	0.85	58.01	0.000	
	6 countries	0.65	0.64	0.84	0.84	52.99	0.000	
	10 countries	0.63	0.62	0.83	0.82	49.91	0.000	
	Low achieving countries	0.54	0.53	0.76	0.75	37.75	0.000	
	High achieving countries	0.76	0.75	0.94	0.93	118.01	0.000	
	Rasch-Rasch	0.54	0.53	0.78	0.77	46.73	0.000	
	Rasch-3PL 2PL GPCM	0.56	0.55	0.83	0.82	68.47	0.000	
	3PL 2PL GPCM -3PL 2PL GPCM	0.63	0.62	0.83	0.82	49.91	0.000	
	Reading	0.83	0.83	0.89	0.88	12.95	0.001	
	Mathematics	0.73	0.72	0.91	0.90	52.85	0.000	
SD(50th percentile)								
	2 countries	0.68	0.67	0.71	0.69	3.30	0.077	0.000
	3 countries	0.72	0.71	0.74	0.73	4.34	0.043	
	4 countries	0.68	0.67	0.70	0.69	3.15	0.083	0.000
	6 countries	0.68	0.67	0.69	0.68	2.10	0.155	0.000
	10 countries	0.70	0.69	0.70	0.69	0.89	0.351	0.000
	Low achieving countries	0.68	0.67	0.70	0.69	3.96	0.053	0.000
	High achieving countries	0.56	0.55	0.56	0.54	0.08	0.777	0.000
	Rasch-Rasch	0.64	0.63	0.71	0.70	10.53	0.002	
	Rasch-3PL 2PL GPCM	0.66	0.65	0.72	0.70	8.67	0.005	
	3PL 2PL GPCM -3PL 2PL GPCM	0.70	0.69	0.70	0.69	0.89	0.351	0.000
	Reading	0.52	0.50	0.61	0.58	6.03	0.021	
	Mathematics	0.43	0.40	0.78	0.76	41.57	0.000	
SD(90th percentile)								
	2 countries	0.03	0.00	0.56	0.54	51.49	0.000	
	3 countries	0.07	0.05	0.71	0.69	91.05	0.000	
	4 countries	0.05	0.03	0.64	0.63	69.25	0.000	

Statistic of interest	Condition	R^2_{lin}	$adj\ R^2_{lin}$	R^2_{qua}	$adj\ R^2_{qua}$	F	p	p_{lin}
	6 countries	0.07	0.05	0.64	0.62	66.75	0.000	
	10 countries	0.04	0.02	0.66	0.64	76.26	0.000	
	Low achieving countries	0.14	0.12	0.29	0.26	8.95	0.005	
	High achieving countries	0.00	-0.02	0.79	0.78	161.87	0.000	
	Rasch-Rasch	0.02	0.00	0.24	0.20	12.04	0.001	
	Rasch-3PL 2PL GPCM	0.11	0.09	0.36	0.33	16.60	0.000	
	3PL 2PL GPCM -3PL 2PL GPCM	0.04	0.02	0.66	0.64	76.26	0.000	
	Reading	0.05	0.01	0.81	0.80	105.86	0.000	
	Mathematics	0.05	0.02	0.79	0.77	89.62	0.000	
SD(95th percentile)								
	2 countries	0.45	0.44	0.66	0.64	25.07	0.000	
	3 countries	0.55	0.53	0.77	0.76	42.04	0.000	
	4 countries	0.53	0.52	0.74	0.73	34.76	0.000	
	6 countries	0.55	0.54	0.74	0.72	30.00	0.000	
	10 countries	0.52	0.51	0.73	0.72	32.83	0.000	
	Low achieving countries	0.28	0.26	0.32	0.29	2.51	0.120	0.000
	High achieving countries	0.12	0.10	0.67	0.66	71.45	0.000	
	Rasch-Rasch	0.25	0.23	0.48	0.46	19.25	0.000	
	Rasch-3PL 2PL GPCM	0.30	0.29	0.50	0.47	16.39	0.000	
	3PL 2PL GPCM -3PL 2PL GPCM	0.52	0.51	0.73	0.72	32.83	0.000	
	Reading	0.13	0.10	0.64	0.61	36.17	0.000	
	Mathematics	0.16	0.13	0.90	0.89	192.45	0.000	

Note: R^2_{lin}= coefficient of determination for a linear regression model, *adj* R^2_{lin}= adjusted coefficient of determination for a linear regression model, R^2_{qua}= coefficient of determination for a quadratic regression model, *adj* R^2_{qua} = adjusted coefficient of determination for a quadratic regression model, F = F-test statistic when comparing the linear and quadratic model with ANOVA and the corresponding p-value, p_{lin} = in case of insignificant difference between linear and quadratic model the significance of the linear coefficient is presented.

The results for the association of standard deviation of MRSD with achievement were highly similar to the previous results. For the SD of mean MRSD values in all conditions, the quadratic model fit significantly better than the linear one, except in the content domain of reading (where the difference was not significant). The biggest change in the estimated coefficient of determination was again for the content domain of mathematics, and this time also for the Rasch-3PL 2PL GMCM condition.

The standard deviations of 5th and 10th percentiles showed a quadratic association with achievement in all conditions. The same holds true for the 90th and 95th percentiles (except in the reading domain, where no association of achievement and standard deviation of MRSD was found).

The linear association was present for standard deviation of MRSD in the 50th percentile in all conditions, except in Rasch-Rasch, Rasch-3PL 2PL GPCM, and the reading and mathematics domains (where the quadratic model fitted better). The increase of the estimated coefficient of determination was largest in the domain of mathematics (from 0.43 to 0.78).

In general, the results show that MRSDs and their standard deviation are consistent across conditions. Moreover, these results point towards a quadratic association with achievement at extreme values of the distribution while a linear relation is present at the average points of the distribution. The coefficients from all models and conditions are presented in Appendix A.

5 Discussion

The focus of this dissertation was to observe the effect of the calibration sample on achievement scores of countries that participate in LSA studies. As a unit, countries were chosen since students' achievement at a country level represents the focus of interest in international LSAs (for example PIRLS, TIMSS and PISA). This explains why all the achievement results were observed on the country level (and also in some subgroups within countries). Students' achievement scores were conditioned on their background data, and the achievement scores were obtained in such a way that they present reliable group estimates within countries. Therefore, countries can be used as sample units. The calibration sample always included a certain number of countries based on the characteristic of interest.

If the IRT model fits the data, a salient feature of IRT models is that trait level estimates with invariant meaning may be obtained from any set of items, and item parameters do not depend on sample characteristics. Parameter invariance was observed under different sample sizes, countries' average achievement scores, the IRT model used and the content domain assessed. In all cases, the main interest was observing the invariance in IRT model parameters and the achievement scores of participating countries. This information is of considerable importance since many researchers and policy makers in participating countries are using the data to inform national institutions about education. Klemenčič (2010) reports that data from international comparative research are increasingly used in the definition of quality indicators of national educational systems, and points out that another aim of these studies is to provide descriptions of the different activities performed in the education systems and their connections with students' achievements. In the following text, the already presented results are explained in the context of previous research in this field.

Usually, the sample sizes in LSA studies are large. In the case of a large sample, even small differences can turn out to be significant. Consequently, we additionally observed and reported the effect size of the statistic of interest. Nevertheless, a statistically significant difference may not be practically relevant, if the effect size is too small. Moreover, if the difference is not significant, the reason could lie in the inadequate power

of the test to prove it. Determining a critical value in social sciences is extremely difficult. Especially for the investigated topic there are not enough studies available to determine an effect size that is practically important. For this reason, the differences and the effect sizes are interpreted in line with the critical value which was determined as a function of α and 1-β, when α=0.05 and 1-β=0.80 (see also chapter 3.8) and according to categories as proposed by Cohen (1992).

5.1 A different number of countries

The first research question was dealing with the sample size of the calibration sample. The main aim was to investigate possible differences in item parameters and proficiency scores when different numbers of countries were included in the calibration sample (countries were treated as units). For different conditions, the calibration sample included two, three, four, six or ten countries. The conditions were selected because of the sample guidelines in the literature as already reported in the methods section. Estimation of more model parameters requires larger samples. Du Toit (2003) reports that sample sizes of 250 examinees are acceptable in research applications and that 500 to 1000 examinees are perfectly sufficient in operational use (more complex models require larger samples). The additional precision gained with bigger sample sizes (beyond 1000) may not justify for the additional computational time or data-collecting costs. In our case, the ten-country condition represents a sample size of approximately 1000, although there were more students participating in each country. Considering the weights used (that sum up to 500 in each country) and the booklet design (where items have approximately 80% of missing values on every item) one country is represented by 100 examinees per item.

Firstly, item parameters were observed. From the results of item parameter comparisons, we can conclude that the most invariant item parameters are step parameters and the difficulty parameter. The item discrimination parameter was less invariant across calibration samples than the item difficulty parameter. Finally, the guessing parameter was found to be the most variant under different conditions. Moreover, in our case, the invariance of the item discrimination parameter increased with an increase in the number of countries included in the calibration sample (when more countries were included in the item parameter estimations). Nevertheless, in all conditions, the correlations of item

parameters with the reference (except for the guessing parameter) were above 0.87. If we take into account the effect size of the correlations, the ten-country condition appeared to be significantly more invariant compared to the two- and three-country conditions for all item parameters except the $step_3$ parameter. The effect sizes were medium and large, and most of them exceeded the predetermined critical values. Even the non-significant differences in $step_3$ parameter were medium or large (according to Cohen's criterion). These findings are in line with the results reported by other researchers. Fan (1998) determined that the item difficulty indexes of IRT were most invariant compared to other parameters in random samples. Adedoyin et al. (2008) reported that item difficulty parameter estimates based on IRT are invariant across varying sample sizes. It should be noted at this point that their sample sizes were of 1000 examinees and larger. Furthermore, Macdonald and Paunonen (2002) concluded from their results that the invariance of the item difficulty estimate is high regardless of the number of items in the test as well as the range of difficulty levels or discrimination values used.

In the opinion of Galdin and Laurencelle (2010) the threshold for Pearson's correlation coefficients that indicate invariance is the value 0.90. Taking this value as a criterion, estimates of all parameters except guessing show sufficient invariance when at least six countries were included in the parameter estimations.

Secondly, the countries' achievement scores were observed. Our main finding was that the average achievement scores of medium and high achieving countries are relatively invariant across conditions. From Figure 4.1 to Figure 4.6, it can be seen that as well as the MRSD, the standard deviation of MRSD is also higher in lower achieving countries (with average achievement below 450 points). The smallest differences were observed for countries with average achievement scores between 500 and 550. The greater variability of the lower achievement scores can be explained by the fact that there are only a few countries with extremely low achievement. The majority (37 out of 45) of countries are crowded between 465 and 557 score points. As Embretson and Reise (2000) report, the accuracy of estimating two different trait levels from test data differs between item sets. If an item set is easy, a low trait level will be more accurately estimated than a high trait level. Similarly, if the calibration sample has relatively low trait levels, the difficulty of easy items will be more accurately estimated than hard items.

Interestingly, there is a country that has low average achievement (350) and lower MRSD values and variability than we would expect due to its position on the lower continuum of the distribution of countries' achievement. In attempting to determine a reason for this result, we reviewed the technical report for PIRLS (Martin et al. 2007). Observing the reported standard errors of countries, the reason for the lower MRSD and lower variability could lie in the sampling design of that country. As can be seen from the international PIRLS 2006 technical report, this country has the smallest standard error of the average achievement among all countries. Apparently, the sampling in this country was extremely efficient, leading to a smaller sampling error.

Furthermore, we could observe the same pattern of MRSD and standard deviation of MRSD across different percentiles. In the 5^{th} and 10^{th} percentile the MRSDs are decreasing when country's average achievement is increasing (in all conditions), and the same holds true for the standard deviation of MRSD, which is also decreasing in increasing average achievement. The same is present for the 50^{th} percentile of countries. In 90^{th} and 95^{th} percentiles the MRSDs are low (on average below 5 score points), but they seem to be increasing when the achievement rises above 600 points. The same happened to the variability. This follows previous research that suggests that extreme levels of abilities are measured with smaller precision (Embretson and Reise 2000, Foy et al. 2010).

The lowest MRSD values for the ten-country condition were also found in the investigated background variables of gender and number of books. The highest MRSD value was found in the least frequent category in number of books. This result is most probably due to the lowest achievement scores of students in this category.

In comparing the achievement scores across conditions, significantly lower differences than the reference scores were observed in the ten-country condition compared to the other conditions. If we consider the effect size, the MRSDs in the ten-country condition were significantly lower than in the two- and three-country conditions for all statistics of interest (except the mean country achievement, 10^{th} percentile, gender and in one category of a number of books at home, i.e. 101 to 200 books). The effect sizes of the differences were small according to Cohen's criterion for non-significant differences, and small, medium or large when differences were significant.

Macdonald and Paunonen (2002) were observing person parameters based on varying item parameter values. They concluded that the results for person statistics accurately estimate

the true abilities of the examinees across all levels of item difficulty values and item discrimination values. Their results suggest that regardless of the measurement framework, test-based decisions regarding person ability estimates will be consistent and accurate. Based on our results from the different number-of-countries conditions, we would suggest caution when interpreting lower achievement scores or extreme values of achievement score distributions, especially at the lower end, when the calibration sample is smaller.

However, the general conclusion from this empirical evidence is that ten countries is a recommended sample size for a calibration sample in LSA studies, in order that the obtained results are invariant. In other words, the average achievement of countries shows negligible differences to the reference scores. When ten countries were included in the item parameter estimations, we obtained remarkably similar item parameters (all correlations were above 0.78) as well as similar achievement scores compared to the reference estimates (the MRSD in the condition of ten countries was 4.18 and the standard deviation was 5.05; in the condition of six countries, the mean difference was almost twice as high: 7.00 with a standard deviation of 7.66).

5.2 Low and high achieving country conditions

The second research question was dealing with the average achievement of countries included in the calibration sample and their effect on the proficiency scores and item parameters. Two conditions were selected for comparisons. Groups of countries were composed based on the average country achievement (one group representing lower achieving countries and the other representing higher achieving countries), each with 15 countries. From each group, a sample of ten countries was then examined. Once again, first, the item parameters were observed.

The correlations in both conditions with the reference were high (above 0.80, except for the guessing parameter), and we could not find any significant differences in the correlation coefficients between conditions, except for the slope parameter. The effect sizes were small for the asymptote and location parameters, medium for the $step_1$ and $step_2$ parameters, and large for the $step_3$ parameter. The correlation of the slope parameter with the reference was significantly higher in higher achieving countries than in lower achieving countries (the effect size did not exceed the critical value but according to Cohen's

criterion it was small). This is in line with the findings of Adedoyin et al. (2008) who reported that item difficulty parameter estimates based on IRT are invariant across different ability groups. When they examined the effect of different ability groups in connection to sample size, they did not find any differences in item parameters. Their sample sizes varied from 1000 to 1900 examinees. In contrast, Hambleton and Swaminathan (1985) report that based on a study the difficulty parameter in different ability groups fails to meet the same invariance as in random groups. Furthermore, Fan (1998) reports that the discrimination parameter across different ability groups fails to meet the same invariance as in random samples.

In the next step, the achievement estimates were observed. The MRSDs between the results obtained by higher achieving countries in the calibration sample were significantly higher than the MRSDs based on the lower achieving sample. The MRSDs when lower achieving countries were included were remarkably similar to the ten-country condition (in the investigation of the sample size). This finding might be regarded as surprising since it was expected that when lower achieving countries were included in the calibration sample, lower achievements would be reproduced more accurately, and when the higher achieving countries were included in the calibration sample, higher achievements would be estimated more accurately. However, it seems that this is not the case. The first thing to consider is the range of achievement scores in the group of lower and higher achieving countries. It turns out that the lower achieving group of countries' achievement scores ranged from 303 to 506 points on average and the higher group from 539 to 557 points. The higher achieving group was much more homogenous in achievement than the lower achieving group. This is most probably the reason for more variant achievement scores in the condition of high achieving countries.

In addition, the differences in percentiles were also significant. The average MRSD values in both conditions were following the same pattern as in the condition of a different number of countries for the 5^{th} and the 10^{th} percentiles. Namely, the MRSD values were decreasing together with the variability while the average values of the percentiles were increasing. However, in the 50^{th}, 90^{th} and 95^{th} percentiles a clear pattern of stable and low MRSD estimates based on lower achieving countries across whole score range (300 to 700) was observed, while an irregular pattern was present in the condition of higher achieving countries.

The lower achieving countries seemed to be the more efficient calibration sample in use when observing achievement scores also across subgroups. The MRSD values based on higher achieving countries were in most categories at least three times higher than in the MRSD values based on the lower achieving countries. Furthermore, the effect sizes for all differences were medium or large and exceed the predetermined critical value. The most surprising result was that the higher achievements (90^{th} and 95^{th} percentiles) were also more efficiently estimated with the lower achieving calibration sample, and the effect sizes were unusually large. This is probably again due to the higher variability of achievement scores in the lower achievement calibration sample and a decidedly homogenous achievement sample in the higher achieving countries. The relationship of achievement and MRSD is described in greater detail in chapter 5.5.

One possible conclusion based on the results is that it is necessary to include lower achieving countries in the calibration sample. It seems that the current cognitive items function differently among higher and lower achieving students. In any case, because of the estimation of trends (and because about half of the items are repeated in the next cycle), sufficient participation of lower achieving countries should be guaranteed. Special attention should be given in avoidance of using an overly homogeneous calibration sample since this results in a greater bias later, if the level of the trait in the overall sample is remarkably different to the level of the trait in the calibration sample.

5.3 Using different IRT models

The third research question was dealing with different IRT models. Reliable trend measurement is one of the major goals of (inter)national assessments. Fostering stability between consecutive assessments is one of the leading considerations in designing an assessment. Essentially, all possible factors that may affect results should be as similar as possible, to rule out these sources of uncertainty. Hickendorff et al. (2009) identify a few of these factors: the sample of students, time of measurement, the assessment format and instruction. Changing any of these factors would require additional and expensive studies. As a result, choices made in the design of a first assessment affect the design of all consecutive assessments. Among the choices to be made prior to the first assessment, the model used is also a fundamental decision.

A comparison of different models is of particular interest since the PISA study uses Rasch models, whereas TIMSS and PIRLS use 3PL and 2PL models (with a generalized partial credit model). Furthermore, scaling of TIMSS was initially done in 1995 based on the Rasch model, but in 1999 the data were rescaled according to 3PL and 2PL models and direct comparisons were never published by the IEA. There is wide discussion about which models are more appropriate for which case; ultimately, however, tradition plays a significant role in these cases, along with the initial choice of models when the study was conducted for the first time.

We used the same data as for previous research questions (PIRLS with 45 countries). In this case, we compared three different conditions. One of the conditions represents the result of the sample size question with ten countries. For the other two conditions, PIRLS data were rescaled using Rasch family models (1PL and partial credit model). The calibration sample included ten countries. In the second condition, the results were compared to the Rasch reference (all countries used in the calibration sample and the items were calibrated using Rasch family models) and in the third condition, the same data were compared to the 3PL 2PL GPCM reference (the same reference also used in previous sections; all countries in calibration sample using 2PL, 3PL and generalized partial credit models).

First, item parameters were compared across conditions. As expected, the Rasch-Rasch condition showed the highest degree of invariance in all item parameter estimates. The invariance of difficulty parameter in 3PL 2PL GPCM -3PL 2PL GPCM condition was lower than in the Rasch-Rasch condition. The effect size exceeded the predetermined critical value and can be regarded as large and the difference was significant. Still all correlations were 0.90 (for $step_3$) or higher. Furthermore, the correlations in the Rasch-3PL 2PL GPCM conditions were high, but the correlation of the difficulty parameter was significantly lower than in both other conditions (also the effect size was important). This is in line with previous research. Macdonald and Paunonen (2002) found the highest correlations of difficulty parameters in 1PL in contrast to 2PL in all test lengths and distributions of true item difficulty values. Moreover, Fan (1998) reported the highest correlation of difficulty parameter in 1PL compared to 2PL and 3PL, although all the correlations were extremely high (above 0.96 in random samples).

When observing the absolute difference in average achievement scores with the reference across conditions, the variation within the 3PL 2PL GPCM -3PL 2PL GPCM condition is the highest in comparison to other two conditions. Although the estimated differences in the Rasch models were highly consistent within conditions (with a particularly low standard deviation of average differences), in the Rasch-3PL 2PL GPCM condition these differences are consistently high especially for lower achieving countries (all the effect sizes exceeded the critical value and are large according to Cohen's criterion). Our finding is consistent with other comparable research done on this topic. Brown et al. (2005) found a change of achievement score distribution in TIMSS 1995 when either a 1PL or 3PL model was used. This did not have an effect on the ranking of countries, which did not change significantly. Similarly, the correlation between achievement scores obtained from different models was high. Nevertheless, they observed greater differences in lower achieving countries. The differences were most probably due to the fact that the 3PL allows for guessing. Furthermore, they report that controlling for guessing yields a poor ability to be better revealed. This leads to a decreased mean achievement, especially in the bottom of the achievement distribution (when a 3PL is used in contrast to 1PL). Moreover, Leeson and Fletcher (2003) reported that the guessing parameter is needed to take into account performance at the low end of the ability continuum (where guessing appears to be a factor in test performance).

The variability of average achievement scores when 3PL was used was much higher compared to the scores obtained by the 1PL model, especially for countries with lower achievement. Brown et al. (2005) reported that the dispersion of scores is far from being robust to the choice of the model. When 3PL was used, the coefficient of variation doubled for some countries compared to the 1PL model.

In practical situations, it is usually of an advantage to use models with fewer rather than more parameters. The interpretability of parameter estimates in 1PL is easier and also the unique mathematical properties speak in favor of this model. Mazzeo and von Davier (2008) recognize that there is some justification for a decision based on an assessment for a more general IRT model (2PL/generalized partial credit model, and 3PL), especially in the light of an assessment that is designed to provide a broad coverage of the domain using multiple item formats and test versions. They stated further that more complex IRT models do accommodate the functioning of items in diverse populations better than the Rasch model, which assumes that all items contribute the same amount of information to the

measurement of student achievements. In the opinion of Mazzeo and von Davier (2008), using a more general IRT model can also help to reduce some of the country-by-item interactions observed in PISA, since the adoption of a more complex measurement model improves model-data-fit considerably. Moreover, this is another critical issue in the choice of models, i.e. model fit.

Progar and Sočan (2008) investigated the IRT model fit in TIMSS 1995 for a subsample of mathematics and science items. They determined that the 2PL model, in comparison to the 1 PL model, fits the data significantly better for all tests and that the 3PL model fits the data better than the 2PL for two (out of three mathematics and three science) tests. Furthermore, they report that the assumption of unidimensionality holds to a reasonable extent in the subsample of math items but is violated in science items.

In the opinion of Klieme and Baumert (2001), it is a robust finding that unidimensional IRT models never show a perfect fit in large samples. The misfit is generally regarded as a negligible specification error of error variance. In addition, Klieme and Baumert (2001) state that complex proficiency syndromes (as they refer to the constructs measured in LSA) surprisingly fit the unidimensional IRT models almost as well as multidimensional models.

In general, the MRSD values for the Rasch - Rasch condition were consistently low in values and variability in all investigated cases (percentiles and subgroups). Rasch models seem to provide the most invariant achievement estimates across all achievement ranges. This can also be seen in the Rasch-3PL 2PL GPCM condition. The MRSD based on Rasch models were either consistently low or consistently high for countries (the variability of the estimates was extremely low; the effect sizes were large especially for the 90^{th} and 95^{th} percentile).

To summarize the results of this research question, there are no substantial differences in achievement scores of countries when their average achievement score is at least 400 points. As soon as a country's achievement is smaller, there are bigger differences between models (the association of achievement and MRSD is in more detail described in chapter 5.5). However, in line with the results of other researchers that suggest a better model fit in more complex models and that the assumption of equal discrimination of all items across countries does not seem particularly reasonable if we take into consideration the number and variety of participating countries, we would suggest the use of more complex models rather than the 1PL model. While the apparent invariance of simpler models cannot be

overlooked it is also necessary to recognize the advantage of the 3PL model controlling for guessing, especially when lower achieving countries are participating. It is true that this item parameter was found to be the least invariant; nevertheless, after including a certain number of countries it showed sufficient invariance. It appeared to be the least invariant when higher achievement countries were included in the calibration sample. This finding suggests that in case of only higher achieving countries a simpler model would be more appropriate to use. In conclusion, the major differences in average achievement scores for extreme achievement scores based on different models show that the choice of model is a fundamental issue to consider.

5.4 Different content domains

The fourth research question was dealing with the invariance of parameters across content domains. Since the hierarchical nature of the cognitive items used in reading assessment in PIRLS (several items are related to a common passage), the local item independence assumption could be violated. Quittre and Monseur (2010) raise the concern of violation of this assumption when tests are organized in units around a common stimulus. Two difficult situations can occur: a local dependence resulting from an unusual level of interest or prior knowledge about the stimulus and a local dependence produced by the fact that information used to answer different items in the unit is interrelated in the stimulus.

The failure of achieving conditional independence in reading comprehension is also recognized by Bock and Moustaki (2007). They report that tests of reading comprehension usually consist of text passages containing information that the intended examinees are unlikely to know. The passage is followed by a number of items based on specific facts mentioned in the text. Because an examinee's understanding of the overall meaning of the passage will affect his or her ability to answer any of the items, it is unlikely that items related to the same passage will have the same degree of association as items related to different passages.

In their study, Quittre and Monseur (2010) were investigating local item (in)dependence and found that PIRLS 2006 texts (for 23 of selected participating countries) generate minor context-related local item dependency, which is generally unusually low. In line with previous concerns and findings based on the study of Quittre and Monseur (2010), the

focus of our interest was in the parameter invariance in assessing reading (in PIRLS) and mathematics (in TIMSS). It was of interest to investigate if the minor context-related local item dependency in PIRLS 2006 also shows an effect in estimated parameters compared to TIMSS 2007 (where item dependencies were not expected).

For this purpose, only common countries participating in PIRLS 2006 and TIMSS 2007 were selected. To observe parameter invariance, ten countries were repeatedly chosen at random for each study and their parameters compared to the reference condition (when all countries were included in the calibration sample). The invariance was observed within the PIRLS study and within the TIMSS study separately, and then the results between studies were compared (this is also the reason only countries that participated in both studies were selected for analyses).

The item parameter correlations for reading were remarkably similar to those obtained in the ten-country condition when observing different numbers of countries in the calibration sample. The only difference between these two conditions was that here countries were sampled from a total number of 29 countries (and their result was compared to the total of 29 countries), and in the different number country condition the countries were sampled from all 45 countries (and also compared to the reference based on the inclusion of 45 countries). The only significant difference between the content domains between item parameters was in the slope parameter, where higher correlations were found in reading assessment. The effect size also exceeded the critical value and was small, according to Cohen's criterion.

Furthermore, the differences in achievement scores were compared between the content domains. From Figure 4.20, it can be seen that the same pattern of MRSDs can be found in both content domains. Countries with lower average achievement show greater MRSD values and also greater variability of MRSD. In general, the differences were found to be significant. The domain of mathematics shoved almost the double MRSD as found in reading. The differences were extremely small and (in mathematics they were on average one score point on a scale with mean 500 and standard deviation of 100) the effect size did not exceed the critical values. The MRSD values did not significantly differ for the 5^{th}, 10^{th}, 90^{th}, 95^{th}, in the category of boys or in the first two categories of the number of books (missing and 0-10 books); the effect sizes were also trivial or small (based on Cohen's criterion). In the 50^{th} percentile, for girls and all remaining categories of the number of

books, the MRSD values in reading were significantly smaller than in mathematics (all these effect sizes were exceeding the value that was set as a critical value before the analyses) and can be categorized as large effect sizes.

It is extremely interesting, that in contrast to our expectations, the smaller invariance was observed in the domain of mathematics. Because of the nature of the cognitive tests, we were expecting smaller invariance in the domain of reading, but the same finding is supported by Monseur et al. (2011). They found greater local item dependencies in the content domain of mathematics compared to reading in PISA. One possible explanation of our result also could be that in TIMSS more items were used in scaling (177 and in PIRLS 125), which itself could also mean a more precise estimate of ability. However, upon examination of the item type, in TIMSS there were 94 multiple choice items and in PIRLS only 63. Since the guessing parameters turned out to be the least invariant across all investigated conditions, the reason of the smaller invariance of achievement scores in mathematics could be based on inclusion of many multiple choice items. Despite the concern of the violation of the local independence assumption in PIRLS, the achievement scores were found to be more invariant in the domain of reading in comparison to the domain of mathematics. Since the guessing parameter showed to be the least invariant, perhaps for future studies in the field of mathematics, fewer multiple choice items should be included in the cognitive part of the assessments to gain more invariant results.

5.5 Association of achievement and MRSD and standard deviation of MRSD

For all conditions, the association of achievement with MRSD and standard deviation of MRSD were investigated. Embretson and Reise (2000) stated that the accuracy of estimating two different trait levels from test data differs between item sets. If an item set is easy, a low trait level will be more accurately estimated than a high trait level. Similarly, if the calibration sample has relatively low trait levels, the difficulty of easy items will be more accurately estimated than hard items. In our case, the items were always the same, but the calibration sample of examinees differed. Since the calibration sample has an effect in estimating the accuracy of item parameters, we further assumed that it can also have an effect on the achievement scores and their variation.

Firstly, the magnitude of the absolute difference (average MRSD) in every condition was observed. The most variant conditions were those with high achieving countries in the calibration sample and when comparing Rasch to the 3PL 2PL GPCM models. Average MRSDs in these two conditions were consistently larger (at least twice as large as in other conditions) for every observed statistic compared to all other conditions, with the biggest differences in extreme percentile values (especially for the 5^{th} and 10^{th} percentile). The highest variability of the MRSD values was observed in the smaller sample sizes (when a few countries were included in the item parameter estimations). These results show that in some conditions we would consistently obtain different achievement scores of countries and that the calibration sample does have an effect on estimated parameters.

A quick review of all presented figures in the results section gives the impression that an association between achievement and MRSD is present. However, the association does not seem to be linear in all cases. For this reason, the association was observed with the help of linear and quadratic regression models. The comparison of models revealed that in most cases the quadratic model fitted better. The association of MRSD values and achievement was better described by a quadratic function for average country achievement, 5^{th}, 10^{th}, 90^{th} and 95^{th} percentiles in almost all conditions. There were two exceptions. In the reading domain, the association seemed to be linear for average country achievement and the MRSD values for 90^{th} and 95^{th} percentiles when low achieving countries were included in the calibration sample did not seem to be related with countries' achievement scores. The association of MRSD values with achievement scores was strong, but the strongest association was found in lower parts of countries distributions. In addition, the quadratic shape of the curve was most prominent in the 90^{th} and 95^{th} percentiles. Very similar results were observed for the association of standard deviation of MRSD and achievement.

From Table A.1 in Appendix A, a change in the direction of association can be observed. When lower distribution points and mean values are observed the association is negative, but for the 90^{th} and 95^{th} percentiles it changes to positive. These results are in line with our expectations. Extreme values are estimated with smaller precision. This is evident also in our results. The quadratic relationship shows that both smaller and larger values within an observed statistic showed bigger absolute differences with the reference scores in all conditions. The only statistic that showed no quadratic association in most conditions was the average country achievement.

From the observed results, we can conclude that achievement scores and MRSD values are associated. Specific caution is advised when a country mean differs a lot from the mean scale score. Usually the average MRSD value is below five score points in most conditions and observed statistics. However, there are two conditions that in general give different results in comparison to the reference scores. A large difference was observed when higher achieving countries were included in the calibration sample and when Rasch family models were compared to more complex models. These results confirm that the calibration sample does have an effect in parameter estimates and that the choice of the model used for estimating parameters plays an important role in scaling.

5.6 Limitations of the present study and suggestions for future work

As there is practical value in including countries as units because the participants in international LSA studies are countries, it might also be useful to include separate students or uncluster the students from the current countries and group them differently into countries based on stricter or more controlled conditions. In our case, we were limited to the characteristics of the participated countries so the results based on this study might not have sufficient generalizability potential. This study should be repeated on different countries or more countries, or replicated on a different cycle of the study. The next cycle of TIMSS and PIRLS was conducted in 2011, and there is an overlap between these two studies. The same 4^{th} grade students participated in the reading as well as in the mathematics and science assessment. Unfortunately, the data were published in December 2012 and were not available when this study was initiated. However, replication of the study on the new available data could enhance the findings and also widen the generalizability of the results since it covers a broader range of countries.

Another limitation of this empirical study is that the parameter values cannot be manipulated, and the true values are not known. Not all possible combinations of conditions were investigated, and we omitted the content domain of science. As “true” scores, we used the scores obtained by the full sample and called them “reference scores”. In contrast, in simulation studies with data generated by some specific rules, we cannot foresee all possible situations that could be potentially meaningful in empirical studies

either. Furthermore, todays' understanding of processes of acquiring knowledge is limited and in this case modeling the "real" situations can also be misleading and not sufficient.

An alternative to the simulation study presents the use of analytical approach, but the analytical approach requires a solid theoretical basis and usually involves more assumptions. It is not suitable for too complex problems. Due to the fact that IRT models are extraordinarily complex, it is more difficult to obtain results with the use of analytical techniques only.

One of the limitations of this study is also that no absolute statement about the quality of parameter estimates could be made in the comparisons of models. This is because we do not know the true parameter estimates to evaluate the bias in using Rasch family models or the 3PL, 2PL, GPCM alternative.

Moreover, we did not separately investigate the model fit, since we assumed all model fits are sufficient for having valid estimates (based on the technical reports and previous studies), but in this case we were not able to compare the variability of the estimates with the degree of model fit. This could also present valuable information that could be included in further studies. Currently, the evaluation of model fit is still in development. In case of including model fit, extensive research should be conducted to determine the most suitable measure of fit, which would probably need a separate investigation.

Another limitation of the procedures used for examining invariance was that imputation error was not incorporated in the results. In LSA studies, the standard error usually also includes an imputation error (that represents the variability in using multiple individual achievement scores). Consequently, the confidence intervals of scores are probably underestimated. In general, this should not affect the interpretation of the results since the imputation error should be similar across conditions.

In this study, the correlation coefficient was used as a measure of invariance of item parameters. According to Rupp and Zumbo (2004) a correlation coefficient is not a sufficient indicator of parameter invariance since it can miss additive group level effects. Instead, their suggestion is to use other measures, such as examining differential item functioning. In our opinion, differential item analysis in LSA can be extremely time consuming. In the context of LSA studies, it is not practical or nearly impossible to search for and remove items exhibiting differential item functioning in all possible subpopulations

that may be of interest to researchers. Another way to investigate invariance could be through assessing the model fit. Thomas and Cyr (2002) report that the NAEP analysts did not assume that the IRT model used was correct, but only that it was realistic enough to provide a good overall summary of the main features of interest. In that sense, there is no perfect measure of invariance, and if some of the assumptions are violated it is still possible for the invariance to hold. The most important is to be aware of the trade-off between the complexity of the analyses and the resulting practical use in empirical situations.

Although the results presented in the thesis were mostly in line with previous research findings, there is still a need to investigate invariance in LSAs in more detail. The invariance should also be investigated under different characteristics of the calibration sample: for example, by the range of abilities that is covered in the calibration sample or by the variation in the achievement in countries. Another question could be how the variability within a country is correlated to the overall achievement estimates. We should certainly also study the effect of the calibration sample on subgroup estimates in greater detail.

In addition, we could attempt to evaluate the effect of every single country in the estimates. That would provide the information if item parameters show differences across countries. In practice, the use of one country does not give sufficient information because it does not provide enough data to reliably estimate all the item parameters (the number of items and missing responses are too large). Even when including two countries in the calibration sample, we frequently experienced no solutions in the parameter estimation phase.

An interesting topic for further research would also be the inclusion and manipulation in included items observed in LSA studies. Items could be manipulated by difficulty, discrimination, number, content or cognitive property. Special caution should be then focused on the contextual effects or the assumption of absence of contextual effects should be fulfilled. In this case, we would obtain an even more detailed picture of invariance including that concerned with items.

6 Conclusions and original contribution to development of the scientific field

The calibration sample in LSA studies usually involves a subsample of participating countries. To obtain trends in achievement, the calibration sample is represented only by countries that participated in subsequent cycles of a particular study. Furthermore, the obtained item parameter estimates are also used to determine achievement scores of other participating countries (that were not included in the calibration sample). Since participating countries in IEA studies rely on the study's results, having some empirical evidence of the invariance of item parameters and achievement score estimates is of immense importance. The focus of the thesis was to observe parameter invariance in LSA under different conditions.

Parameter invariance is an ideal state and is violated if any of the item parameter estimates fail to be identical (up to the same linear transformation) across different examinee populations or measurement conditions. Our interest was in the effect of the calibration sample (size and ability) on parameter estimates. Furthermore, the use of different IRT models was observed, and the invariance was investigated in two content domains. This thesis contributes to a better understanding of the property of the invariance of IRT model estimates in real data.

The investigated conditions are of particular importance because only a small number of countries participate in some LSA studies (e.g. TIMSS Advanced, prePIRLS). Furthermore, in other studies (TIMSS and PIRLS) a wide variety of countries are participating and more and more developing countries, which achievement is usually lower than that from the developed countries join each cycle. The IRT models used in TIMSS and PIRLS differ from the ones used in PISA, and no clearly documented findings that would give sufficient information about comparisons of models exist thus far. Finally, the invariance across content domains is important because of possible violations of the IRT model assumptions in estimating reading achievement.

We observed parameter invariance in item parameters with correlation coefficient and for estimating the invariance in achievement scores we calculated mean root squared distance

of the new scores in comparison to the reference ones. Even though we could not observe perfect parameter invariance, we could observe high correlations among item parameter estimates across different conditions. In general, the correlations with the reference item parameter estimates were very high across all investigated conditions. The most invariant item parameter estimate turned out to be item difficulty, and the least invariant item parameter estimate was the guessing parameter. These findings were in line with the previous research.

From a countries' perspective, we found small and sometimes insignificant differences in the achievement scores for the majority of countries when different countries were included in the calibration sample. In general, extreme values of the distribution were estimated with more difference to the reference scores compared to the mean scores. The biggest differences were observed when higher achieving countries were included in the calibration sample and when Rasch models were compared to the 2PL, 3PL and GPCM models.

Although the mean achievement scores of countries in the middle range of achievement scale were in general remarkably similar to the reference achievement scores, countries with lower achievement showed greater variability in all conditions, especially when fewer countries were included in the item parameter estimations. In addition to the fact that there is not enough information to estimate lower achievement (due to the high rate of incorrect responses), sampling design also plays a role. In contrast, with an efficient sampling design we can, to some extent, overcome the greater variability of lower countries' achievement scores (as could be seen from the example of one country in the condition of different number of countries included). Furthermore, the achievements of lower achieving countries showed less variability and smaller differences in comparison with the reference when more countries were included in the calibration sample. The differences with the reference were also smaller when lower achieving countries were included in the calibration sample. A greater invariance was also observed with the use of simpler models.

The association of MRSD values with achievement scores was found to be the greatest for lower achieving countries and also greater for high achieving countries compared to the countries with average mean achievement score, whereas the standard deviation of MRSD values was the greatest for small sample sizes, it also showed that differences in other conditions were more or less stable. Achievement score invariance across content domains

and in random samples (with varying sample sizes) was the smallest compared to different ability samples and use of different models. The higher achieving calibration sample gave the least invariant achievement scores and also the scores when comparing Rasch models to 2PL, 3PL and GPCM models were less invariant. From these results, we can conclude that the calibration sample plays a role in LSA studies.

It is possible that we missed some important characteristics that could potentially have an effect on the results. In general, we identified a few conditions where the achievement scores turned out to be less invariant as expected. Even if small absolute differences existed, the effect on the achievement scores in relative sense could be large. Nevertheless, the results of lower achieving countries should be interpreted with immense caution since their results might be subject to greater variability (due to sampling design, less accurate estimates due to the high number of incorrect items, and especially when the calibration sample is small).

Since not many studies on this topic exist, it was exceedingly difficult to determine a value that could be interpreted as sufficient effect size, which is why a sensitivity power analysis was used. According to this analysis, we determined the critical effect sizes with power of 0.8 and significance criterion of 0.05 (for a given numerus). To get a sense of the magnitudes of the determined effect sizes, we used Cohen's classification. When comparing the critical effect sizes with Cohen's classification, we can observe that most of the critical values would be regarded as medium in effect size. Most of the insignificant results showed trivial or small effect sizes (according to Cohen's criterion). Based on both criteria we can conclude, if we did not obtain a significant difference, that it is reasonable to conclude that the effect is nonexistent from the practical viewpoint. Furthermore, we could also conclude that the study had a sufficient number of repetitions, and the results can be interpreted with confidence.

Practical guidelines for future studies resulting from the findings would suggest including at least ten countries with broad coverage of abilities in calibration sample. When lower achieving countries participate in a study, it is especially important that these countries are also included in the calibration sample, if possible. Rasch family models were found to be more invariant and show remarkably similar MRSD values for all ranges of countries' achievements. In case of the use of more complex models, the achievements of lower achieving countries are more variable than the rest. The differences between the simpler

and more complex models are in the extreme values. This finding shows that the choice of the model is an important issue to consider. The invariance in reading was found to be greater than in mathematics, which could be due to the number of items included and the item type used. Greater invariance in mathematics could be achieved, if fewer multiple choice items would be included (since the guessing parameter turned out to be the least invariant). Moreover, the calibration sample should consist of as many countries as possible to avoid the non-convergence of item parameter estimates that we experienced during the item calibration phase in case of fewer countries.

Thus far, many research papers have focused on the comparison of IRT and classical test theory estimates, IRT estimates in general, item position effects, the linking error and its relation to the number of linking or trend items included in trend estimation. A few studies have provided empirical evidence on invariance in real data. We examined the invariance in two LSA studies, with the focus on the calibration sample, calibration model and content domain. In addition to the theoretical assumption of invariance at the base of the model fit (that is usually checked and shows sufficient fit), empirical evidence of invariance was provided. Simulation studies can rarely capture the wide variety of conditions that occur in reality. With this information, we were able to evaluate if there are any differences in achievement score estimates and item parameter estimates due to the inclusion of different countries in subsequent cycles on which trend estimates are based. This is of practical importance since it presents guidelines for future LSA studies and also other studies that use IRT models. The original contribution of the dissertation results in a different perspective to parameter invariance in a special case of studies.

The results of this study not only enhance the understanding of invariance property of IRT models, but also enable LSA professionals to raise the efficiency of their item calibration process in case a subset of participating countries turned out to be sufficient for a valid estimation of item parameters. It also provides practical guidelines and points out some situations that should be avoided in item calibration process.

In simulation studies, the clustered nature of the data should definitely be taken into account since many studies (due to costs) use a two-stage sample design. The advantage of this study is the use of real data that are clustered. Furthermore, in future studies the invariance should be investigated in clustered samples and not only in random or ability samples with varying sample sizes.

From the obtained results, we can conclude that the calibration sample does have an effect on the achievement score estimates in some conditions. These findings increase and complement the empirical evidence (knowledge) of parameter invariance in real data. In particular, the findings are useful for international LSA studies and other studies using IRT models.

7 References

Adedoyin, Ombola O. 2010. Investigating the Invariance of Person Parameter Estimates Based on Classical Test and Item Response Theories. *International Journal of Educational Science* 2 (2): 107–113.

Adedoyin, Ombola O., Johnson H. Nenty and Bagele Chilisa. 2008. Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Review* 3 (2): 83–93.

Babcock, Ben and Anthony D. Albano. 2012. Rasch Scale Stability in the Presence of Item Parameter and Trait Drift. *Applied Psychological Measurement* 36 (7): 565–580.

Baker, David P. 1997. Surviving TIMSS: Or, Everything You Blissfully Forgot about International Comparisons. *The Phi Delta Kappan* 79 (4): 295–300.

Baker, Frank B. 2001. *The basics of item response theory*. University of Maryland, College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Beaton, Albert E. and E. G. Johnson. 1992. Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement* 26 (2): 163–175.

Beaton, Albert E. and David F. Robitaille. 2002. A look back at TIMSS: What have we learned about international studies? In *Secondary Analysis of the TIMSS Data*, ed. Albert E. Beaton and David F. Robitaille, 409–417. Dordrecht: Kluwer Academic Publishers.

Bela knjiga o vzgoji in izobraževanju v Republiki Sloveniji [White paper on education in the Republic of Slovenia] 2011. Ljubljana: Zavod RS za šolstvo.

Bock, R. Darrell and Irini Moustaki. 2007. Item response theory in a general framework. In *Handbook of statistics on psychometrics, Vol 26, Psychometrics,* ed. C. Radhakrishna Rao and Sandip Sinharay, 469–513. Amsterdam: North-Holland.

Bond, Trevor G. and Christine M. Fox. 2001. *Applying The Rasch Model: Fundamental Measurement in the Human Sciences*. USA: Lawrence Erlbaum Associates.

Braun, Henry. 2013. Prospects for the future: A Framework and Discussion of Directions for the Next Generation of International Large-Scale Assessments. In *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*, ed. Matthias von Davier, Irwin Kirsch, Kentaro Yamamoto and Eugenio Gonzalez, 149–160. Dordrecht: Springer.

Brennan, Robert L. 2008. A discussion of population invariance. *Applied psychological measurement* 32 (1): 102–114.

Brown, Giorgina, John Micklewright, Sylke V. Schnepf, and Robert Waldmann. 2005. Cross-National Surveys of Learning Achievement: How Robust are the Findings? *IZA Discussion Paper No.* 1652.

Bulle, Nathalie. 2011. Comparing OECD educational models through the prism of PISA. *Comparative Education* 47 (4): 503–521.

Cohen, Jacob. 1992. Statistical power analysis. *Current directions in psychological science* 1 (3): 98–101.

Cook, Linda L., Daniel R. Eignor, and Hessy L. Taft. 1988. A Comparative Study of the Effects of Recency of Instruction on the Stability of IRT and Conventional Item Parameter Estimates. *Journal of Educational Measurement* 25(1): 31–45.

Crocker, Linda and James Algina. 1986. *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

De Ayala, Rafael Jaime. 2009. *The Theory and Practice of Item Response Theory (Methodology in the Social Sciences)*. New York: The Guilford Press.

Direct Estimation Software Interactive (Version 3.23) [computer software and manual] (2009). Princeton: Educational Testing Service.

Du Toit, Mathilda. 2003. *IRT from SSI: BILOG_MG, MULTILOG, PARSCALE, TESTFAC*. Lincolnwood, IL: Scientific Software International.

Embretson, Susan E. and Steven P. Reise. 2000. *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, George. 1994. Historical views of the concept of invariance in measurement theory. In *Objective measurement: Theory into practice (Vol. 2),* ed. Mark Wilson, 73–99. Norwood, NJ: Ablex.

Fan, Xitao. 1998. Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement* 58 (3): 357–381.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39: 175–191.

Foy, Pierre and Alka Arora. 2009. TIMSS Advanced 2008 User Guide for the International Database. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College

Foy, Pierre and John F. Olson, eds. 2009. *TIMSS 2007 International Database and User Guide.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Foy, Pierre, Bradley Brossman, and Joseph Galia. 2012. Scaling the TIMSS and PIRLS 2011 Achievement Data. In *Methods and procedures in TIMSS and PIRLS 2011,* eds., Michael O. Martin and Ina V. Mullis, 1–28. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Foy, Pierre and Kathleen T. Drucker, eds. 2013. *PIRLS 2011 User Guide for the International Database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Foy, Pierre and Ann M. Kennedy, eds. 2008. *PIRLS 2006 User Guide for the International Database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Foy, Pierre, Michael O. Martin and Ina V. Mullis. 2010. The Limits of Measurement: Problems in Estimating Reading Achievement in PIRLS 2006 for Low-performing Countries. Presented on IRC 2010 Gothenburg Sweden. Available at:

http://www.iea-irc.org/fileadmin/IRC_2010_papers/PIRLS/Foy_Martin_Mullis.zip (17th of November, 2010).

Galdin, Marlene and Louis Laurencelle. 2010. Assessing parameter invariance in item response theory's logistic two item parameter model: A Monte Carlo investigation. *Tutorials in Quantitative Methods for Psychology* 6 (2): 39–51.

Goldstein, Harvey. 2004. International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education* 3: 319–330.

Gonzalez, Eugenio. J. 1997. Reporting Student Achievement in Mathematics. In *Third International Mathematics and Science Study Technical Report, Volume II: Implementation and Analysis – Primary and Middle School Years*, ed. Michael O. Martin and Dana. L. Kelly, 147–174. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Gray, John. 1997. A Bit of a Curate's Egg? Three Decades of Official Thinking about the Quality of Schools. *British Journal of Educational Studies* 45 (1): 4–21.

Hambleton, Ronald K. and Hariharan Swaminathan. 1985. *Item Response Theory: Principles and Applications*. Boston MA: Kluwer-Nijhoff Publishing.

Hambleton, Ronald K., Hariharan Swaminathan, and Jane H. Rogers. 1991. *Fundamentals of Item Response Theory.* Newbury Park, Calif.: Sage Publications.

Hencke, Juliane, Leslie Rutkowski, Oliver Neuschmidt, and Eugenio J. Gonzalez. 2009. Curriculum coverage and scale correlation on TIMSS 2003. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* 2: 85–112.

Hickendorff, Marian, Willem J. Heiser, Cornelis M. van Putten, and Norman D. Verhelst. 2009. How to Measure and Explain Achievement Change in Large-Scale Assessments: A Rejoinder. *Psychometrika* 74 (2): 367–374.

Holland, Paul W., Neil J. Dorans and Nancy S. Petersen. 2007. Equating Test Scores. In *Handbook of statistics on psychometrics, Vol 26, Psychometrics,* ed. C. Radhakrishna Rao and Sandip Sinharay, 169–203. Amsterdam: North-Holland.

IBM Corp. 2011. *IBM SPSS Statistics for Windows, Version 20.0.* Armonk, NY: IBM Corp.

IEA. 2013. Current studies of International Association for the Evaluation of Educational Achievement. Available at: http://www.iea.nl/current_studies.html (30th of June, 2013).

Keeves, John P. 2011. IEA – From the Beginning in 1958 to 1990. In *IEA 1958-2008: 50 years of experiences and memories, volume 1*, ed. Constantinos Papanastasiou, Tjeerd Plomp and Elena C. Papanastasiou, 3–39. Nicosia, Cyprus: Cultural Center of the Kykkos Monastery.

Kijima Rie. 2010. Why participate? Cross-national Assessments and foreign aid to education. In *The Impact of International Achievement Studies on National Education Policymaking*, ed. Alexander W. Wieseman, 35–61. Bingley: Emerald.

Kirsch, Irwin, Marylou Lennon, Matthias von Davier, Eugenio Gonzalez, Kentaro Yamamoto. 2013. On the Growing Importance of International Large-Scale Assessments. In *The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*, ed. Matthias von Davier, Irwin Kirsch, Kentaro Yamamoto and Eugenio Gonzalez, 1–12. Dordrecht: Springer.

Klemenčič, Eva. 2010. The Impact of International Achievement Studies on National Education Policymaking: The Case of Slovenia – How many Watches do we need? In *The Impact of International Achievement Studies on National Education Policymaking*, ed. Alexander W. Wieseman, 239–266. Bingley: Emerald.

Klemenčič, Eva and Mojca Rožman. 2009. Knowledge globalization through international studies and assessments. *International conference on social sciences and humanities: The progressive impact of research in social sciences and humanities: towards the regeneration of knowledge*, 54–69. Malaysia: Faculty of Social Sciences and Humanities.

Klieme, Eckhard. 2000. Fachleistungen in voruniversitaren mathematic und physikunterricht: teoretische grundlagen, kompetenzstufen und unterrichtschwerpunkte. In *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie - Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Vol. 2. Mathematische und physikalische Kompetenzen*

am Ende der gymnasialen Oberstufe, ed. Jurgen Baumert, Wilfred Bos and Reiner Lehmann, 57–128. Opladen: Leske und Budrich.

Klieme, Eckhard and Jurgen Baumert. 2001. Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education* 16 (3): 385–402.

Kolen, Michael J. and Robert L. Brennan. 2004. *Test equating, scaling, and linking: methods and practices*. New York: Springer.

Kubow, Particia K. and Paul R. Fossum. 2007. Comparative Education. In *Comparative Education: Exploring Issues in International context*, 3–29. Upper Saddle River; Columbus: Pearson/Merrill/Prentice Hall.

Leeson, Heidi and Richard Fletcher. 2003. An Investigation of Fit: Comparison of the 1-, 2-, 3-Parameter IRT Models to the Project asTTle Data. *NZARE/AARE Conference 2003 Proceedings*. (unpaged). Auckland, NZ: New Zealand Association for Research in Education and Australian Association for Research in Education Conference 2003.

Linn, R. L. 2002. The Measurement of Student Achievement in International Studies. In *Methodological advances in cross-national surveys of educational achievement*, ed. Andrew C. Porter, and Adam Gamoran, 27–57. Washington, DC: National Academy Press.

Lord, Frederic M. 1980. Omitted responses and formula scoring. In *Applications of item response theory to practical testing problems,* 225–234. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, Frederic M. and Melvin R. Novick. 1968. *Statistical theories of mental test scores*. New York: Addison-Wesley.

Macdonald, Paul and Sampo V. Paunonen. 2002. A Monte Carlo Comparison of Item and Person Statistics Based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement* 62 (6): 921–943.

Mair, Patrick, Steven P. Reise and Peter M. Bentler. 2008. IRT Goodness-of-Fit Using Approaches from Logistic Regression. Department of Statistics Papers, Department

of Statistics, UCLA, UC Los Angeles. Available at: http://www.escholarship.org/uc/item/1m46j62q (2nd of April, 2013).

Martin, Michael O., ed. 2005. *TIMSS 2003 User Guide for the International Database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, Michael O., Mullis, Ina V.S. and Kennedy, Ann M. 2007. *PIRLS 2006 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mazzeo, John and Matthias von Davier. 2008. *Review of the Programme for International Student Assessment (PISA) Test Design: Recommendations for Fostering Stability in Assessment Results.* OECD Publishin. Available at: http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=EDU/PISA/GB%282008%2928&docLanguage=En (17th of November, 2010).

Michaelides, Michalis P. and Edvard H. Haertel. 2004. *Sampling of common items: An unrecognized source of error in test linking. CSE Report 636.* Los Angeles: Center for the Study of Evaluation (CSE), University of California.

Mislevy, Robert J. 1991. Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56: 177–196.

--- 1995. What can we learn from international assessment? *Evaluation and Policy Analysis* 17 (4): 419–437.

Mislevy, Robert J., Eugene G. Johnson and Eiji Muraki. 1992. Scaling procedures in NAEP. *Journal of Educational Statistics* 17 (2): 131–154.

Monseur, Christian, Ariane Baye, Dominique Lafontaine, and Valérie Quittre. 2011. PISA test format assessment and the local independence assumption. *IERI Monograph Series: Issues and Methodologies in Large- Scale Assessments* 4: 131–155.

Monseur, Christian and Alla Brezner. 2007. The Computation of Equating Errors in International Surveys in Education. *Journal of applied measurement* 8 (3): 323–335.

Monseur, Christian, Heiko Sibberns and Dirk Hastedt. 2008. Linking errors in trend estimation for international surveys in education. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* 1: 113–122.

Morizot, Julien, Andrew T. Ainsworth, and Steven P. Reise. 2007. Toward modern psychometrics: Application od Item response theory in personality research. In *Handbook of Research Methods in Personality Psychology,* eds. Richard W. Robins, Chris Fraley and Robert F. Krueger, 407–423. New York: The Guilford Press.

Mullis, Ina V. S., Michael O. Martin and Pierre Foy. 2008. *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, Ina V. S., Michael O. Martin, Pierre Foy, and Alka Arora. 2012. *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, Ina V. S., Michael O. Martin, Ann M. Kennedy and Pierre Foy. 2007. *IEA's Progress in International Reading Literacy Study in Primary School in 40 Countries*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, Ina. V., Michael O. Martin, Graham J. Ruddock, Christine Y. O'Sullivan and Corinna Preuschoff. 2009. *TIMSS 2011 Assessment Frameworks*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

OECD. 2005. *PISA 2003 Technical Report OECD*. OECD Publishing.

--- 2009. *PISA 2006 Technical Report*. OECD Publishing.

Olson, John F., Michael O. Martin and Ina V. S. Mullis. 2008. *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Phillips, David and Michele Schweisfurth. 2007. Comparative education research: Survey outcomes and their uses. In *Comparative and International Education: An Introduction to Theory, Method, and Practice,* 118–129. London: Continuum.

Porter, Andrew, C. and Adam Gamoran. 2002. Progress and Challenges for Large-Scale Studies. In *Methodological advances in cross-national surveys of educational*

achievement, ed. Andrew C. Porter, and Adam Gamoran, 3–23. Washington, D.C: National Academic Press.

Postlethwaite, Neville. 1967. *School Organization and Student Achievement. A study based on achievement in mathematics in twelve countries*. Stockholm: Almqvist & Wiksell.

Progar, Špela, and Gregor Sočan. 2008. An empirical comparison of item response theory and classical test theory. *Horizons of Psychology* 17 (3): 5–24.

Puklek Levpušček, Melita, Maja Zupančič and Gregor Sočan. 2012. Predicting Achievement in Mathematics in Adolescent Students: The Role of Individual and Social Factors. *The Journal of Early Adolescence* 33 (4): 523–551.

Quittre, Valérie and Monseur Christian. 2010. Exploring Local Item Dependency for items clustered around common reading passage in PIRLS data. Presented on *IRC 2010 Gothenburg Sweden*.

R Development Core Team 2010. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rudner, Lawrence M. 1977. A Closer Look at Latent Trait Parameter Invariance. Paper presented at the *Annual Meeting of the New England Educational Research Organization*.

Rupp, André A. and Bruno D. Zumbo. 2003. Which Model is Best? Robustness Properties to Justify Model Choice Among Unidimensional IRT Models under Item Parameter Drift. *The Alberta journal of Educational Research* 49 (3): 264–276.

--- 2004. A Note on How to Quantify and Report Whether Irt Parameter Invariance Holds: When Pearson Correlations are Not Enough. *Educational and Psychological Measurement* 65 (4): 588–599.

--- 2006. Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement* 66 (1): 63–84.

Rutkowki, Leslie, Eugenio Gonzalez, Marc Joncas and Matthias von Davier. 2010. International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educational Researcher* 39 (2): 142–151.

Scientific Software International, Inc. (2003). *Parscale for Windows (Version 4.1)*. Chicago: Scientific Software International, Inc.

Stevens, Stanley Smith. 1946. On the Theory of Scales of Measurement. *Science* 103 (2684): 677–680.

Stone, Clement A. and Bo Zhang. 2003. Assessing Goodness of Fit of Item Response Theory Models: A Comparison of Traditional and Alternative Procedures. *Journal of Educational Measurement* 40 (4): 331–352.

Sykes, Robert C. and Anne R. Fitzpatrick. 1992. The stability of IRT b values. *Journal of educational measurement* 29 (3): 201–211.

Štraus, Mojca, Eva Klemenčič, Barbara Brečko, Mojca Čuček, and Alenka Gril. 2006. *Metodološka priprava mednarodno primerljivih kazalnikov spremljanja razvoja vzgoje in izobraževanja v Sloveniji. Raziskovalno poročilo.* Ljubljana: Pedagoški inštitut.

Štraus, Mojca. 2004. Mednarodne primerjave kot podlaga za oblikovanje strategije razvoja izobraževalnega sistema. *Sodobna pedagogika* 55 (5): 12–27.

Taylor, Catherine S. and Yoonsun Lee. 2010. Stability of Rasch Scales Over Time. *Applied Measurement in Education* 23 (1): 87–113.

Thomas, Roland D. and Andre Cyr. 2002. Applying item response theory methods to complex survey data. *SSC Annual Meeting, May 2002, Proceedings of the Survey Methods Section*.

Van der Linden, Wim J. and Ronald K. Hambleton. 1997. *Handbook of modern item response theory*. New York. Springer-Verlag.

Von Davier, Matthias, Sandip Sinharay, Andreas Oranje, and Albert Beaton. 2007. The statistical procedures used in national assessment of educational progress: recent developments and future directions. In *Handbook of statistics on psychometrics,*

Vol 26, Psychometrics, ed. C. Radhakrishna Rao and Sandip Sinharay, 1039–1056. Amsterdam: North-Holland.

Von Davier, Matthias, Eugenio J. Gonzalez, and Robert Mislevy. 2009. What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large- Scale Assessments* 2, 9–36.

Wagemaker, Hans. 2011. IEA: International studies, Impact and Transition. In *IEA 1958-2008: 50 years of experiences and memories*, ed. Constantinos Papanastasiou, Tjeerd Plomp and Eelena C. Papanastasiou, 253–272. Nicosia, Cyprus: Cultural Center of the Kykkos Monastery.

Wells, Craig S., Michael J. Subkoviak, and Ronald C. Serlin. 2002. The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement* 26: 77–87.

Wiseman, Alexander W. 2010. Introduction: The Advantages and Disadvantages of National Education Policymaking Informed by International Achievement Studies. In *The Impact of International Achievement Studies on National Education Policymaking*, ed. Alexander W. Wieseman, xi–xxii. Bingley: Emerald.

Wu, Margaret. 2005. The role of plausible values in large-scale surveys. *Studies in Educational Evaluation* 31: 114–128.

--- 2010. *Comparing the similarities and differences of PISA 2003 and TIMSS. OECD Education working paper no. 32*. OECD Publishing.

8 Author index

9 Subject index

10 Expanded abstract in Slovene

Učinek sestave vzorca pri ocenjevanju parametrov postavk in dosežkov v mednarodnih raziskavah znanja (razširjen povzetek)

Pravica do izobraževanja je zagotovljena v Evropski konvenciji o človekovih pravicah in je omenjena tudi v veliko drugih mednarodnih konvencijah, sprejetih v zadnjih desetletjih. Vse države podpisnice se zavežejo, da bodo zagotovile enake možnosti izobraževanja za vse. V tem kontekstu je zelo pomembna tudi kakovost izobraževanja. Indikatorji kakovosti izobraževanja se skozi čas spreminjajo. Pomembni indikatorji kakovosti izobraževanja so bili včasih na primer število šol in učencev na določeni stopnji izobraževanja, število učiteljev na določeno število učencev in podobno. V zadnjem času dostopnost šolanja ni več najpomembnejši pokazatelj kakovosti šolskega sistema (Gray 1997; Štraus in drugi 2006; Klemenčič in Rožman 2009). Zgolj dostopnost šolanja še ne zagotavlja tudi učinkovitega in kakovostnega izobraževalnega sistema.

Eden izmed pomembnejših dejavnikov, ki zagotavlja kakovost izobraževanja, je evalvacija izobraževanja (Štraus 2004). Pri evalvaciji izobraževanja sta pomembna tako nacionalni kot tudi mednarodni vidik (Mislevy 1995; Bela knjiga 2011). V veliko državah izvajajo nacionalno preverjanje, vendar pa ima tudi mednarodno preverjanje znanja pomembno vlogo pri evalvaciji izobraževalnega sistema. Nacionalno in mednarodno preverjanje znanja sledita različnim ciljem in služita različnemu namenu. Nacionalno preverjanje znanja se običajno izvaja na celotni populaciji učencev (na primer ob zaključku osnovne ali srednje šole) in je pogosto pogoj za vstop na naslednjo stopnjo šolanja. Mednarodno preverjanje znanja pa posreduje informacije o različnih izobraževalnih sistemih in se izvaja na reprezentativnem vzorcu populacije v sodelujočih državah. Rezultati posameznim državam omogočajo, da bolje spoznajo svoj izobraževalni sistem (Porter in Gamoran 2002).

10.1 Mednarodne raziskave znanja

Trenutno je zaključenih in v teku več mednarodnih raziskav znanja, ki preverjajo znanje učencev na različnih področjih. Raziskavi pod okriljem Mednarodne zveze za evalvacijo izobraževalnih dosežkov (v nadaljevanju IEA) sta na primer Mednarodna raziskava trendov znanja matematike in naravoslovja (v nadaljevanju TIMSS, ki preverja znanje matematike in naravoslovja četrtošolcev in osmošolcev) ter Mednarodna raziskava bralne pismenosti (v nadaljevanju PIRLS, ki preverja bralno pismenost četrtošolcev). Raziskave se razlikujejo glede na vsebino, ki jo raziskujejo in populacijo, ki je vključena v raziskavo. Najbolj razširjena raziskava pod okriljem Organizacije za ekonomsko sodelovanje in razvoj (v nadaljevanju OECD) je Program mednarodne primerjave dosežkov učencev (v nadaljevanju PISA), ki ocenjuje uporabo znanja s področja matematike, naravoslovja in bralne pismenosti v vsakdanjem življenju. IEA raziskave se osredotočajo na preverjanje vsebin, določenih z učnim načrtom, medtem ko se OECD raziskave (na primer PISA) usmerjajo na preverjanje uporabnega znanja, pridobljenega tekom izobraževalnega procesa na populaciji, ki se vključuje na trg dela v državah članicah OECD.

Mednarodne raziskave slonijo na empiričnih podatkih in so zasnovane neeksperimentalno in prečno. Veliko truda vlagajo v zagotavljanje primerljivosti, zanesljivosti in veljavnosti rezultatov. Še posebej je to pomembno pri mednarodnih raziskavah, ki morajo poleg omenjenega zagotavljati tudi primerjalno veljavnost (angl. comparative validity). Za primerjalno veljavnost mora biti izpolnjen še pogoj, da so podatki primerljivi tudi na mednarodni ravni. To pomeni, da je mogoče razlike v dosežkih med državami pripisati dejanskim razlikam med učenci. Goldstein (2004) poroča, da se vzorec odgovorov na postavke v nekaterih državah v raziskavi PISA razlikuje (npr. Anglija in Francija), kar onemogoča podajanje kakršnihkoli zaključkov v zvezi s primerjavami teh držav na podlagi enotne lestvice.

Čeprav so si mednarodne raziskave v specifičnih postopkih različne, pa raziskave IEA in OECD v glavnem uporabljajo podobno metodologijo. Ena od skupnih lastnosti raziskav je, da za ugotavljanje dosežkov uporabljajo teorijo odgovora na postavko (v nadaljevanju TOP).

10.2 Teorija odgovora na postavko

V družboslovnih znanostih merjene spremenljivke običajno niso neposredno dostopne. Latentne spremenljivke (spremenljivke, ki so predmet merjenja, a jih ne moremo neposredno izmeriti) merimo s pomočjo številnih manifestacij vedenja (spremenljivke, za katere predvidevamo, da so odraz latentne spremenljivke). TOP, znana tudi kot teorija latentnih potez, sloni na modelu, kjer je ocena latentne lastnosti odvisna od odgovorov udeležencev in tudi od lastnosti postavk (Embretson in Reise 2000). Model teorije odgovora na postavko vključuje stopnjo lastnosti posameznika in značilnosti postavk, ki so povezane z odgovori udeležencev na postavke. Sloni na dveh predpostavkah: karakteristična krivulja postavke ima določeno obliko in dosežena mora biti lokalna neodvisnost postavk.

Karakteristična krivulja postavke opisuje, kako se spreminja verjetnost odgovora na podlagi spremembe stopnje latentne lastnosti. Predstavlja verjetnost pravilnega odgovora na postavko in je običajno označena kot $P_i(\theta)$, ki je funkcija stopnje lastnosti θ (van Der Linden in Hambleton 1997).

Postavke so lokalno neodvisne pod pogojem, da pri isti vrednosti latentne lastnosti med seboj niso povezane. Zaradi možnega kršenja predpostavke o neodvisnosti postavk so Monseur in drugi (2011) v raziskavi PISA raziskovali lokalno neodvisnost postavk. Ugotovili so, da v primeru odvisnosti precenjujemo variabilnost držav z nižjim dosežkom in podcenjujemo variabilnost držav z višjim dosežkom.

Kot osnova za ocenjevanje stopnje latentne lastnosti posameznika služi matrika odgovorov na postavke (matrika podatkov). Na osnovi odgovorov izračunamo parametre postavk, ki odslikujejo značilnosti oseb in postavk (težavnost, diskriminativnost, popravek za ugibanje …). Nato združimo izračunane lastnosti postavk z odgovori posameznikov ter tako ocenimo stopnjo latentne lastnosti. Latentno lastnost ocenimo na podlagi modela, zato v tem kontekstu o TOP govorimo kot o teoriji, osnovani na modelu.

10.2.1 Modeli TOP

Modeli TOP se razlikujejo glede na matematično obliko karakteristične krivulje postavke in s tem tudi glede na število parametrov, ki jih določa model. Med najpogosteje uporabljenimi so logistični modeli, ti slonijo na logistični funkciji. Na primer Raschev model (ali logistični model z enim parametrom – v nadaljevanju 1PL) transformira surove podatke na lestvico z enakimi intervali. Enakost intervalov je zagotovljena s pomočjo logaritemske transformacije razmerij obetov podatkov, posploševanje pa je mogoče zaradi verjetnostnih enačb (Bond in Fox 2001). Predpostavka Raschevega modela je, da imajo vse postavke enako diskriminativno moč, medtem ko je v logističnem modelu z dvema parametroma (v nadaljevanju 2PL) in v modelu s tremi parametri (v nadaljevanju 3PL) dodan še parameter, ki upošteva razlike v diskriminativnosti postavk (v obeh modelih) ter popravek za ugibanje (samo v 3PL).

Različni modeli se uporabljajo za različne tipe postavk. 3PL uporabljamo za izbirne odgovore, kjer se posameznik odloča med ponujenimi alternativami odgovorov. 2PL uporabljamo za odprte odgovore z možnostjo pravilnega ali napačnega odgovora; model z delnim točkovanjem pa za odprte odgovore, kjer je za pravilni odgovor možnih več kot ena točka (točkovanje ni dihotomno). Glede na izbran model za karakteristično krivuljo postavke uporabljamo in ocenjujemo različno število parametrov (različne kombinacije v različnih modelih):

- b_i $(-\infty, \infty)$ – težavnost – točka na lestvici lastnosti, kjer je verjetnost pravilnega odgovora udeleženca na postavko i 0,5 oziroma v primeru 3PL, točka na polovici med vrednostjo spodnje asimptote in 1 (parameter lokacije);

- a_i $(0, \infty)$ – diskriminativnost – proporcionalna glede na naklon tangente krivulje odgovora v točki 0,5 oziroma b (parameter naklona za postavko i);

- c_i $(0, 1)$ – spodnja asimptota (običajno predstavlja korekcijo za možnost ugibanja pri vprašanjih izbirnega tipa z več možnimi odgovori);

- $d_{i,l}$ (l=0, m_i-1) – pražni parameter kategorije (m je število kategorij odgovorov na postavko i).

TOP se uporablja pri konstrukciji lestvice latentne lastnosti. Modele TOP lahko uporabimo za merjenje osebnostnih potez, razpoloženjskih stanj, vedenjskih vzorcev in stališč, kakor

tudi kognitivnih lastnosti (Embretson in Reise 2000). »Merske instrumente moramo najprej ustvariti in določiti enote tako, da v ponovitvah pridemo do enakega rezultata« (Bond in Fox 2001, 3). Šele tedaj lahko uporabimo instrumente za merjenje želene lastnosti. Pred uporabo TOP moramo biti prepričani, da instrument (ali preizkus znanja ali test) meri želeno lastnost. Instrument v naslednjem koraku preizkusimo na reprezentativnem vzorcu populacije, kateri je namenjen. Slednje je še posebej pomembno, če želimo rezultate posploševati na populacijo.

10.3 Merjenje znanja v mednarodnih raziskavah znanja

V mednarodnih raziskavah znanja merijo znanje učencev s pomočjo objektivnih preizkusov znanja na vzorcu učencev, ki so bili izbrani kot predstavniki nacionalne populacije. Za oceno znanja učencev uporabljajo metodo matričnega razvrščanja postavk, kjer vsak učenec reši le določen nabor postavk. Postavk je namreč veliko več, kot jih lahko v razumnem času reši posamezni učenec. Zato postavke združijo v določeno število blokov (v TIMSS 2007 na primer 28, od tega 14 matematičnih blokov in 14 naravoslovnih blokov), ki jih v parih združujejo v zvezke (14 zvezkov nalog v TIMSS 2007). Bloki so sestavljeni tako, da so v njih enakomerno zastopana vsebinska področja, kognitivna področja in vrste postavk. Posamezni učenec odgovarja na postavke v enem zvezku. Vsak blok se ponovi običajno v dveh zvezkih, kar omogoča uporabo iste lestvice za vse učence, saj so zvezki med seboj povezani.

Pri posameznih učencih se pojavlja veliko manjkajočih vrednosti, ker vsak učenec rešuje le del postavk. Pri izračunu dosežkov se je za učinkovito izkazala kombinacija lestvičenja s TOP in metodologije večkratnega vstavljanja (angl. multiple imputation; posamezne vrednosti večkratnega vstavljanja se v mednarodnih raziskavah znanja imenujejo verjetnostne vrednosti). Lestvičenje, ki se sedaj uporablja v mednarodnih raziskavah znanja, so najprej uporabljali za ameriško nacionalno preverjanje znanja (v nadaljevanju NAEP, Beaton, in Johnson 1992). Ta pristop se je razvil v metodologijo verjetnostnih vrednosti, ki se danes uporablja v mednarodnih raziskavah znanja. Zaradi uporabljenega postopka zanesljivo ocenjevanje dosežka posameznika ni mogoče, ocene dosežkov pa so se izkazale kot zelo zanesljive za skupine udeležencev.

10.3.1 Postopek lestvičenja v mednarodnih raziskavah znanja

Postopek pridobitve verjetnostnih vrednosti je naslednji. Najprej je potrebno umeriti postavke v preizkusih znanja (izračunajo se ocene parametrov za vsako postavko, pri čemer so podatki uteženi tako, da vsaka država k parametrom prispeva enako). Nadalje rezultate metode glavnih komponent na spremenljivkah iz vprašalnika uporabijo v pogojevanju (angl. conditioning). Običajno upoštevajo toliko glavnih komponent, da skupaj pojasnijo 90% variance spremenljivk. Te glavne komponente se imenujejo pogojne spremenljivke (Olson in drugi 2008). Dosežek posameznika je predstavljen s pomočjo petih naključno izbranih vrednosti iz pogojne porazdelitve učenca, glede na njegove odgovore iz preizkusa znanja, vprašalnika in parametre postavk. Z vključitvijo vseh razpoložljivih podatkov v model (pogojevanje) se v verjetnostnih vrednostih ohrani odnos med spremenljivkami iz vprašalnika in ocenjenimi dosežki.

Ko so lestvico dosežkov uporabili prvič, sta bili določeni arbitrarni konstanti za aritmetično sredino (500 točk) in standardni odklon (100 točk). Izbrana lestvica se tako izogne negativnim vrednostim ter uporabi decimalnih mest pri poročanju o učenčevem dosežku (Gonzalez 1997). Da bi bila mogoča primerjava med različnimi ponovitvami raziskave, vse dosežke v nadaljnjih raziskavah umestijo na to lestvico.

Lestvice trendov (spremembe dosežkov v času) temeljijo na pristopu, ki se imenuje sočasno umerjanje postavk. Običajno sestoji iz treh korakov, ki vzpostavijo povezavo med trenutnim in prejšnjim umerjanjem.

Najprej vzpostavimo skupen set parametrov postavk na podlagi podatkov preteklega in trenutnega cikla (vključene so le države, ki so sodelovale v obeh ciklih in postavke, ki so bile uporabljene v obeh ciklih). Nato ocenimo parametre postavk, izračunamo aritmetično sredino in standardni odklon porazdelitve latentne lastnosti za posamezni cikel in opazujemo razlike med porazdelitvami:

- razliko med prejšnjim ciklom v sočasni kalibraciji ter trenutnim ciklom v sočasni kalibraciji – sprememba v dosežku;
- razliko med prejšnjim ciklom v sočasni kalibraciji ter prejšnjim ciklom v prejšnji kalibraciji – sprememba v ocenah parametrov postavk.

Zaradi spremembe v ocenah parametrov postavk predstavlja drugi korak iskanje linearne transformacije. Linearna transformacija je potrebna za popravek razlik v parametrih postavk, ki so posledica dejstva, da so bili podatki v prejšnjem ciklu v kalibraciji združeni z drugimi podatki (približno polovica postavk se ponovi, polovica pa je novih). Linearna transformacija uskladi porazdelitev podatkov preteklega cikla v sočasni kalibraciji s podatki preteklega cikla v prejšnji kalibraciji.

Zadnji korak je uporaba te linearne transformacije na podatkih trenutnega cikla, ki so bili lestvičeni v sočasni kalibraciji za vse sodelujoče države in ne samo za države, ki so sodelovale v obeh ciklih. S to transformacijo so podatki trenutnega cikla na enaki lestvici kot podatki prejšnjega cikla (Olson in drugi 2008).

10.4 Invariantnost parametrov v TOP

10.4.1 Opredelitev

Ena od glavnih prednosti TOP je invariantnost parametrov. To pomeni, da so parametri postavk neodvisni od podvzorca udeležencev in da so parametri lastnosti neodvisni od podvzorca uporabljenih postavk. Oboje sledi v primeru, ko so predpostavke za uporabo TOP izpolnjene in se podatki prilegajo modelu (Hambleton, Swaminathan in Rogers 1991). Lastnost invariantnosti pa ne pomeni, da bodo imele ocene iz podatkov identične lastnosti ne glede na vključene postavke ali udeležence. Zadostuje že, če so parametri postavk v linearni zvezi. Tudi natančnost ocen dveh različnih stopenj lastnosti se razlikuje med podvzorci postavk. Če so v podvzorec izbrane lažje postavke, bo nižja stopnja lastnosti bolj natančno ocenjena kot visoka stopnja. Podobno je tudi, če imajo v kalibracijskem vzorcu udeleženci v povprečju nižjo stopnjo lastnosti. V tem primeru bo težavnost lažjih postavk bolj natančno ocenjena kot težavnost težjih postavk (Embretson in Reise 2000).

10.4.2 Pretekle raziskave s področja invariantnosti parametrov postavk in dosežkov

Embretson in Reise (2000) navajata, da le Raschev model omogoča aditivnost in druge temeljne merske lastnosti. Številni psihometriki zavračajo kompleksnejše modele zaradi tega, ker naj ne bi zagotavljali objektivnega merjenja. Vendar pa zagovorniki kompleksnejših modelov pogosto poudarjajo, da se Raschev model slabo prilega empiričnim podatkom.

Primerjava postavk v modelih z dvema in tremi parametri ne dosega enake invariantnosti kot je pričakovana v Raschevem modelu (Embretson in Reise 2000). Fan (1998) poroča, da je parameter težavnosti v 1PL bolj invarianten od težavnosti v 2PL ali 3PL. Brown in drugi (2005) poročajo o razlikah v dosežkih učencev v raziskavi TIMSS 1995, če je uporabljen različen model (1PL ali 3PL). Mesto na lestvici držav se za posamezno državo skoraj ne spremeni, povezanost med dosežki učencev glede na uporabo različnega modela pa je zelo močna. Večje razlike med rezultati obeh modelov so opazili v državah, ki imajo nižji povprečni dosežek.

Nekateri raziskovalci (Hencke in drugi 2009) so proučevali, ali je izbor postavk v raziskavi TIMSS 2003 povezan s povprečnim dosežkom učencev v sodelujočih državah. V ocenjevanje parametrov so vključili le tiste postavke, za katere so v posamezni državi poročali, da so vključene v njihov učni načrt (in izključili postavke, katerih vsebin niso poučevali). Zaključili so, da se relativni položaj držav bistveno ne spremeni, ko so vključili le naloge, ki so skladne z učnimi načrti v posamezni državi (države z visokim dosežkom so imele visok dosežek ne glede na vključene naloge; države z nizkim dosežkom so ohranile nizek dosežek; v državah s srednjim dosežkom tudi ni bilo večjih razlik). Ugotovili so tudi, da se je v petih državah dosežek pomembno zvišal, če so bile vključene le naloge, ki so jih obravnavali po učnem načrtu v primerjavi z dosežkom, kjer so bile vključene vse naloge. V štirih državah se je povišal relativni rang za ena in v eni državi celo za šest. Vendar so bile razlike v dosežkih med državami, ki so spremenile pozicijo, majhne in statistično neznačilne.

Drugi raziskovalci (Monseur in Brezner 2007; Monseur in drugi 2008) so proučevali povezovalno napako pri oceni trendov v mednarodnih raziskavah znanja. Pod predpostavkami TOP bi morali v vsakem primeru dobiti podobno povezovalno funkcijo, ne glede na izbor postavk. Ugotovili so, da se povezovalna napaka povečuje, če se zmanjšuje

število skupnih postavk (negotovost indikatorjev trendov je inverzno proporcionalna številu skupnih postavk). Monseur in drugi (2008) nadalje navajajo, da preizkusi z manj nalogami dajo večjo povezovalno napako.

10.5 Raziskovalni problem in raziskovalna vprašanja

Modeli TOP se v raziskavah pogosto uporabljajo zaradi prednosti, ki jih nudijo pred klasično testno teorijo. Med drugim omogočajo primerjavo različnih preizkusov (niso potrebne vzporedne verzije testov) in identifikacijo različnega vedenja postavk znotraj podskupin. Ena najpomembnejših prednosti je ta, da je stopnja izraženosti posamezne lastnosti neodvisna od uporabljenih postavk, parametri postavk pa so neodvisni od vzorca udeležencev.

Mednarodne raziskave, v katerih ocenjujejo znanje učencev na različnih vsebinskih področjih, uporabljajo TOP zaradi njenih prednosti. Vsaka država, ki se odloči za sodelovanje v kateri od mednarodnih raziskav, mora slediti strogim smernicam, ki so določene za vsak del raziskave (prevod preizkusov znanja in vprašalnikov, vzorčenje, zahtevana odzivnost šol in učencev, mednarodno in nacionalno preverjanje izvedbe ...). Tako je zagotovljena enotnost standardov merjenja v vseh vključenih državah. To zagotavlja tudi vzdrževanje visokih standardov kakovosti za vse sodelujoče v raziskavi in zmanjšuje možnosti, da bi bile razlike v izsledkih posledica uporabe različnih postopkov. Nekatere države tem standardom ne zadostijo in zato niso vključene v mednarodno poročilo, prav tako pa tudi ne v ocenjevanje parametrov postavk (na primer Mongolija v raziskavi TIMSS 2007). Prav tako so iz skupine držav, na podlagi katerih se ocenjujejo parametri postavk, izključeni sodelujoči šolski sistemi (regionalne entitete, ki sledijo enakim postopkom kot države).

V ocenjevanje parametrov postavk so vključene samo države, ki so sodelovale v zaporednih izvedbah in ki s kakovostjo izvedbe raziskave sledijo mednarodnim postopkom. V zaporednih izvedbah sodelujejo različne države, zato so v ocenjevanje parametrov postavk vsakič vključene druge države. Pri ocenjevanju parametrov je pomembna neodvisnost dosežkov učencev in parametrov postavk. Bi bili izsledki oziroma zaključki enaki, če bi bile vključene druge države? Za izračun dosežkov je pomemben tudi izbor modela za ocenjevanje parametrov postavk. Do sedaj nam ni poznana celostna

študija, ki bi proučevala učinek uporabe modela na ocenjevanje latentne lastnosti. Na področju ocenjevanja branja pa se zaradi narave sestave preizkusov pojavlja možnost kršenja predpostavke o lokalni neodvisnosti postavk. Ker se pri branju postavke nanašajo na isto besedilo oziroma kontekst, nekateri avtorji (Monseur in drugi 2011) v tem primeru opozarjajo na možnost kršenja predpostavk za uporabo TOP.

V doktorskem delu smo se osredotočili na opazovanje invariantnosti parametrov postavk in dosežkov učencev v primerih, ko so v ocenjevanje parametrov vključene različne skupine držav. To vprašanje pa se ne zastavlja v raziskavi PISA, saj je sodelovanje držav v tej raziskavi bolj določeno in stabilno (OECD 2005, OECD 2009). Nadalje PISA v ocenjevanje parametrov postavk vključuje le države članice OECD. Je pa to vprašanje pomembno v mednarodnih raziskavah znanja, ki jih izvaja IEA, saj v ponovitvah raziskav sodelujejo različne države oziroma je sodelovanje držav v teh raziskavah spremenljivo. Zanimala nas je tudi invariantnost parametrov, ocenjenih s pomočjo različnih modelov in na različnih vsebinskih področjih, od katerih smo se omejili na matematiko in branje.

V tehničnem poročilu o raziskavah TIMSS in PIRLS je zelo malo podatkov o prileganju podatkov modelu. Zaslediti je le podatke o tem, da so bila po končani kalibraciji postavk izvedena preverjanja, ali parametri postavk primerno opisujejo opazovano porazdelitev odgovorov učencev vzdolž kontinuuma dosežkov (Olson in drugi 2008, 249). Ker pa se model v praksi nikoli popolnoma ne prilega podatkom, ostaja vprašanje o invariantnosti parametrov postavk in dosežkov odprto.

V doktorskem delu smo zato skušali odgovoriti na naslednja štiri raziskovalna vprašanja:

- Ali obstajajo razlike v ocenah parametrov in dosežkov, če v ocenjevanje parametrov postavk vključimo različno število držav?

- Ali ima povprečni dosežek vključenih držav kakšen učinek na ocenjene parametre postavk in dosežke učencev?

- Ali sta je invariantnost ocen parametrov postavk in dosežkov neodvisna od modela TOP, ki ga uporabimo za računanje parametrov postavk?

- Ali lahko enako invariantnost ocen parametrov postavk ter dosežkov ugotovimo na različnih vsebinskih področjih (področju matematike in bralne pismenosti)?

10.6 Opis raziskovalne metode

1.6.1 Podatki

Uporabili smo podatke dveh mednarodnih raziskav PIRLS 2006 in TIMSS 2007 ter ponovno izračunali dosežke učencev glede na različne pogoje. Referenčne parametre postavk smo dobili tako, da smo v ocenjevanje vključili vse sodelujoče države, pri čemer je vsaka država enako prispevala k parametrom postavk (uporabili smo utež SENWGT, ki je dostopna v mednarodni bazi in utežili učence tako, da jih je v vsaki državi po 500). Ti rezultati so nam služili kot osnova za primerjavo na novo izračunanih parametrov postavk in dosežkov učencev.

1.6.2 Opis pogojev

Za reševanje raziskovalnih vprašanj smo na novo ocenili parametre postavk in dosežke pod različnimi pogoji. Kalibracijski vzorec oziroma vzorec udeležencev, na katerem smo umerili (izračunali) parametre postavk, se tako med pogoji razlikuje. Vsakič smo najprej umerili postavke in nato izračunali dosežke učencev v vseh državah. Nove rezultate smo primerjali z referenčnimi vrednostmi parametrov.

Najprej smo v ocenjevanje parametrov postavk vključili različno število držav. Število vključenih držav je bilo 2, 3, 4, 6 in 10. Navzgor smo se omejili zaradi predpostavke, da se po določenem številu vključenih držav parametri postavk več ne bodo bistveno spreminjali. Tudi v teoriji se ne priporoča vzorcev večjih od 1000 (kar pri uporabljenih utežeh in številu manjkajočih podatkov predstavlja 10 držav), saj bistveno ne doprinesejo k natančnosti ocen parametrov (de Toit 2003). Države smo v vsakem izmed pogojev izbrali naključno in postopek ponovili stokrat znotraj vsakega pogoja ter tako pridobili tudi informacijo o variabilnosti rezultatov. Prav tako smo v vseh ponovitvah izračunali tudi dosežke učencev v vsaki državi.

Pri preverjanju naslednjega raziskovalnega vprašanja smo ocenjevali parametre postavk glede na vključevanje držav v kalibracijski vzorec po njihovem povprečnem dosežku. Za ugotavljanje povprečnega dosežka držav smo uporabili referenčne vrednosti. Države smo razvrstili po povprečnem dosežku ter nato izbrali zgornjo tretjino držav (15 držav) in spodnjo tretjino držav (15 držav). Nato smo ponovno ocenili parametre postavk za vsako skupino držav (višji in nižji dosežek) tako, da smo vsakič naključno izbrali 10 (izmed 15)

držav v skupini. Znotraj vsakega pogoja smo postopek ponovili 100 krat, na podlagi česar smo lahko sklepali o variabilnosti rezultatov. Na novo izračunane parametre postavk ter dosežke smo primerjali z referenčnimi.

Pri preverjanju tretjega raziskovalnega vprašanja smo za kalibracijo postavk uporabili različne modele. Primerjali smo modele iz družine Rasch (1PL in model z delnim točkovanjem) z logističnimi modeli z več parametri (3PL, 2PL, posplošeni model z delnim točkovanjem). V vsakem pogoju smo izmed 45 sodelujočih držav naključno izbrali 10 držav in jih vključili v kalibracijski vzorec. Znotraj vsakega pogoja pa smo izvedli 100 ponovitev.

Pri zadnjem raziskovalnem vprašanju smo se osredotočili na primerjavo med različnimi vsebinskimi področji. Za primerjavo smo izbrali bralno razumevanje (raziskava PIRLS) in matematiko (raziskava TIMSS). Raziskava TIMSS vključuje in preverja znanje učencev četrtih in osmih razredov, vendar smo se zaradi primerjave z raziskavo PIRLS odločili za upoštevanje zgolj četrtošolcev. V raziskavi PIRLS namreč sodelujejo učenci enake starosti. Zato smo izbrali zgolj države, ki so sodelovale v obeh ciklih raziskav (PIRLS 2006 in TIMSS 2007). Teh držav je bilo 29. Izmed teh držav smo neodvisno za PIRLS in za TIMSS izbrali po 10 držav v kalibracijski vzorec in na novo izračunali ocene parametrov. Znotraj vsakega pogoja smo izvedli 100 ponovitev.

1.6.3 Postopek

Za vsak pogoj smo najprej primerjali parametre postavk s pomočjo Pearsonovega koeficienta korelacije. V rezultatih prikazujemo in poročamo aritmetično sredino korelacijskega koeficienta za vsak parameter postavke med novo izračunanim in referenčnim parametrom. Rezultat predstavlja povprečno korelacijo med enakimi parametri postavk (težavnost, diskriminativnost, ugibanje in pražni parametri kategorij) za 100 ponovitev.

Nadalje smo za vsako državo primerjali tudi odstopanja na novo izračunanih dosežkov učencev od referenčnih. Ker je lestvica dosežkov poljubna, smo dosežke za namene primerjav v vsaki ponovitvi standardizirali na lestvico s povprečjem 500 in standardno deviacijo 100 točk. Nato smo primerjali aritmetično sredino in različne percentile (5., 10., 50., 90. in 95.) za posamezne države. V pogojevanje smo vključili tudi dve spremenljivki

iz vprašalnika za učence (spol in število knjig doma), kar nam je omogočilo tudi primerjavo povprečnih dosežkov različnih podskupin za sodelujoče države.

Za lestvičenje podatkov ter vzorčenje verjetnostnih vrednosti (ki predstavljajo dosežke učencev) smo uporabili enake postopke, kot jih uporabljajo v mednarodnih centrih. Za ocenjevanje parametrov postavk smo uporabili program PARSCALE 4.1. (2003), za generiranje dosežkov pa program DESI (2009).

10.7 Rezultati in interpretacija

Pri prvem raziskovalnem vprašanju nas je zanimal učinek velikosti vzorca na ocene parametrov. Ugotovljeni korelacijski koeficienti so zelo visoki za vse ocene parametrov postavk razen za ugibanje (v primeru dveh držav znaša 0.52 in naraste do 0.81 v primeru desetih držav) in naraščajo skladno z večanjem števila držav v vzorcu. Rezultati kažejo večja odstopanja od referenčnih vrednosti v primeru manjšega števila držav v kalibracijskem vzorcu.

Velikost učinka korelacijskih koeficientov nakazuje, da so parametri postavk v pogoju z desetimi vključenimi državami statistično pomembno bolj invariantni kot v pogoju z dvema ali s tremi državami v kalibracijskem vzorcu. To velja za vse parametre postavk razen za pražni parameter kategorije 3 (kar je verjetno posledica zelo majhnega števila postavk s tem parametrom). Kljub temu se kot najbolj invarianten izkazuje parameter težavnosti.

Do podobne ugotovitve je prišel že Fan (1998), ki poroča, da je parameter težavnosti v slučajnih vzorcih najbolj invarianten. Adedoyin, Nenty and Chillisa (2008) poročajo, da je parameter težavnosti invarianten tudi v primeru vzorcev različnih velikosti. Pri tem je potrebno opozoriti, da so sami raziskavo opravili na vzorcih z vsaj 1000 udeleženci (kar pa že predstavlja zadostno velikost vzorca za zanesljivo oceno). Tudi Macdonald in Paunonen (2002) zaključujeta, da je invariantnost parametra težavnosti visoka ne glede na število vključenih postavk in ne glede na razpon teoretičnih vrednosti težavnosti ali diskriminativnosti.

Ocene povprečnih dosežkov v državah niso pokazale velikih razlik med pogoji z različnim številom držav v kalibracijskem vzorcu. Ugotovljeni povprečni dosežki držav se med

pogoji statistično značilno niso razlikovali. Ocene percentilov v državah se med pogoji razlikujejo. Rezultati kažejo, da so razlike med percentili dosežkov po pogojih, v primerjavi z referenčnimi vrednostmi, statistično značilni. Opazili smo, da so razlike nižje, če je v ocenjevanje parametrov postavk vključenih več držav. Iz tega je mogoče zaključiti, da percentilne vrednosti izkazujejo večjo invariantnost, ko je v kalibracijski vzorec vključenih več držav (v našem primeru je bilo to deset držav). Opazili smo tudi večja odstopanja pri ekstremnih percentilih. Ocene bolj oddaljenih točk od povprečja so se izkazale za manj invariantne.

Omenjeni rezultati so v skladu z dosedanjimi ugotovitvami raziskovalcev, ki prav tako ugotavljajo invariantnost ocen latentnih lastnosti. Macdonald in Paunonen (2002) ugotavljata, da so ocenjene lastnosti oseb invariantne glede na vključeno težavnost in diskriminativnost postavk. Zaključujeta, da bodo testni rezultati za oceno posameznikove lastnosti v vsakem primeru konsistentni in točni.

Za preverjanje drugega raziskovalnega vprašanja smo ocenjevali parametre postavk glede na vključevanje držav po njihovem povprečnem dosežku. Parametre smo ponovno ocenili na podvzorcu vsake skupine držav (višji in nižji dosežek) in te parametre uporabili za izračun dosežkov vseh sodelujočih držav. Ponovno smo primerjali na novo izračunane parametre postavk ter dosežke z referenčnimi.

Pri korelacijah parametrov postavk nismo opazili večjih razlik med pogojema. Parametri postavk so se v obeh pogojih izkazali za zadovoljivo invariantne. Pri ocenah dosežkov učencev v sodelujočih državah pa so se pokazale pomembne razlike. Izkazalo se je, da na podlagi vključenosti držav z nižjim dosežkom v kalibracijski vzorec, učinkoviteje ocenimo dosežek vseh držav. Slednji izkazujejo statistično pomembno nižje razlike v primerjavi z referenčnimi vrednostmi, tako za povprečne kot tudi ekstremne percentilne vrednosti ter ocene podskupin, ki smo jih ocenjevali, kot pa dosežki, dobljeni na podlagi vključenosti držav z višjim dosežkom v kalibracijski vzorec. Razlike v dosežkih, izračunanih na podlagi vključenih držav z višjim dosežkom in referenčnimi dosežki, so vsaj trikrat večje kot tiste, dobljene na podlagi držav z nižjim dosežkom.

Ugotovljeni rezultati so nekoliko presenetljivi. Ob podrobnejšem pogledu na dosežke obeh skupin držav (iz katerih smo vzorčili države v kalibracijski vzorec) lahko ugotovimo, da so dosežki v skupini držav z višjim dosežkom mnogo bolj homogeni kot dosežki držav z nižjim dosežkom. Razpon dosežkov držav v skupini z višjim dosežkom je med 539 in 557

točkami, medtem ko je razpon držav v skupini z nižjim dosežkom mnogo večji (med 303 in 506 točk). Ti rezultati nakazujejo, da je v kalibracijskem vzorcu zelo pomembna pokritost oziroma razpršenost dosežkov. Slednja mora biti večja, saj tako bolj zanesljivo ocenimo dosežek v sodelujočih državah. Prav posebno pozornost moramo nameniti državam z nižjim povprečnim dosežkom, ki tudi v tem primeru v obeh pogojih nakazujejo večja odstopanja od referenčnih vrednosti.

Naslednje raziskovalno vprašanje obravnava invariantnost parametrov postavk in dosežkov učencev glede na uporabo različnih modelov TOP. Postopek izračuna dosežkov smo ponovili s pomočjo uporabe različnih modelov (modeli iz družine Rasch ter logistični modeli - 3PL, 2PL vključno s posplošenim modelom z delnim točkovanjem). Pri parametrih postavk smo lahko primerjali zgolj parameter težavnosti, saj z Raschevimi modeli ocenjujemo zgolj ta parameter. Parameter težavnosti se je izkazal za zelo invariantnega ne glede na uporabljeni model, kar je tudi v skladu z dosedanjimi rezultati drugih raziskovalcev (Macdonald in Paunonen 2002).

Rezultati ocen dosežkov v državah pa kažejo drugačno sliko. V primeru dosežkov, višjih od 400 točk, so razlike med modeli majhne. Večje razlike med modeli se pojavijo v primeru držav z nižjimi dosežki. Ko primerjamo kompleksnejše modele z Raschevimi, dobimo večje razlike v dosežkih, ki tudi zelo malo variirajo. Raschevi modeli so se izkazali sicer za zelo invariantne, ne glede na dosežke držav. Razlike med uporabljenimi modeli so na spodnjem delu porazdelitve dosežkov večje in hkrati stabilne. Slednje je ugotovil tudi Brown s sodelavci (2005).

Nazadnje smo opazovali in preverjali še invariantnost ocenjenih parametrov postavk ter dosežkov na različnih vsebinskih področjih (matematika in bralna pismenost). Med različnimi vsebinskimi področji smo izbrali branje in matematiko, saj ti dve spretnosti smatramo za osnovni in sta tudi bolj primerljivi med državami (saj naravoslovje v različnih državah poučujejo pri različnih predmetih).

Pri preverjanju in opazovanju različnih vsebinskih področij smo pričakovali večjo invariantnost na področju branja. Ugotovitve pa so pokazale, da je večja invariantnost dosežkov prisotna na področju matematike. Čeprav so bile razlike v absolutnih vrednostih, v primerjavi z referenčnimi, zelo majhne, na področju matematike v povprečju samo eno točko (na lestvici s povprečjem 500 in standardnim odklon 100), smo ugotovili srednjo velikost učinka. Večje razlike na področju matematike so odkrili tudi Monseur in drugi

(2011). Možna razlaga za takšne rezultate je tudi sestava kognitivnih preizkusov znanja, ki na področju matematike vsebujejo več postavk z izbirnimi odgovori. Pri slednjih ocenjujemo še parameter ugibanja, ki pa se je v vseh primerih izkazal za bolj variabilnega od ostalih parametrov.

10.8 Zaključki

Invariantnost parametrov predstavlja idealno stanje in ji ne bo zadoščeno, če katerakoli ocena parametra ne bo identična v različnih populacijah udeležencev ali merskih pogojih (v praksi gledamo na identičnost blažje in sicer kot na linearno transformacijo). Čeprav v našem primeru nismo opazili popolne invariantnosti parametrov, kažejo rezultati zelo visoke korelacije med parametri postavk v različnih pogojih. Kot najbolj invarianten se je izkazal parameter težavnosti, kot najbolj varianten pa parameter ugibanja.

V primeru naključnega izbora držav (ko smo opazovali učinek velikost vzorca) smo ugotovili majhne velikosti učinka za povprečne vrednosti. Čeprav so bili povprečni dosežki držav blizu referenčnim dosežkom, smo opazili večja odstopanja v primeru držav z nižjimi dosežki. Prav tako smo v teh državah opazili večjo variabilnost dosežkov, še posebej, če je bilo v kalibracijski vzorec vključeno manjše število držav.

Ko smo v kalibracijski vzorec vključili države z nižjim ali višjim dosežkom, se je pokazala nekoliko drugačna slika. Države z nižjimi dosežki so se izkazale za veliko učinkovitejše, saj so bili ocenjeni dosežki mnogo bližje referenčnim vrednostim za države, kot v primeru držav z višjim dosežkom. To se je izkazalo tudi pri ocenah vseh percentilov in dosežkov podskupin. Države z nižjimi dosežki so izkazovale večji razpon različnih dosežkov, kar se je potrdilo za ugodno. Zaključiti je mogoče, da je priporočljivo v kalibracijski vzorec vključiti države širokim razponom povprečnih dosežkov.

Raschevi modeli omogočajo zelo stabilno ocenjevanje parametrov. V primerjavi s kompleksnejšimi modeli pa so ocene dosežkov različne. Te razlike so zelo stabilne. Razlike v dosežkih držav so bodisi stabilno nizke bodisi stabilno visoke. Navedeno navaja na to, da je izbor modela pomemben faktor. Razlike so zelo verjetno nastale kot posledica vključenega parametra za ugibanje. Ta parameter je pomemben predvsem, ko ocenjujemo

nižje dosežke, saj je ugibanje pri učencih z manj znanja pogostejše. Zato je kontrola tega dejavnika pomembna.

V nasprotju s pričakovanji se je matematično vsebinsko področje izkazalo za manj invariantno v primerjavi z branjem. Absolutne razlike so sicer majhne, a je velikost učinka visoka. Razlog za večjo variantnost matematičnih dosežkov je lahko ponovno parameter ugibanja. Na področju matematike je namreč več nalog izbirnega tipa, ki dopuščajo ugibanje pravilnega odgovora.

Zaključimo lahko, da ima kalibracijski vzorec učinek predvsem pri ocenjevanju dosežkov v državah z nižjim povprečnim dosežkom in manj pri ocenjevanju parametrov, ki so se izkazali za relativno invariantne v vseh pogojih. V državah z nižjim dosežkom smo opazili večje razlike v primerjavi z referenčnimi dosežki kot pri državah s povprečnim dosežkom ali višjim dosežkom v vseh opazovanih pogojih. Na podlagi dobljenih rezultatov torej priporočamo posebno previdnost pri interpretaciji izjemno nizkih dosežkov. Vsekakor pa za kalibracijski vzorec v prihodnjih raziskavah priporočamo vključitev vsaj deset držav, ki imajo velik razpon različnih dosežkov. Dosežki učencev z nižjimi dosežkom bodo prav tako bolj učinkovito ocenjeni s pomočjo bolj zapletenih modelov, ki vključujejo več parametrov postavk in kontrolirajo možnost ugibanja. Prav tako rezultati nakazujejo stabilnejšo sliko v primeru manjšega števila postavk z možnostjo ugibanja.

10.9 Izvirni prispevek

Do sedaj se je veliko raziskovalcev osredotočalo predvsem na učinke položaja postavk, napako povezave ter njeno zvezo s številom postavk, ki so vključene v ocenjevanje trendov. Raziskovali so tudi invariantnost ocen parametrov na simuliranih podatkih. Za razliko od preteklih raziskav smo v doktorski disertaciji opazovali, kako k ocenam parametrov postavk prispeva sodelovanje različnih držav v mednarodnih raziskavah znanja. S tem smo ocenili, ali so parametri postavk in dosežki občutljivi na to, da v različnih ponovitvah mednarodnih raziskav sodelujejo različne države. Mere, povezane s postavkami in rezultate simulacijskih študij, smo dopolnili in ugotovili, kako izbor vzorca vpliva na parametre postavk in dosežke na realnih podatkih.

Izvirni prispevek doktorskega dela je torej v proučevanju invariantnosti ocen parametrov in dosežkov na realnih podatkih, ki uporabljajo specifično metodologijo. Ta nam omogoča globlji vpogled v invariantnost dosežkov učencev v mednarodnih raziskavah znanja, prispeva pa tudi k boljšemu razumevanju rezultatov mednarodnih raziskav znanja. Slednje je še posebej pomembno, ker v sodelujočih državah podatke uporabljajo tudi za pomembne odločitve pri reformah izobraževanja. Tako IEA kot OECD raziskave poleg preverjanja znanja zberejo tudi veliko dodatnih informacij, ki so povezane z znanjem učencev in nudijo dodatno osvetlitev stanja izobraževanja v posamezni državi. Primarni namen IEA raziskav (Mullis in drugi 2009) je pridobiti čim več informacij o dejavnikih, ki se povezujejo s stališči in z dosežki učencev, na katere je mogoče vplivati in tako izboljšati učinkovitost izobraževalnih sistemov.

Glavni namen doktorske naloge pa je bil v okviru raziskave navedenih štirih raziskovalnih vprašanj določiti učinek sestave kalibracijskega vzorca na ocene parametrov postavk in dosežkov na različnih vsebinskih področjih, uporaba različnih modelov in podati predloge o upoštevanju sestave kalibracijskega vzorca pri naslednjih raziskavah. Menimo, da smo z ugotovitvami in rezultati naloge prispevali k boljšemu razumevanju pojma oziroma lastnosti invariantnosti modelov TOP na realnih podatkih, ki se nujno ne prilegajo modelom. Hkrati pa rezultati nudijo veliko uporabnih podatkov, posebej za nadaljnje ponovitve mednarodnih raziskav znanja ter za druge raziskave, ki uporabljajo modele TOP.

Appendix A

Table A.1: Regression of MRSD on achievement in different conditions for selected statistics of interest

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
Average country achievement						
	2 countries	intercept	1.43	0.19	7.61	0.000
		linear	-10.63	1.26	-8.43	0.000
		quadratic	4.46	1.26	3.53	0.001
	3 countries	intercept	1.23	0.15	8.27	0.000
		linear	-8.83	1.00	-8.85	0.000
		quadratic	3.69	1.00	3.70	0.001
	4 countries	intercept	1.15	0.15	7.72	0.000
		linear	-8.34	1.00	-8.31	0.000
		quadratic	3.52	1.00	3.51	0.001
	6 countries	intercept	0.94	0.12	7.72	0.000
		linear	-6.93	0.82	-8.48	0.000
		quadratic	2.80	0.82	3.43	0.001
	10 countries	intercept	0.66	0.08	8.67	0.000
		linear	-4.49	0.51	-8.74	0.000
		quadratic	1.99	0.51	3.88	0.000
	Low achieving countries	intercept	0.77	0.12	6.45	0.000
		linear	-5.80	0.80	-7.24	0.000
		quadratic	2.33	0.80	2.91	0.006
	High achieving countries	intercept	2.59	0.40	6.52	0.000
		linear	-22.87	2.66	-8.59	0.000
		quadratic	8.20	2.66	3.08	0.004
	Rasch-Rasch	intercept	0.08	0.00	19.17	0.000
		linear	-0.45	0.03	-16.12	0.000
		quadratic	0.25	0.03	8.77	0.000
	Rasch-3PL 2PL GPCM	intercept	3.58	0.33	10.72	0.000
		linear	-26.07	2.24	-11.63	0.000
		quadratic	10.75	2.24	4.79	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	0.66	0.08	8.67	0.000
		linear	-4.49	0.51	-8.74	0.000
		quadratic	1.99	0.51	3.88	0.000
	Reading	intercept	0.57	0.06	10.31	0.000
		linear	-2.94	0.30	-9.84	0.000
		quadratic	0.46	0.30	1.54	0.136
	Mathematics	intercept	0.98	0.05	20.06	0.000
		linear	-2.52	0.26	-9.60	0.000
		quadratic	2.49	0.26	9.50	0.000
5^{th} percentile						

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
	2 countries	intercept	6.70	0.54	12.40	0.000
		linear	-40.05	3.62	-11.06	0.000
		quadratic	27.78	3.62	7.67	0.000
	3 countries	intercept	5.49	0.43	12.64	0.000
		linear	-32.53	2.91	-11.17	0.000
		quadratic	23.37	2.91	8.03	0.000
	4 countries	intercept	5.16	0.43	11.93	0.000
		linear	-30.76	2.90	-10.60	0.000
		quadratic	22.75	2.90	7.84	0.000
	6 countries	intercept	4.08	0.36	11.39	0.000
		linear	-25.14	2.40	-10.47	0.000
		quadratic	18.10	2.40	7.54	0.000
	10 countries	intercept	2.81	0.22	12.76	0.000
		linear	-16.40	1.48	-11.10	0.000
		quadratic	12.01	1.48	8.13	0.000
	Low achieving countries	intercept	2.36	0.50	4.72	0.000
		linear	-21.81	3.36	-6.50	0.000
		quadratic	15.63	3.36	4.66	0.000
	High achieving countries	intercept	14.18	1.01	13.99	0.000
		linear	-82.22	6.80	-12.09	0.000
		quadratic	61.74	6.80	9.08	0.000
	Rasch-Rasch	intercept	0.41	0.01	34.59	0.000
		linear	-0.90	0.08	-11.30	0.000
		quadratic	0.60	0.08	7.52	0.000
	Rasch-3PL 2PL GPCM	intercept	16.63	0.72	23.22	0.000
		linear	-60.58	4.80	-12.61	0.000
		quadratic	45.36	4.80	9.44	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	2.81	0.22	12.76	0.000
		linear	-16.40	1.48	-11.10	0.000
		quadratic	12.01	1.48	8.13	0.000
	Reading	intercept	2.75	0.18	15.57	0.000
		linear	-12.00	0.95	-12.63	0.000
		quadratic	4.47	0.95	4.71	0.000
	Mathematics	intercept	2.35	0.11	22.18	0.000
		linear	-8.30	0.57	-14.53	0.000
		quadratic	4.54	0.57	7.95	0.000
10^{th} percentile						
	2 countries	intercept	5.33	0.46	11.65	0.000
		linear	-33.71	3.07	-10.98	0.000
		quadratic	20.02	3.07	6.52	0.000
	3 countries	intercept	4.42	0.36	12.16	0.000
		linear	-27.46	2.44	-11.26	0.000
		quadratic	16.94	2.44	6.95	0.000
	4 countries	intercept	4.15	0.37	11.33	0.000

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
		linear	-25.98	2.46	-10.58	0.000
		quadratic	16.35	2.46	6.66	0.000
	6 countries	intercept	3.29	0.30	10.91	0.000
		linear	-21.03	2.02	-10.41	0.000
		quadratic	13.07	2.02	6.47	0.000
	10 countries	intercept	2.29	0.19	12.15	0.000
		linear	-13.68	1.26	-10.84	0.000
		quadratic	8.87	1.26	7.03	0.000
	Low achieving countries	intercept	1.99	0.39	5.05	0.000
		linear	-18.10	2.64	-6.86	0.000
		quadratic	10.78	2.64	4.09	0.000
	High achieving countries	intercept	11.36	0.82	13.84	0.000
		linear	-70.47	5.51	-12.79	0.000
		quadratic	45.08	5.51	8.18	0.000
	Rasch-Rasch	intercept	0.33	0.01	36.46	0.000
		linear	-0.67	0.06	-10.95	0.000
		quadratic	0.52	0.06	8.41	0.000
	Rasch-3PL 2PL GPCM	intercept	14.39	0.65	22.19	0.000
		linear	-59.75	4.35	-13.74	0.000
		quadratic	40.70	4.35	9.36	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	2.29	0.19	12.15	0.000
		linear	-13.68	1.26	-10.84	0.000
		quadratic	8.87	1.26	7.03	0.000
	Reading	intercept	2.17	0.14	15.16	0.000
		linear	-9.57	0.77	-12.41	0.000
		quadratic	2.82	0.77	3.65	0.001
	Mathematics	intercept	2.04	0.09	23.28	0.000
		linear	-6.84	0.47	-14.54	0.000
		quadratic	3.98	0.47	8.46	0.000
50^{th} percentile						
	2 countries	intercept	1.42	0.12	12.29	0.000
		linear	-7.08	0.77	-9.14	0.000
		quadratic	0.97	0.77	1.25	0.217
	3 countries	intercept	1.21	0.09	13.22	0.000
		linear	-5.98	0.62	-9.71	0.000
		quadratic	0.77	0.62	1.24	0.220
	4 countries	intercept	1.12	0.09	12.43	0.000
		linear	-5.61	0.61	-9.26	0.000
		quadratic	0.76	0.61	1.26	0.215
	6 countries	intercept	0.93	0.07	13.84	0.000
		linear	-4.81	0.45	-10.67	0.000
		quadratic	0.65	0.45	1.43	0.159
	10 countries	intercept	0.66	0.04	15.92	0.000
		linear	-3.22	0.28	-11.49	0.000

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
		quadratic	0.50	0.28	1.79	0.080
	Low achieving countries	intercept	0.78	0.06	12.16	0.000
		linear	-3.95	0.43	-9.23	0.000
		quadratic	0.88	0.43	2.06	0.045
	High achieving countries	intercept	2.28	0.35	6.52	0.000
		linear	-14.70	2.34	-6.27	0.000
		quadratic	-1.20	2.34	-0.51	0.612
	Rasch-Rasch	intercept	0.16	0.01	31.12	0.000
		linear	-0.36	0.03	-10.49	0.000
		quadratic	0.10	0.03	3.00	0.004
	Rasch-3PL 2PL GPCM	intercept	4.11	0.40	10.30	0.000
		linear	-26.14	2.68	-9.75	0.000
		quadratic	2.71	2.68	1.01	0.317
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	0.66	0.04	15.92	0.000
		linear	-3.22	0.28	-11.49	0.000
		quadratic	0.50	0.28	1.79	0.080
	Reading	intercept	0.52	0.04	11.84	0.000
		linear	-1.61	0.24	-6.76	0.000
		quadratic	-0.38	0.24	-1.61	0.119
	Mathematics	intercept	1.07	0.06	19.21	0.000
		linear	-1.96	0.30	-6.55	0.000
		quadratic	2.34	0.30	7.80	0.000
90^{th} percentile						
	2 countries	intercept	2.50	0.05	50.37	0.000
		linear	0.69	0.33	2.07	0.045
		quadratic	2.86	0.33	8.62	0.000
	3 countries	intercept	2.19	0.05	46.30	0.000
		linear	1.26	0.32	3.97	0.000
		quadratic	2.83	0.32	8.91	0.000
	4 countries	intercept	2.04	0.05	40.87	0.000
		linear	0.94	0.34	2.80	0.008
		quadratic	2.50	0.34	7.47	0.000
	6 countries	intercept	1.61	0.04	44.80	0.000
		linear	0.86	0.24	3.55	0.001
		quadratic	2.07	0.24	8.55	0.000
	10 countries	intercept	1.18	0.03	42.29	0.000
		linear	0.85	0.19	4.53	0.000
		quadratic	1.43	0.19	7.64	0.000
	Low achieving countries	intercept	0.84	0.05	17.81	0.000
		linear	0.55	0.32	1.74	0.089
		quadratic	0.31	0.32	0.98	0.332
	High achieving countries	intercept	5.79	0.17	34.04	0.000
		linear	2.51	1.14	2.20	0.033
		quadratic	12.37	1.14	10.85	0.000

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
	Rasch-Rasch	intercept	0.28	0.01	26.83	0.000
		linear	0.07	0.07	1.06	0.294
		quadratic	0.26	0.07	3.75	0.001
	Rasch-3PL 2PL GPCM	intercept	8.24	0.29	28.66	0.000
		linear	4.05	1.93	2.10	0.042
		quadratic	20.74	1.93	10.75	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	1.18	0.03	42.29	0.000
		linear	0.85	0.19	4.53	0.000
		quadratic	1.43	0.19	7.64	0.000
	Reading	intercept	1.13	0.03	43.08	0.000
		linear	-0.02	0.14	-0.12	0.908
		quadratic	1.10	0.14	7.80	0.000
	Mathematics	intercept	1.24	0.04	29.90	0.000
		linear	0.51	0.22	2.27	0.032
		quadratic	1.90	0.22	8.52	0.000
95^{th} percentile						
	2 countries	intercept	3.03	0.05	65.52	0.000
		linear	3.28	0.31	10.56	0.000
		quadratic	1.62	0.31	5.22	0.000
	3 countries	intercept	2.66	0.04	61.52	0.000
		linear	3.41	0.29	11.77	0.000
		quadratic	1.60	0.29	5.52	0.000
	4 countries	intercept	2.41	0.05	51.81	0.000
		linear	2.78	0.31	8.88	0.000
		quadratic	1.27	0.31	4.08	0.000
	6 countries	intercept	1.93	0.04	54.83	0.000
		linear	2.36	0.24	9.99	0.000
		quadratic	1.10	0.24	4.64	0.000
	10 countries	intercept	1.41	0.03	50.28	0.000
		linear	1.87	0.19	9.94	0.000
		quadratic	0.74	0.19	3.95	0.000
	Low achieving countries	intercept	0.96	0.05	17.99	0.000
		linear	0.24	0.36	0.68	0.498
		quadratic	-0.14	0.36	-0.39	0.695
	High achieving countries	intercept	7.14	0.19	38.59	0.000
		linear	9.29	1.24	7.48	0.000
		quadratic	8.84	1.24	7.12	0.000
	Rasch-Rasch	intercept	0.37	0.01	34.07	0.000
		linear	0.34	0.07	4.60	0.000
		quadratic	0.30	0.07	4.16	0.000
	Rasch-3PL 2PL GPCM	intercept	10.32	0.34	30.61	0.000
		linear	13.10	2.26	5.79	0.000
		quadratic	19.10	2.26	8.44	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	1.41	0.03	50.28	0.000

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
		linear	1.87	0.19	9.94	0.000
		quadratic	0.74	0.19	3.95	0.000
	Reading	intercept	1.35	0.03	45.01	0.000
		linear	0.72	0.16	4.46	0.000
		quadratic	0.76	0.16	4.73	0.000
	Mathematics	intercept	1.49	0.03	50.31	0.000
		linear	1.16	0.16	7.25	0.000
		quadratic	1.96	0.16	12.26	0.000

Table A.2: Regression of standard deviation of MRSD on achievement in different conditions for selected statistics of interest

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
SD(Average country achievement)						
	2 countries	intercept	0.94	0.08	12.08	0.000
		linear	-5.67	0.52	-10.87	0.000
		quadratic	2.60	0.52	5.00	0.000
	3 countries	intercept	0.88	0.09	9.64	0.000
		linear	-5.80	0.61	-9.52	0.000
		quadratic	2.78	0.61	4.56	0.000
	4 countries	intercept	0.83	0.09	9.57	0.000
		linear	-5.51	0.58	-9.46	0.000
		quadratic	2.52	0.58	4.32	0.000
	6 countries	intercept	0.70	0.08	8.83	0.000
		linear	-5.07	0.53	-9.57	0.000
		quadratic	2.17	0.53	4.09	0.000
	10 countries	intercept	0.54	0.07	8.23	0.000
		linear	-3.97	0.44	-9.04	0.000
		quadratic	1.60	0.44	3.64	0.001
	Low achieving countries	intercept	0.31	0.03	10.44	0.000
		linear	-1.69	0.20	-8.45	0.000
		quadratic	0.73	0.20	3.66	0.001
	High achieving countries	intercept	0.25	0.01	17.76	0.000
		linear	-1.21	0.09	-12.86	0.000
		quadratic	0.56	0.09	5.95	0.000
	Rasch-Rasch	intercept	0.06	0.00	21.31	0.000
		linear	-0.30	0.02	-16.84	0.000
		quadratic	0.15	0.02	8.60	0.000
	Rasch-3PL 2PL GPCM	intercept	0.09	0.01	17.23	0.000
		linear	-0.53	0.04	-14.48	0.000

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
		quadratic	0.28	0.04	7.71	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	0.54	0.07	8.23	0.000
		linear	-3.97	0.44	-9.04	0.000
		quadratic	1.60	0.44	3.64	0.001
	Reading	intercept	0.49	0.05	9.13	0.000
		linear	-2.65	0.29	-9.15	0.000
		quadratic	0.10	0.29	0.36	0.724
	Mathematics	intercept	0.73	0.04	20.43	0.000
		linear	-2.01	0.19	-10.43	0.000
		quadratic	1.49	0.19	7.73	0.000
SD(5^{th} percentile)						
	2 countries	intercept	4.14	0.21	19.74	0.000
		linear	-21.81	1.41	-15.52	0.000
		quadratic	12.87	1.41	9.16	0.000
	3 countries	intercept	3.95	0.24	16.65	0.000
		linear	-22.70	1.59	-14.26	0.000
		quadratic	14.27	1.59	8.96	0.000
	4 countries	intercept	3.77	0.23	16.68	0.000
		linear	-21.35	1.52	-14.08	0.000
		quadratic	13.39	1.52	8.83	0.000
	6 countries	intercept	3.28	0.22	15.09	0.000
		linear	-19.53	1.46	-13.38	0.000
		quadratic	12.47	1.46	8.54	0.000
	10 countries	intercept	2.44	0.18	13.62	0.000
		linear	-15.47	1.20	-12.86	0.000
		quadratic	10.18	1.20	8.46	0.000
	Low achieving countries	intercept	0.95	0.09	11.03	0.000
		linear	-5.25	0.58	-9.04	0.000
		quadratic	3.73	0.58	6.42	0.000
	High achieving countries	intercept	1.08	0.03	42.01	0.000
		linear	-4.35	0.17	-25.26	0.000
		quadratic	2.14	0.17	12.41	0.000
	Rasch-Rasch	intercept	0.29	0.01	32.65	0.000
		linear	-0.59	0.06	-9.85	0.000
		quadratic	0.33	0.06	5.41	0.000
	Rasch-3PL 2PL GPCM	intercept	0.48	0.01	34.97	0.000
		linear	-1.12	0.09	-12.28	0.000
		quadratic	0.73	0.09	7.96	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	2.44	0.18	13.62	0.000
		linear	-15.47	1.20	-12.86	0.000
		quadratic	10.18	1.20	8.46	0.000
	Reading	intercept	2.93	0.18	16.34	0.000
		linear	-13.81	0.96	-14.32	0.000
		quadratic	4.67	0.96	4.84	0.000

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
	Mathematics	intercept	1.99	0.08	24.40	0.000
		linear	-6.59	0.44	-14.97	0.000
		quadratic	3.27	0.44	7.43	0.000
SD(10th percentile)						
	2 countries	intercept	3.32	0.17	19.46	0.000
		linear	-18.55	1.15	-16.20	0.000
		quadratic	9.17	1.15	8.01	0.000
	3 countries	intercept	3.14	0.20	15.50	0.000
		linear	-19.09	1.36	-14.06	0.000
		quadratic	10.50	1.36	7.73	0.000
	4 countries	intercept	3.00	0.19	15.63	0.000
		linear	-18.12	1.29	-14.08	0.000
		quadratic	9.80	1.29	7.62	0.000
	6 countries	intercept	2.58	0.19	13.95	0.000
		linear	-16.46	1.24	-13.24	0.000
		quadratic	9.05	1.24	7.28	0.000
	10 countries	intercept	1.94	0.15	12.54	0.000
		linear	-12.92	1.04	-12.48	0.000
		quadratic	7.31	1.04	7.06	0.000
	Low achieving countries	intercept	0.81	0.07	11.64	0.000
		linear	-4.54	0.47	-9.72	0.000
		quadratic	2.87	0.47	6.14	0.000
	High achieving countries	intercept	0.91	0.02	37.52	0.000
		linear	-3.66	0.16	-22.53	0.000
		quadratic	1.77	0.16	10.86	0.000
	Rasch-Rasch	intercept	0.24	0.01	35.45	0.000
		linear	-0.47	0.05	-10.13	0.000
		quadratic	0.32	0.05	6.84	0.000
	Rasch-3PL 2PL GPCM	intercept	0.40	0.01	37.16	0.000
		linear	-0.85	0.07	-11.83	0.000
		quadratic	0.60	0.07	8.27	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	1.94	0.15	12.54	0.000
		linear	-12.92	1.04	-12.48	0.000
		quadratic	7.31	1.04	7.06	0.000
	Reading	intercept	2.26	0.15	15.32	0.000
		linear	-11.04	0.79	-13.93	0.000
		quadratic	2.85	0.79	3.60	0.001
	Mathematics	intercept	1.69	0.07	23.86	0.000
		linear	-5.54	0.38	-14.56	0.000
		quadratic	2.77	0.38	7.27	0.000
SD(50th percentile)						
	2 countries	intercept	0.96	0.06	16.60	0.000
		linear	-3.84	0.39	-9.86	0.000
		quadratic	0.71	0.39	1.82	0.077

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
	3 countries	intercept	0.87	0.06	15.58	0.000
		linear	-4.05	0.38	-10.78	0.000
		quadratic	0.78	0.38	2.08	0.043
	4 countries	intercept	0.80	0.06	13.73	0.000
		linear	-3.86	0.39	-9.84	0.000
		quadratic	0.70	0.39	1.78	0.083
	6 countries	intercept	0.67	0.05	13.03	0.000
		linear	-3.31	0.34	-9.63	0.000
		quadratic	0.50	0.34	1.45	0.155
	10 countries	intercept	0.52	0.04	13.15	0.000
		linear	-2.63	0.27	-9.93	0.000
		quadratic	0.25	0.27	0.94	0.351
	Low achieving countries	intercept	0.34	0.02	17.24	0.000
		linear	-1.32	0.13	-9.81	0.000
		quadratic	0.27	0.13	1.99	0.053
	High achieving countries	intercept	0.30	0.02	14.31	0.000
		linear	-1.04	0.14	-7.37	0.000
		quadratic	-0.04	0.14	-0.28	0.777
	Rasch-Rasch	intercept	0.12	0.00	31.50	0.000
		linear	-0.24	0.02	-9.63	0.000
		quadratic	0.08	0.02	3.25	0.002
	Rasch-3PL 2PL GPCM	intercept	0.18	0.01	27.69	0.000
		linear	-0.44	0.04	-9.90	0.000
		quadratic	0.13	0.04	2.94	0.005
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	0.52	0.04	13.15	0.000
		linear	-2.63	0.27	-9.93	0.000
		quadratic	0.25	0.27	0.94	0.351
	Reading	intercept	0.43	0.05	9.45	0.000
		linear	-1.42	0.24	-5.84	0.000
		quadratic	-0.60	0.24	-2.46	0.021
	Mathematics	intercept	0.78	0.04	19.47	0.000
		linear	-1.53	0.22	-7.07	0.000
		quadratic	1.39	0.22	6.45	0.000
SD(90^{th} percentile)						
	2 countries	intercept	1.69	0.04	39.29	0.000
		linear	0.45	0.29	1.55	0.128
		quadratic	2.07	0.29	7.18	0.000
	3 countries	intercept	1.56	0.04	42.75	0.000
		linear	0.77	0.24	3.14	0.003
		quadratic	2.33	0.24	9.54	0.000
	4 countries	intercept	1.46	0.04	40.39	0.000
		linear	0.61	0.24	2.53	0.015
		quadratic	2.01	0.24	8.32	0.000
	6 countries	intercept	1.23	0.03	40.46	0.000

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
		linear	0.58	0.20	2.86	0.007
		quadratic	1.66	0.20	8.17	0.000
	10 countries	intercept	0.91	0.02	40.47	0.000
		linear	0.34	0.15	2.22	0.032
		quadratic	1.32	0.15	8.73	0.000
	Low achieving countries	intercept	0.50	0.02	30.15	0.000
		linear	0.32	0.11	2.89	0.006
		quadratic	0.34	0.11	2.99	0.005
	High achieving countries	intercept	0.56	0.01	42.09	0.000
		linear	0.01	0.09	0.14	0.890
		quadratic	1.13	0.09	12.72	0.000
	Rasch-Rasch	intercept	0.21	0.01	27.10	0.000
		linear	0.05	0.05	1.06	0.294
		quadratic	0.18	0.05	3.47	0.001
	Rasch-3PL 2PL GPCM	intercept	0.33	0.01	30.43	0.000
		linear	0.19	0.07	2.68	0.010
		quadratic	0.30	0.07	4.07	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	0.91	0.02	40.47	0.000
		linear	0.34	0.15	2.22	0.032
		quadratic	1.32	0.15	8.73	0.000
	Reading	intercept	1.11	0.03	41.36	0.000
		linear	-0.37	0.15	-2.53	0.018
		quadratic	1.49	0.15	10.29	0.000
	Mathematics	intercept	0.98	0.03	32.04	0.000
		linear	0.41	0.16	2.48	0.020
		quadratic	1.55	0.16	9.47	0.000
SD(95^{th} percentile)						
	2 countries	intercept	2.06	0.04	47.35	0.000
		linear	2.16	0.29	7.41	0.000
		quadratic	1.46	0.29	5.01	0.000
	3 countries	intercept	1.92	0.04	51.92	0.000
		linear	2.49	0.25	10.03	0.000
		quadratic	1.61	0.25	6.48	0.000
	4 countries	intercept	1.79	0.03	52.61	0.000
		linear	2.12	0.23	9.28	0.000
		quadratic	1.35	0.23	5.90	0.000
	6 countries	intercept	1.51	0.03	53.45	0.000
		linear	1.77	0.19	9.37	0.000
		quadratic	1.04	0.19	5.48	0.000
	10 countries	intercept	1.09	0.02	52.08	0.000
		linear	1.28	0.14	9.08	0.000
		quadratic	0.81	0.14	5.73	0.000
	Low achieving countries	intercept	0.57	0.02	30.16	0.000
		linear	0.53	0.13	4.15	0.000

Statistic of interest	Condition	*Coeff*	*Estimate*	*se*	*t*	*p*
		quadratic	0.20	0.13	1.59	0.120
	High achieving countries	intercept	0.68	0.02	38.87	0.000
		linear	0.46	0.12	3.92	0.000
		quadratic	0.99	0.12	8.45	0.000
	Rasch-Rasch	intercept	0.27	0.01	34.10	0.000
		linear	0.24	0.05	4.47	0.000
		quadratic	0.23	0.05	4.39	0.000
	Rasch-3PL 2PL GPCM	intercept	0.44	0.01	32.35	0.000
		linear	0.46	0.09	5.03	0.000
		quadratic	0.37	0.09	4.05	0.000
	3PL 2PL GPCM -3PL 2PL GPCM	intercept	1.09	0.02	52.08	0.000
		linear	1.28	0.14	9.08	0.000
		quadratic	0.81	0.14	5.73	0.000
	Reading	intercept	1.33	0.03	39.97	0.000
		linear	0.55	0.18	3.09	0.005
		quadratic	1.08	0.18	6.01	0.000
	Mathematics	intercept	1.19	0.02	54.14	0.000
		linear	0.76	0.12	6.45	0.000
		quadratic	1.64	0.12	13.87	0.000