Methodological Issues in Large-Scale Educational Surveys

Khurrem Jehangir

METHODOLOGICAL ISSUES IN LARGE-SCALE EDUCATIONAL SURVEYS

Khurrem Jehangir

Graduation Committee

Chairman:	**prof. dr. T.A.J. Toonen**
Promotor:	**prof. dr. C.A.W. Glas**
Members:	**prof. dr. J. Hartig**
	prof. dr. R.R. Meijer
	prof. dr. A. Need
	prof. dr. B.P. Veldkamp
	prof. dr. A.J. Visscher

ISBN 978-90-365-3959-3
DOI-number is: 10.3990/1.9789036539593
printed by Ipskamp Drukkers, Enschede

METHODOLOGICAL ISSUES IN LARGE-SCALE EDUCATIONAL SURVEYS

Dissertation

to obtain
the degree of doctor at the University of Twente
on the authority of the rector magnificus
prof. dr. H. Brinksma
on account of the decision of the graduation committee
to be publicly defended on
Thursday, October 29th, 2015, at 16.45

By

Khurrem Jehangir
Born on January 22nd, 1977
in Neuilly-sur-Seine, France

This dissertation has been approved by the promotors:

Prof. dr. C.A.W. Glas

ACKNOWLEDGEMENTS

This thesis is the fruit of research that was carried out at the department of Measurement, Research Methodology and Data Analysis of the University of Twente under the supervision of Prof. Dr. C.A.W. Glas. It was a privilege to have him as my mentor: first as his research assistant during the PISA project and subsequently while I was doing my PhD. My thanks and deep gratitude are due to him for his guidance and help in the writing of the thesis.

I thank Henk Moelands, Jose Nijons and Joke Kordes of CITO International with whom I had an excellent cooperation when I was the coordinator between the University of Twente and the stakeholders in the PISA project. In this respect I also like to thank Eveline Gebhardt of ACER in Melbourne, Australia.

Many thanks are due to my friends and colleagues in the OMD department and I like to mention particularly Jean Paul Fox for his invaluable advice during the last period and Wim Tielen who helped me a lot by correcting faults in the software. Further I like to thank Stephanie van den Berg who has critically appraised the research during my final year and Bernard Veldkamp for his comments and suggestions. I also thank Birgit and Lorette who have been taking care of many things and Naveed Khalid and Hanneke Geerlings for their willingness to help out whenever it was needed.

A special thanks goes to my uncle Prof.Dr. S.A.P.L. Cloetingh. He convinced my parents to send my brother and me for further education to the Netherlands and I remember the day that he went with us to the University of Twente for the admission.

The late Piet Grootswagers and Annemieke Grootswagers have been very kind and I like to thank Annemieke.

I am very thankful to my parents Anika Cloetingh and Khalid Jehangir who never failed to encourage me and to my brother Assed Jehangir and his wife Tania Tariq for their support.

Khurrem Jehangir
29 October, 2015, Enschede, the Netherlands

Table of Contents

Chapter 1

Introduction

This thesis focuses on the application of item response theory (IRT) in the context of large scale international educational surveys like PISA 2009 (OECD). Although IRT methodology has been widely used in educational applications such as test construction, norming of examinations, detection of item bias and computerized adaptive testing, the context of large scale surveys presents a number of specific problems. A number of these problems are addressed in this thesis. The procedures are illustrated using student questionnaire data of the 2006 and 2009 cycle of the PISA study.

The first problem in international comparative educational tests relates to the detection of cultural bias over countries. In this thesis, we target a problem know as country-specific Differential Item Functioning (CDIF) or as country-by-item-interaction. Statistical tests to detect differential item functioning are available, but the huge number of students and countries presents feasibility problems related to the power of the tests and presentation and interpretation of the results. The power problem is related to the fact that with a sample size of students exceeding half a million even the tiniest model violation becomes significant. Still, many well-founded test statistics for IRT models (see, for instance, Glas & Suárez-Falcón, 2003) are based on residuals (differences between predictions from the model used and actual observations) that can shed light on the severity of the model violation. Further, this information can be used to model CDIF using country specific item parameters. In this approach, it is assumed that a scale consist of both items which are free of CDIF and items that may be subject to CDIF. The first set of items ensures the validity of the measure across countries. The second set of items is calibrated concurrently with the first set of items and both sets of items contribute to measurement precision. Tests of model fit are used to establish that the two sets of items relate to the same latent variable, that is, the same construct, yet with different item parameters across countries.

In Chapter 2, this methodology is outlined and applied to the field trial of the background questionnaires of the PISA 2009 cycle (this chapter was published in the *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, edited by Rutkowski, von Davier, and Rutkowski, as Glas & Jehangir (2014). However, besides in the background questionnaires, CDIF can also play a role in the assessment of cognitive outcomes. In fact, in an article in *Psychometrika*, titled '*Analysis of model fit and robustness, a new look at the PISA scaling model underlying ranking of countries according to reading literacy*', Svend Kreiner and Karl Bang Christensen (K&C) heavily criticize both the methodology of the PISA project, both with respect to the use of the Rasch model (Rasch, 1960) and the presence of CDIF. According to K&C, their analysis provides strong evidence of misfit of the PISA scaling model and especially very strong evidence of CDIF in the PISA 2006 reading dataset. Based on these findings they assert that the country rankings reported by PISA 2006 are not reliable. In Chapter 3, K&C's main criticism concerning the impact of CDIF on the ranking of countries in PISA 2006 is investigated with the conclusion that the K&C critique is inappropriate. In Chapter 4, the practical significance of modeling CDIF on the background questionnaire scales is studied not only in terms of the ordering of countries on the respective scales but also its impact on the results of regression analyses with latent variables in survey research. Chapter 4 was published in *Measurement: Journal of the International Measurement Confederation* (Jehangir, Van Den Berg, & Glas, 2015).

Another problem related to using IRT in large-scale international educational surveys pertains to issues of test administration. IRT gives flexibility in managing practical issues that a large scale survey entails. IRT separates person and item parameters and thus allows for the use of incomplete item administration designs (in educational measurement usually referred to as booklet designs) that support domain coverage through the administration of a large number of items while limiting the response burden of students. However, early in the history of the PISA project, it became clear that the positon of an item in a booklet influenced the item difficulty parameters. The problem was addressed by the introduction of so-called booklet parameters. These booklet parameters are assumed to be valid for all countries. In Chapter 5, the validity of this approach is evaluated by comparing it to alternatives, one that allows for booklet-by-country interaction, one using position parameters, and one using position-by-country interaction parameters.

The final topic relates to the combination of the results of IRT measurement models with multilevel structural models to relate cognitive outcomes to background variables. Several

procedures are available. A much used procedure is to generate so-called plausible values from the measurement model, that is, the IRT, conditionally on principle components of background variables, and to estimate latent regression models conditional on these plausible values. Alternatives are concurrent estimation of the measurement and latent regression model and a two-step procedure where the measurement model is estimated first and the latent regression model is estimated using the item parameters obtained in the first step as covariates. The motivation for the plausible values approach is that concurrent and two-step estimation methods are complicated and require dedicated software that is not generally available to practitioners. Therefore, in datasets like the PISA dataset, plausible values for outcome variables and maximum likelihood estimates for latent background variables are already provided in the dataset for use by secondary researchers. In Chapter 6, a study is reported that investigates if the results of the different estimation procedures lead to comparable inferences in latent regression analyses. Finally, Chapter 7 gives an example of an advanced latent regression model based on plausible value methodology that explores the relation between socio-economic status and reading achievement in PISA 2009 through an intercepts-and-slopes-as-outcomes paradigm. Chapter 7 was published in the *Journal of Educational Research* (Jehangir, Glas, & Van den Berg, 2015).

Chapter 2

Modeling Country-specific Differential Item Functioning in Large-Scale Surveys

Fit to item response theory (IRT) models in large-scale surveys that transcend national and cultural boundaries, such as the PISA project can be compromised by the presence of country-specific or culture-specific differential item functioning (CDIF). The current chapter proposes methods to detect CDIF and explores the feasibility of improving the fit of the measurement model by using country-specific item parameters to model CDIF. In this approach, it is assumed that a scale consist of both items which are free of CDIF and items with CDIF. The first set of items ensures the validity of the measure across countries. The second set of items is calibrated concurrently with the first set of items and both sets of items contribute to measurement precision. Tests of model fit are used to establish that the two sets of items relate to the same latent variable, that is, the same construct, yet with different item parameters. The procedure is illustrated using student questionnaire data of the 2009-cycle of the PISA study. Using data of OECD countries, concurrent maximum marginal likelihood (MML) estimates of the parameters the partial credit model (PCM) and the generalized partial credit model (GPCM) are obtained. Then information on observed and expected response frequencies is used to identify CDIF items. Country-specific item parameters are introduced for the items with the largest effect sizes of CDIF and new MML estimates are obtained. The impact of using country-specific item parameters is evaluated by comparing the ordering of the countries on the latent variables measured without and with a model for CDIF.

2.1 Introduction

The growing awareness of the importance of education for knowledge economies has led to even greater emphasis on improving educational systems. Educational surveys play a prominent role in taking stock of the state of educational systems. Educational surveys not only depict the

current state of the educational system but also help identify weaknesses and handicaps that can be addressed with proper policy planning. Large-scale educational surveys enable comparisons of large groups of students within countries and across countries. They allow countries to gauge the performance of their populations on a comparative scale, to evaluate their global position and to get insights into factors which determine the effectiveness of their educational systems.

However, large-scale surveys are a complex undertaking and present many challenges, especially with respect to ensuring that the results are comparable across diverse target groups. An especially important problem has to do with cultural bias. Modern educational surveys not only measure the cognitive abilities of students in areas of interest but also include a set of context or background questionnaires which measure background variables that serve as possible determinants of educational achievement. CDIF may occur both in cognitive items and in items of background questionnaires, but the background questionnaires may be more vulnerable. CDIF in achievement tests may, for example, occur through the content of the context stories in a math or language achievement test. Still, though the framing of a question may influence the response behavior, it is reasonable to assume that the underlying construct, say math achievement or language comprehension, is stable over countries and cultures. In background questionnaires, cultural bias may be more prominent. Firstly, it is no minor task to define constructs such as the socio-economic status or the pedagogical climate in such a way that they allow for comparisons over countries and cultures and, secondly, culture-related response tendencies may bias the comparability between countries and cultures.

Both educational achievement and most of the explanatory variables on student-, parent-, teacher-, classroom and school-level are viewed as latent variables. The data from tests of educational achievement and the background questionnaires are usually analyzed using item response theory models (IRT, see, for instance, Lord, 1980) models. In this chapter, we present statistical methodology to identify CDIF and to account for CDIF. This methodology will be applied to the background questionnaires that were used in the 2009-cycle of the PISA (Program for International Student Assessment) survey. In the PISA study, the data of the background questionnaires were modeled using an exponential family IRT model, that is, the partial credit model (PCM, see, Masters, 1982). The statistical methodology presented here, will be developed in the framework of the PCM, but also in the framework of a more general model, the generalized partial credit model (GPCM, see, Muraki, 1992). To assess the impact of modeling CDIF using the PCM and GPCM, the rank order of the participating countries on the constructs measured by the PISA background questionnaires will be evaluated.

2.2 Item Response Theory

The background questionnaires in the PISA project consist mostly of polytomously scored items, that is, the scores on an item indexed i ($i = 1,2,....,K$) are integers between 0 and m_i, where m_i is the maximum score on item i. In the GPCM the probability of a student n ($n = 1,...,N$) scoring in category j on item i (denoted by $X_{nij} = 1$) is given by

$$P(X_{nij} = 1 \mid \theta_n) = P_{ij}(\theta_n) = \frac{\exp(j\alpha_i\theta_n - \beta_{ij})}{1 + \sum_{h=1}^{M_i} \exp(h\alpha_i\theta_n - \beta_{ih})}, \quad (2.1)$$

for $j = 1,..., M_i$. Note that the probability of a response in category $j = 0$ is thus given by

$$P(X_{ni0} = 1 \mid \theta_n) = P_{i0}(\theta_n) = \frac{1}{1 + \sum_{h=1}^{M_i} \exp(h\alpha_i\theta_n - \beta_{ih})}. \quad (2.2)$$

An example of the category response functions $P_{ij}(\theta)$ for an item with four response categories is given in Figure 1. The graph also shows the item-total score function (ITF)

$$E(T_i \mid \theta) = \sum_{j=1}^{M_i} jE(X_{ij} \mid \theta) = \sum_{j=1}^{M_i} jP_{ij}(\theta), \quad (2.3)$$

where the item-total score is defined as $T_i = \Sigma_j jX_{ij}$. Note that the ITF increases as a function of θ. The location of the response curves are related to location parameters defined by $\delta_{i1} = \beta_{i1}$ and $\delta_{ij} = \beta_{ij} - \beta_{i(j-1)}$, for $j = 2,...,m_i$. The location parameter δ_{ij} is the position on the θ-scale where the curves $P_{i(j-1)}(\theta)$ and $P_{ij}(\theta)$ intersect. Finally, the so-called discrimination parameter α_i gauges the kurtosis of the curves. If the discrimination parameters for all items are constrained to one, the GPCM specializes to the PCM.

The GPCM and the PCM are not the only IRT models giving rise to sets of response curves where a higher level on the latent scale, i.e., the θ-scale, is associated with a tendency to score in a higher response category. The sequential model by Tutz, (1990) and the graded response

model by Samejima (1969) have response curves which can hardly be distinguished in the basis of empirical data (Verhelst, Glas, & de Vries, 1997). Therefore, the choice between the GPCM and these two alternatives is not essential.

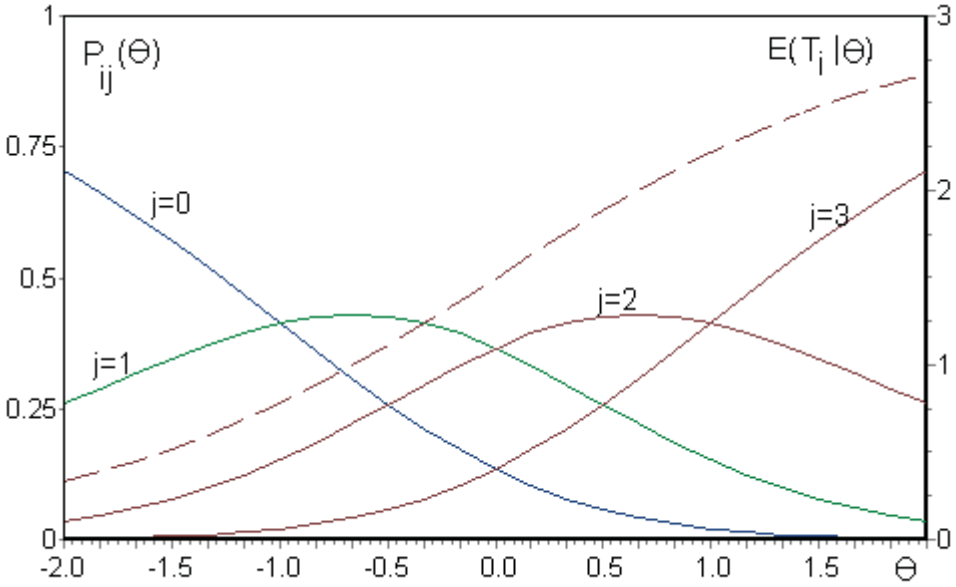

Figure 2.1 Response functions and ITF under the GPCM.

Estimating all the parameters in the GPCM concurrently has both practical and theoretical drawbacks. The practical problem is the sheer amount of parameters (the sample size of the PISA project approaches more than 15 000 students with the analogous number of θ parameters), which renders standard computational methods such as the Newton-Raphson method infeasible. Theoretical problems have to do with the consistency of such concurrent estimates (refer to Haberman, 1977). Depending on the model and the psychometrician's preferences, various alternative estimation methods are available which solve the problem. One of the most used methods, and the method used in the present chapter, is the maximum marginal likelihood (MML, Bock & Aitkin, 1981) estimation method. To apply this method, it is assumed that the θ-parameters have one or more common normal distributions. So we consider populations indexed g ($g = 1,\ldots,$G) and assume that

$$\theta_n \sim N(\mu_{g(n)}, \sigma^2_{g(n)})$$

where *g(n)* is the population to which respondent *n* belongs. Populations may, for instance, be the countries in an educational survey, or gender, or countries crossed with gender, etc. In MML, the likelihood function is marginalized over the θ-parameters, that is, the likelihood

function of all item parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and all means and variances $\boldsymbol{\mu}, \boldsymbol{\sigma}$ given all response patterns $\boldsymbol{x}_n$ ($n=1,\ldots,N$) is given by

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \prod_{n}^{N} \int p(\boldsymbol{x}_n \mid \theta_n, \boldsymbol{\alpha}, \boldsymbol{\beta})\, p(\theta_n; \mu_{g(n)}, \sigma^2_{g(n)}) d\theta_n$$

where $p(\boldsymbol{x}_n \mid \theta_n, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is the probability of response pattern $\boldsymbol{x}_n$ and $p(\theta_n; \mu_{g(n)}, \sigma^2_{g(n)})$ is the normal density related to population *g(n)*.

The likelihood equations are derived by identifying the likelihood equations as if the θ-parameters were observed and then taking the posterior expectation of both sides of the equation with respect to the posterior of the θ-parameters. For instance, if the θ-parameters were observed, the likelihood equation for the mean μ_g would be

$$\mu_g = \sum_{n \mid n(g)=g} \theta_n$$

and after taking posterior expectations of both sides gives the MML estimation equation

$$\mu_g = \sum_{n \mid n(g)=g} E\left(\theta_n \mid \boldsymbol{x}_n; \alpha, \boldsymbol{\beta}, \mu_g, \sigma^2_g\right)$$

$$= \sum_{n \mid n(g)=g} \int \theta_n\, p(\theta_n \mid \boldsymbol{x}_n; \alpha, \boldsymbol{\beta}, \mu_g, \sigma^2_g) d\theta_n. \qquad (2.4)$$

In the next section, the MML framework will be applied to detection and modeling of DIF.

2.3 Detection and Modeling of DIF

Part of the process of establishing the construct validity of a scale may consist of showing that the scale fits a one-dimensional IRT model. This means that the observed responses can be attributed to item and person parameters that are related to some one-dimensional latent dimension. Construct validity implies that the construct to be measured is the same for all respondents. Item bias, or differential item functioning (DIF) violates this assumption. An item displays DIF if the probabilities of responding in different categories vary across sub-populations (say countries or genders) given equivalent levels of the underlying attribute

(Holland & Wainer, 1993; Camilli & Sheppard, 1994). Or, equivalently, an item is biased if the manifest item score, conditional on the latent dimension, differs between sub-populations (Chang & Mazzeo, 1994).

Several techniques for detecting DIF have been proposed. Most of them are based on evaluating differences in response probabilities between groups, conditional on some measure of the latent dimension. The most generally used technique is based on the Mantel Haenszel statistic (Holland & Thayer, 1988), others are based on log linear models (Kok, Mellenbergh & van der Flier, 1985), or on IRT models (Hambleton & Rogers, 1989). The advantage of IRT-based methods over the other two approaches is that IRT offers the possibility of modeling DIF for making inferences about differences regarding the average scale level of sub-populations.

In the present chapter, the logic for the detection of DIF will be based on the logic of the Lagrange multiplier (LM) test (Rao, 1948, Aitchison & Silvey, 1958). Applications of LM tests to the framework of IRT have been described by Glas (1998, 1999), Glas and Falćon (2003) and Glas and Dagohoy (2007). In this chapter, our primary interest is not in the actual outcome of the LM test, because due to the very large sample sizes in educational surveys such as PISA, even the smallest model violation, that is, the smallest amount of DIF, will be significant. The reason for adopting the framework of the LM test is that it clarifies the connection between the model violations and observations and expectations used to detect DIF. Further, it produces comprehensible and well-founded expressions for model expectations, the value of the LM test statistic can be used as an effect size of DIF, and the procedure can be easily generalized to a broad class of IRT models. Before, the general approach to detect DIF is outlined, a special case is presented to clarify the method. Consider two groups labeled the reference group and the focal group. For instance, the reference group may be girls and the focal group may be boys. Define a background variable

$$y_n = \begin{cases} 1 & \text{if person } n \text{ belongs to the focal group,} \\ 0 & \text{if person } n \text{ belongs to the reference group.} \end{cases}$$

For reasons of clarity, the method will be introduced in the framework of the two-parameter logistic model, (the 2PLM) which is the special case of the GPCM pertaining to dichotomously scored items. Consider a model where the probability of a positive response is given by

$$P_i(\theta_n) = \frac{\exp(\alpha_i\theta_n - \beta_i + y_n\delta_i)}{1+\exp(\alpha_i\theta_n - \beta_i + y_n\delta_i)}. \tag{2.5}$$

For the reference population, $y_i = 0$ and the model is analogous to the 2PLM. For the focal population, $y_i = 1$, so in that case the model is also the 2PLM, but the item location parameter β_i is shifted by δ_i.

The LM test targets the null-hypothesis of no DIF, that is, the null-hypothesis $\delta_i = 0$. The LM test statistic is computed using the MML estimates of the null-model, that is, δ_i is not estimated. The test is based on evaluation of the first order derivatives of the marginal likelihood with respect to δ_i evaluated at $\delta_i = 0$ (see Glas, 1999). If the first order derivative in this point is large, the MML estimate of δ_i is far removed from zero, and the test is significant. If the first order derivative in this point is small, the MML estimate of δ_i is probably close to zero and the test is not significant. The actual LM statistic is the squared first order derivative divided by its estimated variance, and it has an asymptotic Chi-square distribution with one degree of freedom. However, as already mentioned above, the primary interest is not so much in the test itself, but on the information it provides regarding the fit between the data and the model. Analogous to the reasoning leading to likelihood equation (4), we first derive the first order derivative assuming that θ_n is observed, and equate it to zero. This results in the likelihood equation

$$\sum_{n=1}^{N} y_n x_{ni} = \sum_{n=1}^{N} y_n P_i(\theta_n).$$

Note that the left-hand side is the number of positive responses given in the focal group and the right-hand side is its expectation if θ_n were observed. Taking expectations with respect to the posterior distribution of θ_n results in

$$\sum_{n=1}^{N} y_n x_{ni} = \sum_{n=1}^{N} y_n E\left(P_i(\theta_n) \mid \boldsymbol{x_n}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_{g(n)}, \sigma^2_{g(n)}\right).$$

So the statistic is based on the difference between the number-correct score in the focal group and its posterior expected value. Note that the difference between the two sides of the likelihood equation can be seen as a residual. Further, if we divide the two sides with the number of respondents in the focal group, that is, with $\Sigma_n y_n$, the expressions become the observed and expected average item score in the focal group. This interpretation provides guidance in judging

the size of the DIF, that is, it provides a framework for judging whether the misfit is substantive or not referenced to the observed score scale.

For a general definition of the approach, which also pertains to polytomously scored items, define covariates y_{nc} ($c = 1,...,C$). Special cases leading to specific DIF statistics will be given below. The covariates may be separately observed person characteristics, but they may also depend on the observed response pattern, but without the response to the item *i* targeted. The probability of a response is given by a generalization of the GPCM, that is,

$$P_{ij}(\theta_n) = \frac{\exp(j\alpha_i\theta_n - \beta_{ij} + j\sum_c y_{nc}\delta_{ic})}{1+\sum_{h=1}^{M_i}\exp(h\alpha_i\theta_n - \beta_{ih} + h\sum_c y_{nc}\delta_{ic})}. \qquad (2.6)$$

For one or more reference populations, the covariates y_{nc} ($c = 1,...,C$) will be equal to zero. These populations serve as a baseline where the GPCM with item parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ holds. In the other populations, one or more covariates y_{nc} are non-zero. The LM statistic for the null-hypothesis $\delta_{ic} = 0$ ($c=1,...,C$) is a quadratic form in the *C*-dimensional vector of first-order derivatives and the inverse of its covariance matrix (for details, see, Glas, 1999). It has an asymptotic Chi-square distribution with *C* degrees of freedom. This general formulation can be translated into many special cases. Three are outlined here and will also be used in the example presented below.

For the first special case, one population serves as the focal population; all other populations serve as reference. The GPCM has only one additional parameter, δ_i, that is, $C = 1$. This leads to the residual

$$r_i = \sum_{n=1}^{N}\sum_{j=1}^{M_i} y_n j X_{ij} - \sum_{n=1}^{N}\sum_{j=1}^{M_i} y_n j E\left(P_{ij}(\theta_n) \mid \boldsymbol{x_n}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_{g(n)}, \sigma^2_{g(n)}\right) \qquad (2.7)$$

Dividing this residual by the number of respondents in the focal group, $\Sigma_n y_n$ produces a residual which is the difference the observed and expected average item-total score in the focal group. The residual gauges so-called uniform DIF, that is, the residual indicates whether the ITF $\Sigma_j jP_{ij}(\theta)$ of the focal group is uniformly shifted over the θ scale relative to the reference group

or not. The associated LM statistic has an asymptotic Chi-square distribution with one degree of freedom.

A second version of the statistic emerges when y_n is a dummy-code for a country. The residuals defined by formula (2.7) then become country-specific, say r_{ic} ($c = 1,\ldots,C$). To assess CDIF, C is equal to the number of countries minus one, because one country must serve as a reference group or base line. The associated LM statistic has an asymptotic Chi-square distribution with degree of freedom equal to the number of countries minus one.

Besides uniform DIF, also non-uniform DIF may occur. In this case, the ITF of focal and reference group may not be just shifted, but they may also cross. That is, in some locations on the θ scale, the ITF of one group is higher, while the reverse is true in other locations. Since θ cannot be directly observed, detection of non-uniform DIF must be based on a proxy for θ. The proxy is a respondent's rest-score, which is the test score on all items except the targeted item i, that is, $\Sigma_n \Sigma_{k \neq i} \Sigma_j jX_{kj}$. The range of these scores is divided into C non-overlapping sub-ranges. Usually, C is between 3 and 6. Residuals are used to evaluate whether the ITF of the focal population and reference populations are different given different rest-scores. So y_{nc} is equal to one if n belongs to the focal population and obtained a rest-score in sub range c ($c = 1,\ldots,C$), and zero otherwise. This leads to the third version of the test based on the residual

$$r_{ic} = \sum_{n=1}^{N}\sum_{j=1}^{M_i} y_{nc}\, jX_{ij} - \sum_{n=1}^{N}\sum_{j=1}^{M_i} y_{nc}\, jE\left(P_{ij}(\theta_n) \mid \boldsymbol{x_n}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mu_{g(n)}, \sigma^2_{g(n)}\right) \qquad (2.8)$$

These are only three examples of the general approach of identifying DIF with (2.6) as an alternative model to the GPCM. The residuals may be based on the frequencies in the response categories rather than on the ITF.

Identification and modeling of DIF is an iterative process where the item with the worst misfit in terms of its value of the LM statistic and its residual is given country-specific item parameters followed by a new concurrent MML estimation procedure and a new DIF analysis. So DIF items are treated one-at-a-time. From a practical point of view, defining country-specific item parameters is equivalent with defining an incomplete design where the DIF item is split into a number of virtual items, where each virtual item is considered as administered in a specific country. The resulting design can be analyzed using IRT software that supports the analysis of

data collected in an incomplete design. Below, items with country-specific parameters will also be referred to as splitted items.

The method is motivated by the assumption that a substantial part of the items function the same in all countries and a limited number of items have CDIF. In the IRT model, it is assumed that all items to pertain to the same latent variable θ. Items without CDIF have the same item parameters in all countries. The items with CDIF have item parameters that are different across countries. This is, these items refer to the same latent variable θ as all the other items, but their location on the scale is different across countries. For instance, the number of cars in the family may be a good indicator of wealth, but the actual number of cars at a certain level of wealth may vary across countries. Or even within countries. Having a car in the inner city of Amsterdam is clearly a sigh of wealth, but the rural eastern part of the Netherlands an equivalent level of wealth will probably result in ownership of three cars.

The number of items given country-specific item parameters is a matter of choice where two considerations are relevant. First, there should remain a sufficient number of anchor items in the scale. Second, the model including the splitted items should fit the data. DIF statistics no longer apply to the splitted items. However, the fit of the item response curve of an individual item, say item *i,* can be evaluated using the test for non-uniform DIF described above, but evaluated using a model including country-specific items parameters. So also in this application, ranges of the rest-score are used as proxies for locations on the θ scale, and the test evaluates whether the model with the country-specific items parameters can properly predict the ITF.

2.4 Examples

Two examples will be given. The objective of the first one is to give the flavor of the model, the second one is meant to show how the approach works in a large-scale international survey. Starting with the first example, the data are taken from the field-trial of the 2009 cycle of PISA. The data emerged from 20 countries and the total sample size was 9522 students. The scale analyzed was “Online Reading Activities” and consisted of 11 items, all scored 0 to 4. Table 2.1 shows results of an analysis where one of the countries served as a focal group, while the rest of the countries served as a reference group. Using the GPCM, concurrent MML estimates were obtained for all item parameters and using a separate population distribution for each

country. The focal group consisted of 586 students; the reference group consisted of 8936 students.

The column labeled "LM" gives the values of the LM statistics based on the residuals r_i defined by formula (2.7). In this case, the LM has one degree of freedom. The significance probabilities are not given: as expected, all tests were significant due to the sample sizes. However, the values indicate that item 8 had the largest misfit in this country. The following four columns give the observed and expected values on which the test is based, for the focal and reference group, respectively. The values are average item scores. It can be seen that for item 1, the observed average in the focal group was 2.88, while the expected value was 2.94. So the focal group scores lower than expected. Since the observed score ranges from 0 and 4, the difference is quite small. Note further that the difference for the reference group is 0.01, which is very small. This is, however, due to the fact that the reference group was much larger and put has far more weight in the estimate of the item parameters. The last column gives the value of r_i as defined by formula (2.7). Again, it can be concluded that item 8 had the worst fit: the focal group scored far too low.

Table 2.1.

Tests for differential item functioning targeted at items within a country

		Focal Group		*Reference Group*		
Item	LM	Obs	Exp	Obs	Exp	r_i
1	39.5	2.88	2.94	2.44	2.43	-0.06
2	86.3	3.57	3.33	2.78	2.79	0.24
3	49.6	2.78	2.59	2.09	2.10	0.19
4	54.2	2.54	2.80	2.39	2.38	-0.26
5	42.6	1.27	1.45	1.30	1.29	-0.18
6	21.7	2.42	2.34	1.97	1.97	0.08
7	14.3	2.70	2.73	2.33	2.33	-0.03
8	136.2	2.80	3.02	2.77	2.76	-0.22
9	3.4	2.04	2.05	1.66	1.66	-0.01
10	62.3	1.17	1.37	1.24	1.23	-0.20
11	31.4	2.39	2.24	1.89	1.90	0.15

Besides information about the interaction between items and countries, also an overall assessment of DIF is of interest. Table 2.2 presents such information. This information is

obtained in the same MML estimation run as used for Table 2.1. The second and third column, labeled LM and Av. Dif, give information aggregated across countries. The LM statistic has 19 degrees of freedom. Again, significance probabilities are not given: all tests were significant due to the large sample size. Further, again item 8 has the largest misfit. The column labeled "Av. Dif" gives an effect size of the DIF aggregated across countries: it is the mean over the countries of the absolute residuals, that is, the absolute differences between observed and expected as defined in formula (2.6).

Table 2.2.

Tests for differential item functioning targeted at items within a country

	No Item Splitted		2 Items Splitted		4 Items Splitted	
Item	LM	Av. Dif	LM	Av. Dif	LM	Av. Dif
1	1107.7	0.20	915.4	0.19		
2	831.0	0.21	581.3	0.21	606.0	0.21
3	664.4	0.18	589.9	0.18	475.8	0.16
4	779.9	0.19	679.9	0.20		
5	1541.2	0.25				
6	414.7	0.14	330.7	0.14	271.5	0.12
7	520.9	0.13	402.9	0.14	355.9	0.12
8	1672.7	0.42				
9	422.1	0.16	396.1	0.16	380.1	0.15
10	384.0	0.10	354.4	0.11	366.1	0.11
11	314.9	0.11	250.7	0.10	232.2	0.10

Next, in an iterative process of splitting items into virtual items, MML estimation and evaluation of LM tests, the items 8, 5, 1 and 4 were splitted, in that order. The columns labeled "2 Items Splitted" give the values after splitting items 8 and 5, the columns labeled "4 Items Splitted" give the values after all four items were splitted. Note that the first analysis does not always determine the order in which the items are splitted: item 2 seems to have more bias than item 4 at first, but their order is reversed in the process. The reason is that the presence of DIF items can also bias the estimates of the parameters of item which are not biased.

What is also needed to justify the procedure is evidence that the complete concurrent model including the link items and the splitted items fits the data for every country. Information that can contribute to such evidence is given in Table 2.3 for the same country as used for Table 2.1

and Table 2.2. The table gives information regarding the fit of the ITF within a country after items are split. For every item, the rest-score range is divided into three sub-ranges and the observed and expected average item scores in the thus formed sub-groups of students are given. The last column gives the means over these subgroups of the residuals defined in formula (2.8), that is, of the absolute difference between observed and expected values in sub-groups. The splitted items are marked with an asterisk. It can be seen that the splitted items fitted the model well. For the items which were not splitted, the table gives information regarding non-uniform DIF. The reason is that the expected values are computed using the assumption that the same item parameters apply in all countries, while the observations may reveal differences in the regression of the item scores on the rest-scores.

Table 2.3. Fit of the ITF within a country

			Group 1		Group 2		Group 3		
Item	LM	Prob	Obs	Exp	Obs	Exp	Obs	Exp	Av. Dif
1*	0.1	0.94	2.43	2.41	2.90	2.89	3.27	3.28	0.01
2	29.8	0.00	3.19	2.95	3.67	3.47	3.79	3.72	0.17
3	5.3	0.07	2.14	2.11	2.82	2.70	3.29	3.20	0.08
4*	1.4	0.48	1.95	1.92	2.45	2.51	3.08	3.05	0.04
5*	1.1	0.56	1.11	1.07	1.26	1.22	1.44	1.47	0.03
6	3.8	0.14	1.85	1.96	2.40	2.38	2.86	2.79	0.07
7	19.7	0.00	2.16	2.36	2.70	2.80	3.16	3.17	0.10
8*	0.4	0.82	2.54	2.50	2.77	2.78	3.04	3.03	0.02
9	7.1	0.03	1.42	1.58	2.04	2.11	2.59	2.69	0.11
10	61.2	0.00	1.01	1.20	1.14	1.37	1.35	1.63	0.23
11	5.0	0.08	1.95	1.91	2.44	2.32	2.82	2.74	0.08

The LM statistics have two degrees of freedom. The sample sizes within the country (586 students) are now such that significance probabilities of the LM tests become informative. Item 2 and 10 show the largest misfit. Consistent with the results in Table 2.1, the ITF of item 2 is too high, while the ITF of item 10 is too low. This is an indication of uniform rather than non-uniform DIF. So it might be worthwhile to also split these items. On the other hand, the link must also remain substantial. There is some tradeoff between these two considerations and some element of arbitrariness cannot be avoided.

The second example pertains to the main study of the 2009 cycle of PISA. The data consisted of samples of 500 students from 31 OECD countries. The analyses consisted of two steps. First the data of all countries were analyzed simultaneously to identify items with country-specific DIF. This was done in an iterative process. In each iteration, MML estimates were obtained and the item with the worst misfit was identified. In the next iteration, this item was given country-specific item parameters and a new MML estimation run was made. This was repeated between two and four times depending on the scale analyzed. Finally, the fit of the resulting model with country-specific item parameters for the DIF items and the parameters of the non-DIF items, which were fixed over the countries, was evaluated. In the second step, the impact of DIF was evaluated by computing the correlations of the countries' mean latent trait values estimated without and with country-specific item parameters. Analyses were done using the PCM and the GPCM. Finally, to evaluate the impact of the choice of the model, the correlations of the countries' mean latent trait values estimated using the PCM and GPCM were computed. The reason is that the PISA project uses the PCM, so as a side line we will make an unassuming comparison between the results obtained using these two models.

Table 2.4 gives the codes and names of the scales which were investigated and the number of items in each scale. Labels starting with ST refer to scales from the student questionnaire and labels starting with IC refer to scales from the ICT questionnaire. To compute the results in table 2.4, MML analyses using the GPCM were made for every scale with all available OECD countries entered in an analysis simultaneously. The number of countries was 31 for the student questionnaire and 26 for the ICT questionnaire. Absolute values of the residuals as defined in formula (2.7) were counted and the percentages of values above 0.25 and 0.20 are displayed in the two last columns of table 2.4, respectively. Note that the scales ST25 ("Like Reading") and IC04 ("Home Usage of ICT") displayed the most DIF. The scales ST27(a) and ST27(b), ST34, ST36, IC02, IC05, IC8 and IC10 were relatively free of DIF.

Table 2.4. Overview of CDIF in the student questionnaire and the ICT questionnaire

Label	Scale	Number of Items	Percentage Item by Country Interaction Residual > 0.25	Residual > 0.20
ST24	Reading Attitude	11	7	12
ST25	Like Reading	5	60	66
ST26	Online Reading Activities	7	18	22
ST27(a)	Use of Control Strategies	4	6	7
ST27(b)	Use of Elaboration strategies	4	1	3
ST27(c)	Use of Memorisation strategies	5	12	16
ST34	Classroom Climate	5	2	4
ST36	Disciplinary Climate	5	2	3
ST37	Stimulate Reading Engagement	7	6	10
ST38	Teacher Structuring Strategies	9	10	12
ST39	Use of Libraries	7	15	22
IC02	ICT availability at school	5	3	4
IC04	Home Usage of ICT	8	24	30
IC05	ICT for School Related Tasks	5	9	14
IC06	Use of ICT at School	9	11	18
IC08	ICT Competency in Different Contexts	5	7	9
IC10	Attitude Towards Computers	4	3	5

In Table 2.5, the results are further broken down to the item level. Items with effect sizes above 0.20 are highlighted. The items causing the DIF can be easily identified. Further breaking down these residuals can lead to interesting insights. It is beyond the scope of this chapter to discuss all item-by-county interactions in detail, so one example must do. As already mentioned, ST25 has the largest bias. ST25 consists of the stem overall question "How often do you read these materials because you want to?" followed by the items "Magazines", "Comic books", "Fiction (novels, narratives)", "Non-fiction books", and "Newspapers". Response categories indexed from 0 to 4 are "Never or almost never", "A few times a year", "About once a month", "Several times a month", and "Several times a week". It turns out that in Finland reading of comic books is much more salient than in other countries. The average observed and expected score over all countries except Finland is 1.25. The average item score in Finland is 2.58, compared to an expected value of 1.78, resulting in a residual of 0.87. The conclusion is that the Finnish students like to read more than the average OECD student, but they are especially fond of comic books. Giving the item regarding comic books country-specific item parameters solved the problem for Finland in the sense that the absolute values of all residuals as defined by formula (2.7) dropped below 0.10.

Table 2.5. Size of residuals on the item level

Scale	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
ST24	0.10	0.07	0.07	0.13	0.09	0.06	0.09	0.11	0.15	0.16	0.12
ST25	0.24	**0.41**	0.20	0.07	**0.35**						
ST26	0.20	0.20	0.19	0.09	0.13	0.15	0.11				
ST27(E)	0.08	0.06	0.08	0.05							
ST27(M)	0.10	0.12	0.17	0.17							
ST27(C)	0.15	0.07	0.06	0.09	0.10						
ST34	0.06	0.09	0.04	0.06	0.08						
ST36	0.07	0.07	0.04	0.09	0.06						
ST37	0.10	0.09	0.09	0.17	0.08	0.10	0.07				
ST38	0.16	0.08	0.14	0.13	0.12	0.16	0.07	0.07	0.13		
ST39	0.14	0.17	0.13	0.10	0.09	0.06	0.16				
IC01	0.13	0.16	0.04	0.18	0.04	0.05	0.06	0.10			
IC02	0.07	0.13	0.02	0.10	0.13						
IC04	0.13	0.16	0.15	0.11	0.09	0.15	0.23	**0.46**			
IC05	0.12	0.20	0.10	0.08	0.13						
IC06	0.15	0.16	0.11	0.07	0.12	0.06	0.13	0.15	0.08		
IC08	0.09	0.19	0.15	0.11	0.05						
IC10	0.09	0.12	0.08	0.14							

The impact of DIF was assessed using both the PCM and GPCM. The countries were rank ordered on their mean value on the latent variable, for both models and without and with items with country-specific parameters. An example pertaining to scale ST26 is given in Table 2.6. The first two items of the scale were splitted. The values in the four last columns are the MML estimated means of the latent trait distributions. Note that at first sight, the rank order of the countries looks quite stable.The Table 2.7 gives the correlations between the estimates obtained using 2 or 4 splitted items. The iterative process of splitting items was stopped when either 4 items were splitted, or 95% of the residuals defined by formula (8) were under 0.25. The first two columns of Tables 2.7 give the rank correlation between order of countries obtained by without and with splitted items using either the PCM or the GPCM as the measurement models. The last two columns present the analogous correlations between the country means. It can be seen that many of the correlations between the country means are quite high except for the scales ‘Like Reading’, ‘Online reading strategies’, and ’Use of memorization strategies’ , which had substantial DIF. Also ‘Use of Libraries’ seems affected by DIF. There are no clear differences between the correlations obtained using the PCM and GPCM.

Table 2.6. Rank order and mean scale level of countries on the scale ST26 for the GPCM and PCM with and without splitted items.

	Rank Order				Mean on Latent Scale			
Country	PCM With	GPCM With	PCM Without	GPCM Without	PCM With	GPCM With	PCM Without	GPCM Without
AUS	8	11	9	10	-0.131	-0.074	-0.109	-0.076
AUT	20	21	21	21	0.067	0.069	0.069	0.066
BEL	4	7	3	5	-0.281	-0.128	-0.260	-0.165
CAN	6	9	6	7	-0.203	-0.103	-0.164	-0.106
CZE	30	30	30	30	0.605	0.546	0.486	0.472
DNK	24	25	24	26	0.149	0.211	0.142	0.183
FIN	11	13	7	9	-0.105	-0.059	-0.134	-0.095
FRA	7	10	8	8	-0.153	-0.094	-0.130	-0.099
DEU	23	23	22	24	0.143	0.167	0.121	0.136
GRC	21	17	20	18	0.067	-0.018	0.054	0.003
HUN	29	29	29	29	0.455	0.437	0.326	0.339
ISL	28	28	25	25	0.237	0.251	0.149	0.173
IRL	2	2	2	2	-0.576	-0.539	-0.486	-0.480
ITA	17	14	17	15	-0.024	-0.053	-0.004	-0.032
JPN	1	1	1	1	-0.673	-0.618	-0.509	-0.500
KOR	25	5	26	14	0.167	0.168	0.158	-0.046
LUX	16	19	15	17	-0.035	0.002	-0.036	-0.017
MEX	3	3	4	3	-0.376	-0.491	-0.248	-0.350
NLD	14	24	13	22	-0.060	0.176	-0.058	0.089
NZL	5	4	5	4	-0.234	-0.267	-0.196	-0.222
NOR	27	27	27	27	0.217	0.223	0.171	0.187
POL	31	31	31	31	0.744	0.562	0.624	0.528
PRT	26	26	28	28	0.217	0.215	0.184	0.190
SVK	13	15	11	11	-0.072	-0.040	-0.098	-0.074
ESP	10	12	12	13	-0.109	-0.073	-0.066	-0.049
SWE	18	20	18	19	0.040	0.024	0.027	0.023
CHE	15	18	14	16	-0.058	-0.007	-0.055	-0.023
TUR	22	16	19	23	0.132	-0.029	0.134	0.026
QUK	19	22	19	23	0.067	0.138	0.049	0.098
USA	9	8	10	6	-0.119	-0.126	-0.104	-0.107
CHL	12	6	16	12	-0.099	-0.135	-0.036	-0.072

The impact of using either the PCM or the GPCM was further evaluated by assessing differences in the estimated means of the countries on the latent scale and also the rank ordering obtained using the two models. These results are given in Table 2.8. The last two columns give the rank correlation and product moment correlation of the latent-scale means of countries obtained using PCM and GPCM when no items were splitted. The two previous columns give the analogous correlations for the number of splitted items

Table 2.7. Correlations between country means of latent distributions estimated with and without splitted items

		Items	Rank Correlation		Correlation	
Label	Scale	Split	PCM	GPCM	PCM	GPCM
ST24	Reading Attitude	2	0.847	0.964	0.978	0.991
ST25	Like Reading	2	0.589	0.861	0.610	0.968
ST26	Online Reading Activities	2	0.616	0.819	0.936	0.962
ST27(a)	Use of Control Strategies	2	0.646	0.706	0.914	0.934
ST27(b)	Use of Elaboration strategies	2	0.838	0.919	0.969	0.973
St27(c)	Use of Memorization strategies	2	0.510	0.616	0.784	0.922
ST34	Classroom Climate	2	0.870	0.870	0.973	0.967
ST36	Disciplinary Climate	2	0.885	0.906	0.979	0.979
ST37	Stimulate Reading Engagement	2	0.933	0.966	0.982	0.991
ST38	Teacher Structuring Strategies	2	0.951	0.958	0.979	0.989
ST39	Use of Libraries	2	0.883	0.880	0.954	0.954
IC02	ICT availability at school	2	0.851	0.823	0.923	0.901
IC04	Home Usage of ICT	2	0.876	0.894	0.980	0.981
IC05	ICT for School Related Tasks	2	0.850	0.844	0.969	0.969
IC06	Use of ICT at School	2	0.969	0.969	0.995	0.995
IC08	ICT Competency	2	0.829	0.822	0.959	0.953
IC10	Attitude Towards Computers	2	0.801	0.743	0.985	0.960
ST24	Reading Attitude	4	0.804	0.919	0.996	0.984
ST26	Online Reading Activities	4	0.606	0.798	0.857	0.905
ST37	Stimulate Reading Engagement	4	0.767	0.829	0.922	0.996
ST38	Teacher Structuring Strategies	4	0.888	0.889	0.956	0.966
ST39	Use of libraries	4	0.788	0.853	0.927	0.945
IC04	Home Usage of ICT	4	0.879	0.894	0.980	0.981
IC06	Use of ICT at School	4	0.976	0.920	0.995	0.862

given in the column labeled “Item Split”. In general the correlations are high. The main exception is ST25. Therefore, given our criteria for comparing model fit, it can be concluded that there is little support for preferring the GPCM over the PCM as an analysis model.

Table 2.8. Correlations between country means of latent distributions estimated using the PCM and GPCM

Label	Scale	Items Split	With Splitted Items		Without Splitted Items	
			Rank Correlation	Correlation	Rank Correlation	Correlation
ST24	Reading Attitude	2	0.940	0.993	0.913	0.988
ST25	Like Reading	2	0.643	0.897	0.574	0.666
ST26	Online Reading Activities	2	0.962	0.994	0.879	0.988
ST27(a)	Use of Control Strategies	2	0.960	0.993	0.959	0.992
ST27(b)	Use of Elaboration strategies	2	0.954	0.994	0.968	0.997
St27(c)	Use of Memorization strategies	2	0.988	0.998	0.805	0.966
ST34	Classroom Climate	2	0.983	0.996	0.987	0.998
ST36	Disciplinary Climate	2	0.976	0.997	0.986	0.997
ST37	Stimulate Reading Engagement	2	0.993	0.998	0.981	0.996
ST38	Teacher Structuring Strategies	2	0.977	0.998	0.978	0.996
ST39	Use of Libraries	2	0.981	0.993	0.990	0.998
IC02	ICT availability at school	2	0.998	0.997	0.968	0.987
IC04	Home Usage of ICT	2	0.959	0.980	0.941	0.978
IC05	ICT for School Related Tasks	2	0.974	0.993	0.980	0.996
IC06	Use of ICT at School	2	0.992	0.998	0.994	0.998
IC08	ICT Competency	2	0.942	0.990	0.972	0.994
IC10	Attitude Towards Computers	2	1.000	0.983	0.980	0.997
ST24	Reading Attitude	4	0.968	0.995		
ST26	Online Reading Activities	4	0.964	0.996		
ST37	Stimulate Reading Engagement	4	0.978	0.994		
ST38	Teacher Structuring Strategies	4	0.985	0.997		
ST39	Use of libraries	4	0.972	0.988		
IC04	Home Usage of ICT	4	0.959	0.980		
IC06	Use of ICT at School	4	0.936	0.880		

2.5 Conclusions

Large-scale educational surveys often give rise to an overwhelming amount of data. Simple unequivocal statistical methods for assessing the quality and structure of the data are hard to design. The present chapter presents diagnostic tools to tackle at least one of the problems which emerge in educational surveys, the problem of differential item functioning. Given the complicated and large data, it comes as no surprise that the tools presented here have both advantages and drawbacks. On the credit side, concurrent MML estimation is well founded, practical and quick. Further, in combination with LM statistics, few analyses are needed to gain

insight in the data. Above, searching for DIF was presented as an iterative procedure, but this procedure can be easily implemented as one automated procedure. On the other hand, the advantage that, contrary to most test statistics for IRT, the LM statistics have a known asymptotic distribution losses much of its impact, because of the power problem in large data sets. What remains is a procedure which is transparent with respect to which model violations are exactly targeted and the importance of the model violation in terms of the actual observations. Further, the procedure is not confined to specific IRT models, but can be generally applied. Finally, the procedure supports the use of group-specific item parameters. The decision of whether group-specific item parameters should actually be used depends on the inferences that are to be made next. In that sense, the example where the order of countries on a latent scale is evaluated is just an example. Often, other inferences are made using the outcomes of the IRT analyses, such as multilevel analyses relating background variables to educational outcomes. Also in these cases, the impact of the using country-specific item parameters can be assessed by comparing different analyses.

The present chapter was mainly written to present statistical methodology and not to draw ultimate conclusions regarding the PISA project. Still, some preliminary conclusions can be drawn. The analyses showed that certain scales of the student background questionnaire and the ICT questionnaire are indeed affected by the presence of CDIF. The scale most affected by CDIF was ST25 'Like Reading'. Other scales where DIF was evident were ST26 'Online Reading activities', ST27c 'Memorization strategies', ST39 'Use of libraries' and IC04 'Home use of ICT'. Correlations between ordering of countries showed that the detected CDIF did indeed have an impact. However, other criteria for impact may be more relevant.

Finally, using either the PCM or GPCM had little impact. Overall, the discrimination parameters were quite high and differences between these indices within the scales probably cancelled when evaluating the order of the countries. Also the conclusions regarding CDIF items were not substantially affected by the model used.

Chapter 3

Methodological Issues of the PISA Scaling Model: Comments on Kreiner & Christensen 2014.

This article is a comment on the article by Svend Kreiner and Karl Bang Christensen (K&C) titled 'Analysis of model fit and robustness, a new look at the PISA scaling model underlying ranking of countries according to reading literacy' (Kreiner & Christensen, 2014). In their article, the authors examine methodological issues concerning the scaling model used for the PISA 2006 reading assessment with specific reference to whether PISA's ranking of countries is flawed due to model misfit and, particularly, by country-specific differential item functioning (CDIF). According to K&C, their analysis provides strong evidence of misfit of the PISA scaling model and especially very strong evidence of CDIF. Based on these findings they assert that the country rankings reported by PISA are not reliable. In the present article the methodological approach to scaling the data as utilized in the PISA project is outlined and an argument is made that the investigations by K&C ignore or misrepresent some important methodological choices made in the PISA approach. It is further shown that the findings by K&C are based on a very limited subset of the data and that their findings do not generalize to the PISA data at large. More specifically, K&C's main criticism concerning the impact of CDIF on the ranking of countries in PISA 2006 is investigated. K&C came to their conclusions based on analysis conducted on data from only one booklet of PISA 2006. According to them, their results can be extrapolated to the entire PISA 2006 reading dataset. This assertion is tested by modeling CDIF using data from other samples from the PISA 2006 data set including the sampled dataset used to calibrate the 2006 reading test. Results show that the impact of CDIF on the ranking of countries is far less prominent than suggested and becomes almost negligible when the statistical uncertainty regarding the country means is properly taken into account. The article ends with the conclusion that the K&C critique is both inappropriate and biased.

3.1 Critique of PISA

K&C studied the fit of the Rasch model (Rasch, 1980) which is the basis of the analytic model of PISA (Adams & Wu, 2007), to the PISA 2006 reading data. They lay particular emphasis on the impact of CDIF on the rank ordering of countries on the PISA 2006 reading test. Their conclusions are that the Rasch model does not fit the PISA data and that results are particularly distorted by the presence of a substantial number of items with CDIF in the reading scale.

According to Raymond Adams, who was head of the PISA consortium from 2000 till 2012 (http://www.oecd.org/pisa/47681954.pdf), the K&C line of argument concerning the use of the Rasch model is strongly based upon tests of statistical significance rather than the substance of the effects detected. As Box (1979) reminds us no statistical model will fit data perfectly, but some statistical models are useful. It has also been shown repeatedly in the past that a rejection of a model based on statistically significant results may be less than meaningful (Berkson, 1942; Gardner & Altman, 1986). For a result that directly applies to the Rasch model, Molenaar (1997) has shown that the rejection of the Rasch model and use of more general IRT models instead may often have surprisingly little impact. The majority of K&C's findings could be summarized with the simple observation that PISA has a large sample, and hence, most model assumptions will be rejected based on a very large power of almost any statistical test that is conducted. The sample sizes in PISA are such that the fit of any scaling model, particularly a simple model like the Rasch model, will almost surely be rejected. PISA has taken the view that it is unreasonable to adopt a slavish devotion to tests of statistical significance (e.g. Gardner, & Altman, 1986) concerning fit to a scaling model. The more fundamental question is whether the scaling approach that has been adopted in PISA is useful. There is nothing in K&C's paper that speaks to the issue of the utility of the scaling approach that has been used or the implications of it use.

K&Cs second argument about the misfit of the Rasch model is about the strong presence of CDIF and local dependence. As an alternative to the Rasch model employed in PISA, K&C first pose a more general Rasch-type model (GLLRM) that permits dependence and CDIF terms. At this point K&C indicate with examples from five countries that adding the CDIF term has substantial impact on the rankings but further adding the local independence term to the CDIF term hardly has any impact (see Table 7 of K&C psychometrika article). However K&C still reach the conclusion that it will not be feasible to apply such an approach to PISA: "…despite the complexity of these models, item fit statistics also reject them ". PISA has also explored

alternative models, the two-parameter model has been applied to PISA data and it has been shown that the outcomes are identical to those obtained when fitting a Rasch model (Macaskill, 2008). Further, the dependency between PISA items that Kreiner mentions has also been modeled, and no implications for the rankings have been observed (Macaskill, 2008). As far as the impact of CDIF is on the rankings is concerned, in the following section a comparative study is done about the impact of CDIF on the constrained sample used by K&C and samples which are representative of the entire PISA dataset. The results show that the impact on CDIF is highly inflated when using data with a limited set of items like K&C have done. Furthermore if the effects of the uncertainty due to measurement error regarding the country means are taken into account, the impact of CDIF becomes negligible.

3.2 Further Analyses of the Country Rankings

K&C rank order the country means using the responses to 20 out of the 28 items in booklet 6. They exclude 8 items because their statistical methodology cannot accommodate to the missing responses to these items. In the present section, their findings will be replicated and an investigation will be done whether the findings generalize to the same booklet 6 including all 28 items, as well as to the entire data set, and to the reading data set obtained in the 2009 PISA cycle. Further, the effects of the uncertainty due to measurement error regarding the country means will be assessed because as far as the change in rankings is concerned K&C undertake no technically appropriate statistical testing of the differences in the ranking.

The analyses started by obtaining concurrent estimates of the item parameters and the means and standard deviations of the ability distributions of all countries under the Rasch model. Estimates were generated using the marginal maximum likelihood approach (MML, Adams & Wu, 2007, Adams, Wu, Carstensen, 2007). CDIF was evaluated using Lagrange Multiplier tests (Glas, 1999). CDIF was modeled by the introduction of country-specific item parameters for the item with the highest value of the LM statistic and by rerunning the concurrent MML estimation procedure. This process was repeated 8 times, splitting up one item at a time. After 8 cycles, the effects on the ordering of countries became negligible. For a detailed motivation and description of the procedure, refer to Glas and Verhelst (1995). All computations were made using public domain software (Glas, 2010).

Table 3.1. Average total and DIF equated scores and rankings of countries for booklet 6 data on ***all the 28*** reading items from the PISA 2006 reading test

	Model						
	Normal Items		8 Virtual Items		Normal Items	Virtual Items	Difference Rank
Country	Mean	s.d	Mean	s.d	Rank	Rank	
ARG	-.52	.05	-.59	.05	48	49	1
AUS	.72	.04	.83	.04	11	10	1
AUT	.56	.06	.71	.06	19	18	1
AZE	-1.15	.08	-.95	.08	54	53	1
BEL	.72	.05	.86	.05	8	7	1
BGR	-.40	.05	-.40	.05	46	45	1
BRA	-.67	.06	-.77	.06	49	53	4
CAN	.79	.03	.85	.03	5	9	4
CHE	.52	.04	.64	.04	22	22	0
CHL	.17	.06	.07	.06	37	40	3
COL	-.36	.06	-.48	.06	44	47	3
CZE	.72	.07	.79	.07	9	12	3
DEU	.65	.07	.79	.07	13	11	2
DNK	.64	.05	.90	.05	14	6	8
ESP	.37	.03	.49	.03	28	27	1
EST	.76	.06	.86	.06	7	8	1
FIN	1.23	.06	1.34	.06	2	2	0
FRA	.59	.07	.72	.07	18	17	1
GBR	.60	.04	.69	.04	16	19	3
GRC	.31	.07	.30	.07	33	35	2
HKG	.92	.06	1.14	.06	3	3	0
HRV	.35	.03	.37	.03	31	33	2
HUN	.47	.06	.56	.06	24	25	1
IDN	-.74	.08	-.75	.08	52	52	0
IRL	.72	.06	.78	.06	10	13	3
ISL	.44	.07	.58	.07	25	24	1
ISR	-.05	.05	.08	.05	41	39	2
ITA	.35	.03	.40	.03	29	32	3
JOR	-.39	.06	-.38	.06	45	44	1
JPN	.55	.06	.77	.06	21	14	7
KGZ	-1.75	.30	-1.80	.30	56	56	0
KOR	1.27	.06	1.39	.06	1	1	0
LIE	.27	.06	.45	.06	34	30	4
LTU	.24	.06	.27	.06	35	36	1
LUX	.37	.07	.50	.07	27	26	1
LVA	.56	.05	.61	.05	20	23	3
MAC	.50	.06	.65	.06	23	21	2
MEX	-.15	.02	-.14	.02	42	42	0
MNE	-.76	.07	-.72	.07	53	51	2
NLD	.78	.06	.97	.06	6	4	2
NOR	.23	.07	.47	.07	36	29	7
NZL	.83	.07	.94	.07	4	5	1
POL	.70	.06	.76	.06	12	15	3
PRT	.35	.06	.31	.06	30	34	4
QAT	-1.71	.06	-1.62	.06	55	55	0
ROU	-.70	.05	-.70	.05	51	50	1
RUS	.13	.06	.13	.06	38	38	0
SRB	-.47	.05	-.46	.05	47	46	1
SVK	.42	.07	.43	.07	26	31	5
SVN	.34	.04	.48	.04	32	28	4
SWE	.60	.07	.75	.07	17	16	1
TAP	.63	.05	.67	.05	15	20	5
THA	-.24	.06	-.20	.06	43	43	0
TUN	-.70	.07	-.58	.07	50	48	2
TUR	.05	.06	.20	.06	39	37	2
URY	.00	.00	.00	.00	40	41	1
Average	.19	.06	.27	.06			2.04

As a first step, the results by K&C were replicated. The average number of places a country's ranking shifted after modeling 8 CDIF items out of the 20 selected items for scaling was 3.57 places. This confirms the published K&C shift of 3.5 places. Next, the Booklet 6 data is further investigated by including all the 28 reading items present in the booklet. As a result, the average shift dropped to 2.04.The results are presented in Table 3.1.

From Table 3.2 it can be seen that the average number of places a country's ranking shifts after modeling 8 DIF items out of the all the available 28 items in booklet 6 is 2.04 places. This is already much less than the shift of 3.57 encountered when scaling only a segment of booklet 6 which Kreiner did. This effect of a reduction in the ranking shifts when the set of items is increased is consistent what results reported by Weeks et al. (2013), and it is what should be expected when the set of items covering the assessment framework becomes more comprehensive. This theory is tested further by scaling the dataset of each booklet separately to see if there is a lot of DIF in individual booklets. There are 13 booklets in PISA 2006. However only 7 booklets out of 13 contained reading items. Besides booklet 6, the other booklets are booklet 2,7,9,11,12 and 13. Booklet 6 was unique in the sense that it contained all 28 items. The other six booklets contain only 14 reading items each. Booklets 2, 7 and 12 (which will be called set A booklets) contain the same set of 14 reading items and booklets 9,11 and 13 (which will be called set B booklets) contain the other set of 14 reading items. Together these 6 booklets covered the entire range of 28 reading items in PISA 2006. DIF is modeled in each of these 6 booklets. Table 3.2 gives a summary of the number of average number of places a country shifts after modeling 6 out of the 14 items with largest DIF. From Table 3.2 it can be seen that all the booklets have a sizeable average shift of between 2.79 and 3.39.

Table 3.2: Average shift in country rankings per booklet after modeling 6 DIF items

	Booklet 2	Booklet 7	Booklet 9	Booklet 11	Booklet 12	Booklet 13
Average shift in country rankings	2.86	3.04	2.79	2.86	3.39	3.25

However all these booklets contain only half the total set of reading test items. To check if administering only a subset of items conveys an advantage or disadvantage to certain countries which subsequently results in DIF, further analyses are done where first the 3 booklets from the same set of booklets are combined (which cover only half the total number of reading items)

and then booklets from set A are cross combined with booklets from set B (which cover the entire set of reading items). The results are presented in Table 3.3.

Table 3.3: Average shift in country rankings per booklet after modeling 6 DIF items for booklets combined from the same booklet set (set A or set B)

	Booklets 9&11&13	Booklets 2&7&12	Booklets 7&12	Booklets 9&13	Booklets 2&12	Booklets 9&11
Average shift in country rankings	2.79	2.89	3.21	2.71	2.79	3.00

Table 3.4: Average shift in country rankings per booklet after modeling 6 DIF items for booklets combined from the different booklet sets (set A & set B)

	Booklets 2&13	Booklets 7&11	Booklets 9&12	Booklets 2&9	Booklets 7&13	Booklets 11&12
Average shift in country rankings	1.54	1.14	1.68	1.46	1.36	1.82

.

From the results in Tables 3.3 and 3.4, it becomes clear that administering the entire set of 28 reading items to various combinations of student populations, results in lesser DIF than when administering only a subset of (14) items to different combinations of student populations.

However, in PISA 2006 sampling was done from the complete booklet set to calibrate the reading items. The calibration process employed by PISA 2006 was replicated to check the impact of CDIF on the PISA 2006 results. PISA calibrated the cognitive reading tests like Reading, Mathematics and Science in 2006 using randomly sampled data of 500 students from the OECD countries only. The resulting sample size was 15,000 students. The results of the reanalysis of these data are presented in Table 3.5. The columns under the heading *Original Rasch Model* give the means of the ability distributions as well as the accompanying standard errors and the rankings obtained using the Rasch model. The columns under the heading *CDIF model* give analogous information after the introduction of country-specific parameters to 8 items. The last column provides the differences in rankings. The average difference was equal to 1.24.

Table 3.5
Rankings of countries for the randomly sampled OECD countries based sample of 15,000 students (500 students from each of the 30 OECD countries) used by PISA for item calibration of the PISA 2006 reading test. Average change of rank is 1.24.

	Original Rasch Model			CDIF Model			Change
Country	Mean	s.e.	Rank	Mean	s.e.	Rank	Rank
AUS	-.07	.11	11	-.05	.11	10	1
AUT	-.25	.11	19	-.26	.11	19	0
BEL	-.09	.12	13	-.13	.12	15	2
CAN	.11	.12	4	.13	.12	4	0
CHE	-.22	.11	16	-.23	.11	18	2
CZE	.00	.00	9	.00	.00	7	2
DEU	-.37	.13	25	-.42	.13	24	1
DNK	-.01	.11	10	-.01	.11	9	1
ESP	-.35	.11	21	-.29	.11	20	1
FIN	.39	.11	2	.49	.11	1	1
FRA	-.22	.12	17	-.13	.12	16	1
GBR	-.09	.12	14	-.12	.12	13	1
GRC	-.68	.12	27	-.93	.12	27	0
HUN	-.36	.11	22	-.37	.11	22	0
IRL	.09	.10	5	.06	.10	5	0
ISL	-.13	.12	15	-.09	.12	12	3
ITA	-.36	.12	24	-.43	.12	25	1
JPN	.04	.11	7	.05	.11	6	1
KOR	.52	.12	1	.42	.12	2	1
LUX	-.36	.13	23	-.40	.13	23	0
MEX	-.87	.10	29	-.98	.10	29	0
NLD	.01	.11	8	-.07	.11	11	3
NOR	-.23	.11	18	-.18	.11	17	1
NZL	.34	.11	3	.38	.11	3	0
POL	-.07	.11	12	-.12	.11	14	2
PRT	-.27	.11	20	-.31	.11	21	1
SVK	-.49	.12	26	-.52	.12	26	0
SWE	.05	.11	6	-.01	.11	8	2
TUR	-.80	.11	28	-.97	.11	28	0

The results in Table 3.5 are based on the calibration of the complete set of 28 reading items from PISA 2006 for the OECD countries sample. Though the ranking of the countries in Table 3.5 is based on the PISA 2006 calibration sample which is a random sample representative of the entire population of a country, the rankings obtained on this calibration sample (before modeling CDIF) are consistent with the PISA 2006 reading test rankings published by the OECD (see, OECD, 2007, page 298).

An analogous analysis was also carried out for the partner countries based on a random sample of 13,500 students (500 students from each of the 27 partner countries) from the PISA 2006 reading test. This resulted in an average shift of 0.44. Running the procedure on a combined sample of OECD and partner countries resulted in a shift of 2.03 places.

In 2006, reading was a minor domain in PISA. Next, the impact of CDIF on the reading assessment in 2009 was studied, when reading was a major domain. In 2009 PISA used a two-step procedure for calibrating item parameters. The PISA 2009 reading test consists of a set of core items, as well as additional standard and easy item blocks. The core items are given to all countries. The OECD countries except Mexico and Chile get the standard items and the easy items are given to 20 countries whose mean scores were lower on the PISA reading scale based on evidence from previous data collections. In step 1, the core and standard items are calibrated simultaneously. In step 2 the easy items are calibrated while keeping the core item parameters fixed at their estimates in Step 1. Once again random sampled data of 500 students from each OECD country is used to scale the core and standard items. Step 1 of this process was recalibrated (in which all item parameters are free) to show the impact of CDIF on the core and standard items dataset that is given to the OECD countries. A respondent sample of 500 respondents per OECD country was also selected, to study the impact of CDIF on the scaling model. The numbers of items selected for analysis are the core and standard items, altogether 101 items out of a total of 131 items in the 2009 reading test. The results of the analysis are presented in Table 3.6 below. 18 items with highest CDIF out of the 101 total items were modeled items using the procedure described above. The format of Table 3.6 is similar to the format of Table 3.5. It can be seen that the average number of places a country's ranking shifts for the PISA 2009 OECD reading calibration sample is 0.58. In the context of CDIF and shifting of country rankings, this indicates that the PISA 2009 model with reading as a major domain was more robust than the PISA 2006 reading model when reading was a minor domain.

Table 3.6
Rankings of countries for the randomly sampled OECD countries based sample of 15,500 students (500 students from each of the 31 OECD countries) used by PISA for item calibration of the PISA 2009 reading test. Average change of rank is 0.58.

	Original Rasch Model			CDIF Model			Change
Country	Mean	s.e.	Rank	Mean	s.e.	Rank	Rank
AUS	.16	.06	12	.15	.06	14	2
AUT	-.14	.06	24	-.13	.06	25	1
BEL	.39	.06	4	.39	.06	4	0
CAN	.30	.06	7	.32	.06	6	1
CHE	-.02	.06	23	.00	.06	21	2
CHL	-.53	.05	30	-.55	.05	30	0
CZE	.14	.06	15	.13	.06	15	0
DEU	.00	.06	21	.04	.06	20	1
DNK	-.20	.05	26	-.22	.05	27	1
ESP	-.14	.05	25	-.09	.05	24	1
FIN	.56	.05	2	.57	.05	2	0
FRA	.23	.06	9	.30	.06	8	1
GBR	-.01	.05	22	-.03	.05	23	1
GRC	.07	.06	16	.09	.06	16	0
HUN	.14	.06	14	.16	.06	12	2
IRL	.16	.06	11	.18	.06	11	0
ISL	.15	.06	13	.16	.06	13	0
ITA	.06	.06	17	.09	.06	17	0
JPN	.47	.06	3	.46	.06	3	0
KOR	.68	.05	1	.72	.05	1	0
LUX	-.23	.06	27	-.17	.06	26	1
MEX	-.76	.05	31	-.77	.05	31	0
NLD	.31	.05	6	.32	.05	7	1
NOR	.19	.05	10	.20	.05	10	0
NZL	.36	.06	5	.36	.06	5	0
POL	.25	.06	8	.26	.06	9	1
PRT	.00	.06	19	.04	.06	19	0
SVK	-.26	.06	28	-.24	.06	28	0
SWE	.05	.06	18	.08	.06	18	0
TUR	-.31	.05	29	-.28	.05	29	0
USA	.00	.00	20	.00	.00	22	2

The final analyses pertain to the reliability of the rankings obtained in the analyses. Though the general public (and especially the press) usually views the country ranking as fixed and true outcomes, they come, of course, with some degree of statistical uncertainty. As can be verified from the Tables 3.1, 3.5 and 3.6, the confidence intervals of some of the countries overlap, and, as a consequence, their rank order cannot be determined with great certainty. The PISA reports give confidence ranges for the ranks of the countries, and display the uncertainty by indicating the probability with which country means differ from the overall mean by using color codes in their tables (for reading, refer to Figure 6.8b, p. 298, OECD, 2007). Here, a different approach is chosen. To take this uncertainty into account, the rank orders were computed using the estimated means and standard errors as reported above, but with the difference, that, for every

country, the rank order was determined ignoring countries with a mean inside either its 65% or 99% confidence band. The results are displayed in Table 3.7, together with the previously reported shifts. Note that the average change in rank order drops below one in all cases. Note that the change becomes virtually zero in the analyses carried out on the 2009 data.

Table 3.7
Average change in rank order

	Number	Number	Accounting for Uncertainty		
Data	Items	Countries	No	35%	1%
Booklet 6	20	57	3.57	0.95	0.29
Booklet 6	28	57	2.04	0.57	0.16
All 2006	28	29	1.24	0.10	0.07
All 2006	28	57	2.03	0.41	0.02
All 2009	101	31	0.58	0.00	0.00

3.3 Discussion

The Rasch Model is indeed a very special model that has a range of properties that support a very powerful form of measurement (Rasch, 1960; Glas & Verhelst, 1995; Adams & Wu, 2007). In practice, however, there is often a discrepancy between the behavior of the observed data and the ideal as described by the model assumptions. With samples as large as those in PISA, it is easy to be confident that there are indeed discrepancies between the model and the observed data.

When discrepancies are observed between an adopted model and observed data one could firstly explore alternative scaling models, or secondly, proceed with use of the model on the assumption that the results will still have utility even if the data and model are not fully compatible (e.g. Molenaar, 1997).

K&C have explored one approach but conceded that it was not viable. Goldstein, Bonnet and Rocher (2007) show how alternative approaches can be used for certain analytic purposes. Other large scale studies, for example the TIMSS study since 1999, use the so-called three parameter logistic model (Lord & Novick, 1968). The two- and three-parameter logistic models are more general than the Rasch model, but using K&C's criteria, they still do not fit PISA data. The two-parameter model, which as K&C mention permits differing item discriminations, has been applied to PISA data and it has been shown that the outcomes are identical to those obtained when fitting a Rasch model (Macaskill, 2008) or very close when item treatments

similar to those presented in this paper are applied (Oliveri & von Davier, 2011). Further, the dependency between PISA items that K&C mention has also been modelled, and no implications for the rankings have been observed (Macaskill, 2008).

According to Raymond Adams (http://www.oecd.org/pisa/47681954.pdf), who was head of the PISA consortium from 2000 till 2012, when exploring alternative models one must take into account the full PISA context. In doing so some factors that need to be taken into account when considering alternative models include:

- The need to comprehensively cover the constructs
- The need for analytic techniques that work simultaneously for more than 50 countries;
- The need to integrate with appropriate sampling methodologies;
- The requirement to provide a database that is accessible to and usable by secondary data analysts;
- The need to support the construction of described proficiency scales;
- The need to permit inter-country comparison of overall levels of performance;
- The need to support the construction of scales that retain their meaning over time.

The second option is to proceed with the Rasch model acknowledging that it does not fit perfectly– and no model will – but undertaking work to ensure that the violations of the model are properly recognized in terms of their impact on the outcomes. The work on item choice and alternative scaling models described above is exactly that kind of work. An investigation was done to verfiy the claim by K&C that the data analyzed from booklet 6 is severely affected by CDIF and that the results obtained on the data from booklet 6 can be extrapolated to the entire PISA 2006 data. This assertion was found to be incorrect. The analysis started by analyzing K&C's sample of data from booklet 6 which consisted of 20 out of a total of 28 items in booklet 6. K&C omitted 8 items because of some missing data on those items. It was observed that using only 20 items from booklet 6 is resulting in large CDIF but when booklet 6 data was scaled using all 28 items the size of CDIF in terms of the average shift in country rankings after modeling CDIF reduced sizably. Next, the sampled dataset from all the OECD and partner countries was analyzed. The presence of CDIF in this dataset was found to be almost identical to the CDIF from booklet 6 data when all 28 items were used for calibration. This reinforces the stance that restricting the data to fewer items than in the total test ignores the complexity of the information gained in a realistic assessment situation. Finally, impact of CDIF on the PISA 2009 reading dataset was studied in which reading was the major domain and test length was

much longer than in PISA 2006. The results indicate that the PISA 2009 reading model fit is even more robust than the PISA 2006 reading model fit with respect to CDIF, which again reinforces the assertion that K&C's decision to focus on reading items even though the main subject of PISA 2006 was science was not ideal.

Chapter 4

Correcting for Differential Item Functioning in Multi-level Regression Models in Cross-National Surveys

Results of the PISA project have shown that the school average socio-economic status is an important background variable that explains a lot of variance in the student results. However, if the socio-economic variable which is measured at the student level is biased across countries due to a cultural bias, then the aggregated variable (at school level) is also subject to error. In this article, DIF (Differential Item Functioning, i.e., item bias) is mitigated using country specific item parameters. The effect of using this approach is studied on the results from multilevel regression for different measurement models and person parameter estimation procedures. Results showed that for countries affected by DIF the impact on the regression coefficients cannot be ignored. The effect is shown to be more for the PCM than the GPCM and generally more for the EAP estimates than the WML estimates.

4.1 Introduction

The PISA (Program for International Student Assessment) educational survey of the OECD aims to evaluate students' skills and aptitude for lifelong learning in the areas of reading literacy, mathematics, science and problem solving. The target population of PISA consists of 15 year-old students. The first PISA cycle started in the year 2000 and is repeated every three years with one of the three mentioned domains being the focus in each cycle. Besides collecting responses on cognitive tests, PISA also collects data on background characteristics of the students through so-called context questionnaires. The information provided by these background questionnaires is key to the ongoing success of PISA, since these questionnaires provide valuable information about the factors which affect students' test performance. This helps the participating countries to frame effective educational policies to tackle the shortcomings in their educational school systems that are revealed by PISA. It also gives the

participating countries an opportunity to observe how background variables may relate to student performance in other countries.

One of the key background variables that explains much variation in student performance is the socio-economic status variable. However, the questionnaire items administered to the students are international and there is always the question of whether the items used as a proxy for socio-economic status function consistent across countries. For example, an item like the number of cars owned by a family may be a proxy for socio-economic status in New York or in a suburban town, but the importance given to the number of cars might be different in these two places. The presence of such a bias in the measurement scale can lead to measurement error in the variable across respondents from diverse cultural or geographical groups. This measurement bias where an item is behaving differently across distinct groups is called Differential Item Functioning (DIF). This can have a bearing on the measurement precision of a key variable like socio economic status and subsequently on the amount of variance it explains in the regression model. Since the background variables are consequential for educational policy making in the various participating countries, it is important to take steps to ensure that cross country bias is minimized in the measurement of the variables of interest.

PISA currently uses a composite index for the socio-economic status of students. This index is composed of three attributes of a student's background, that is, possessions at home (HOMEPOS), highest parental education (PARED) and highest parental occupation (HISEI). PISA also makes use of these three attributes of socio-economic status separately in the regression models. HOMEPOS is a latent variable, which means that it is not directly observed but inferred using a measurement model. PARED and HISEI on the other hand, are directly observed scores based on a comparative framework across countries. To form a composite index from these three sub-components a principal component analysis is done to compute a composite scale score. The method to assess DIF is applicable to multi-item scales, so we use the HOMEPOS scale in our analysis. The HOMEPOS scale consists of a common set of 20 items that are administered across all participating countries. The items are both dichotomous and polytomous. The dichotomous items ask the examinee if he or she has a certain household possession at home, like an own room or a computer or a dishwasher etcetera. The polytomous items ask the examinee how many televisions or cars or bathrooms are present in the household. There has been a long debate in PISA whether to use national or international item parameters to measure the HOMEPOS scale. The item parameters of the measurement model used to scale HOMEPOS represent one or more characteristics of the test item. National item parameters are

those which are obtained by using within country scaling and international item parameters are those which are obtained by scaling a sample of students drawn from a group of countries. Both methods have their advantages and disadvantages and so far there is no consensus on which is better. Using international item parameters is impacted by DIF as already explained. On the other hand using national item parameters avoids DIF but the consequence is that meaningful comparisons between countries become harder to make.

The aim of this article is to study the impact of cross-country DIF on multi-level regression of student performance on HOMEPOS and Mean HOMEPOS (average HOMEPOS of all individuals in a school), that is, to assess how the regression coefficients of HOMEPOS and Mean HOMEPOS vary with and without compensating for DIF. We propose to assess the impact of DIF in the latent scale HOMEPOS by using country specific item parameters for any items displaying large DIF. This approach enables the use of international calibration for the item parameters and for making meaningful international comparisons while tackling the cultural bias inherent in international measurement instruments. The effect of using country specific item parameters for DIF items in the HOMEPOS scale is shown for different measurement models and different parameter estimation procedures to show how well this approach functions in different settings.

4.2 Method

This section gives an overview of how the analyses were conducted and discusses the theory they are based on. Firstly, we describe the Item Response Theory (IRT) models which are used as measurement models in our analyses. Secondly, we present the estimation process for our analyses starting with the calibration phase of the estimation process, that is, how item parameters are estimated. Thirdly, we describe how DIF is identified and compensated for with country specific items parameters. Then the scoring methods for person parameters, that is, the latent trait scores of persons are described. Finally the multi-level regression model is presented.

4.2.1 Item Response Theory

IRT has been gaining popularity amongst survey researchers as a tool for analysing survey data, especially in the field of education. IRT relates the item responses on a test to a latent variable, for example reading achievement (see, for instance, Lord, 1980). The use of IRT is especially relevant for the analyses of the PISA scales. The term scales refers to groups of questions/items

which are clustered together and aimed at measuring certain constructs of interest like for instance ‘attitude towards reading’. Most scales in PISA consist of polytomously scored items. There are several reasons for using IRT-based scores rather than conventional sum scores or weighted sum scores for measuring latent constructs. As will become clear in the example given below, one of the main advantages is that IRT-based latent scale scores come with standard errors which reflect the measurement error of the instruments used. Further, IRT-based IRT scale scores remain comparable when the responses are collected in a so-called incomplete design, that is, when respondents are presented different sets of items.

IRT models describe the relationship between an examinee’s standing on a latent variable, say educational achievement or an attitude, and item responses, based on the characteristics of the items of the test. The dependence of the observed responses on the dichotomously or polytomously scored items on the latent person variable is fully specified by the item characteristic function, which is the regression of an item score on latent variable. The item characteristic function allows inference about latent variable to be made from the observed item responses. The item characteristic functions cannot be directly observed because the ability parameter is not observed. But under certain assumptions it is possible to infer the information of interest from the examinee’s responses to the test items (e.g. Lord 1980).

An example of an item characteristic function for a dichotomous item is the one parameter model. The probability that an examinee answers an item correctly depends on his ability and the difficulty of the item. The difficulty parameter is the point on the ability scale where the probability of a correct response is 50%. So the larger the value of the difficulty parameter, the higher the ability that is required to have a 50% chance of getting the item correct. In Figure 3.1, two different item characteristic functions are plotted for a one parameter model. The functions differ by location on the ability scale. Item 2 is more difficult and shifted to the right on the difficulty scale.

The one-parameter model can be extended to a two-parameter model where the probability that an examinee answers an item correctly depends not only on his or her ability and the difficulty of the item but also on the discriminating behaviour of the item. The difficulty parameter is the point on the ability scale where the probability of a correct response is 50%. The discrimination parameter is proportional to the slope of the item response function at the point of the difficulty parameter on the ability scale.

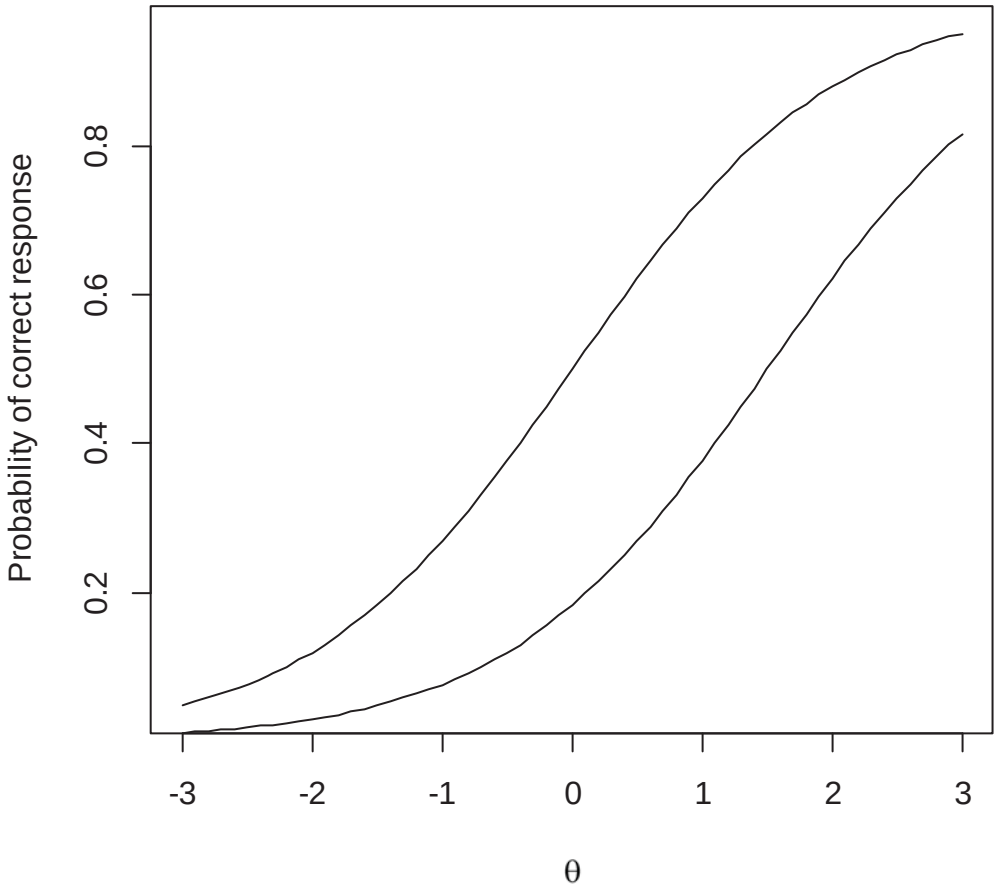

Figure 4.1. Item characteristic function for a one parameter model.

Items with high discrimination parameter values are useful for separating examinees into different ability levels. In Figure 4.2, two different item characteristic functions are plotted. The functions differ by location on the ability scale and by the slopes. Item 2 is more difficult and shifted to the right on the ability scale. The item characteristic curve of Item 2 corresponds with a higher discrimination parameter. As a result, a small increase in ability leads to a higher increase in probability of scoring correct compared to item 1.

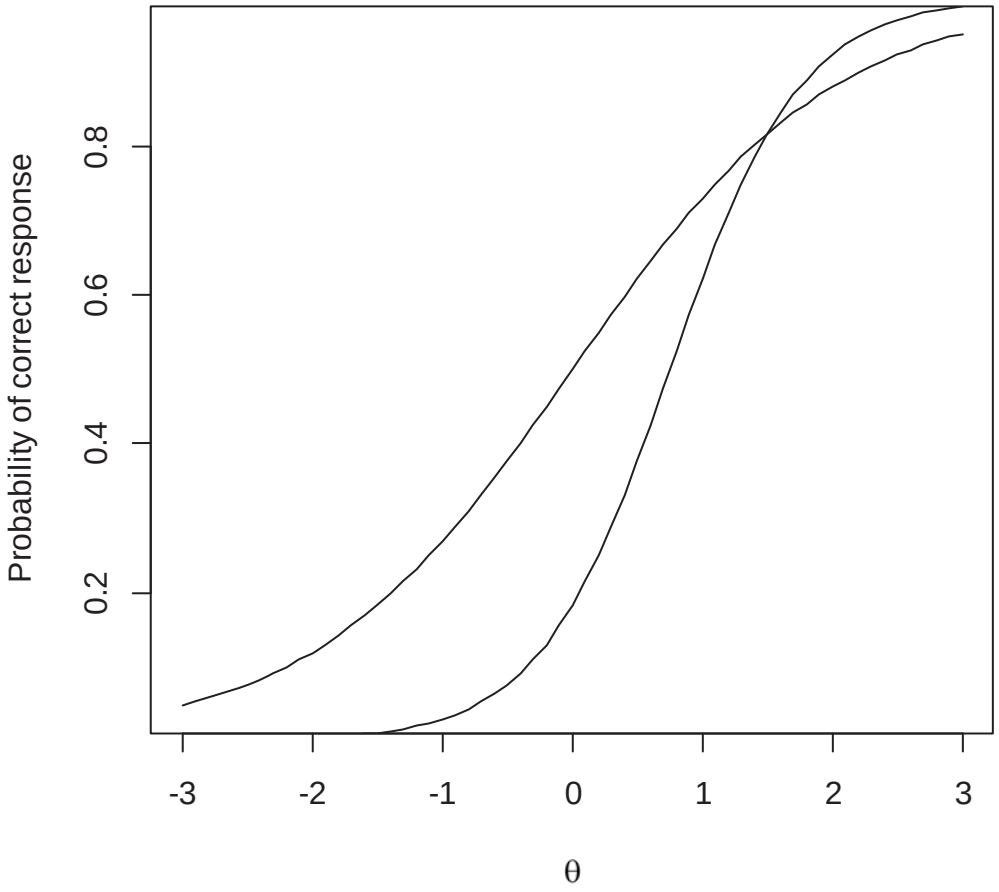

Figure 4.2. Item characteristic function for a two parameter model.

An IRT model may provide an adequate description of the test data. It is essential to test the fit of the model to the data. The examinees' abilities are unobservable but can be estimated. An IRT model provides a framework for the uncertainty regarding the estimate of the ability. So, an IRT model can be used to measure the abilities of the examinees, and quantifies the uncertainty regarding the estimate. For an overview of different IRT models, see, for example, Hambleton and Swaminathan (1985).

An IRT model for polytomous items is the Generalized Partial Credit model GPCM by Muraki (1992). If an item is dichotomous the GPCM reduces to the two-parameter model.

In the GPCM, the probability of a student *n* scoring in category *j* on item *i* (denoted by $X_{nij} = 1$) is given by

$$P(\mathrm{X}_{nij}=1 \mid \theta_n) = P_{ij}(\theta_n) = \frac{\exp(j\alpha_i\theta_n - \beta_{ij})}{1+\sum_{j=1}^{M_i}\exp(j\alpha_i\theta_n - \beta_{ij})}, \quad (4.1)$$

for $j = 1,\ldots, M_i$. As the latent trait level of the student increases the probability of scoring in higher categories increases. For every latent trait level θ (theta) there is a certain category where the probability of responding is the highest. The expected item-total score is given by

$$E(T_i \mid \theta) = \sum_{j=1}^{M_i} jE(X_{ij} \mid \theta) = \sum_{j=1}^{M_i} jP_{ij}(\theta). \quad (4.2)$$

where the item-total score is defined as $T_i = \sum_{j=1}^{M_i} jX_{ij}$. Note that the expected item-total score increases as a function of θ. If the discrimination parameter α_i is constrained to one, the GPCM reduces to the PCM (Masters, 1982).

4.2.2 Estimation Process

The estimation process described below is for the independent variable HOMEPOS that is used in the regression analyses. The dependent variable in the regression analyses is the reading ability of the student. The reading ability, which is a latent trait, is not estimated here, but directly used from the PISA 2009 student questionnaire database. A brief description of how the reading ability was estimated in PISA 2009 is given later in the Multi-level model section.

4.2.2.1 Item calibration

The estimation process for the independent variable HOMEPOS starts with a calibration run to estimate the item parameters. This is done using the marginal maximum likelihood (MML) method (Bock & Aitkin, 1981). In this approach, the person parameters θ are considered as nuisance parameters. They are assumed to be samples from one or more normal distributions and intenerated out of the likelihood function. Thus, the likelihood function no longer depends on the person parameters, but only on the item parameters and the means and standard deviations of the population distributions. In the application presented here, every country has its own distribution.

4.2.2.2 Country-specific item parameters

The next step after obtaining the item parameters was to investigate the presence of DIF or item bias across countries. In this study we focussed on uniform DIF across countries, which means, that only changes in the item difficulty due to DIF are targeted and changes in the item discrimination parameter are ignored. Several options are available in the presence of DIF. One extreme option is to eliminate the DIF items from the measurement instrument. If the number of eliminated items is large, this has the drawback that the measurement precision decreases and the construct validity is threatened. An alternative is to model DIF using country-specific item parameters. The items without DIF identify the latent scale and it is assumed that the items with DIF still load on this latent scale but with specific item parameters for the concerning subgroup(s) (Glas & Verhelst, 1995; Glas, 1998). Thus, these items can be used to estimate the value on the latent variable and contribute to the precision of the estimates. The data of the set of countries in the study is analysed simultaneously to identify items with country-specific DIF and an item which is identified as a DIF item is modelled with country-specific item parameters. This is done in an iterative process where the item with the worst misfit is replaced with country specific item parameters, followed by a new overall analysis. So the DIF items were treated one-at-a-time until eight items had been replaced. The number of items that are calibrated using country specific item parameters in a scale is an arbitrary choice. The only thing that has to be ensured is that there is a sufficient number of anchor items remaining in the scale. The scale consisted of a total of twenty items, so replacing eight items with country specific item parameters meant that there were twelve items that were used to anchor the scale.

In order to identify DIF in the HOMEPOS scale we proceeded as follows. A set of observed and expected responses for an item in the scale was constructed over the entire student population in a country, while controlling for the different marginal distribution in each country. The observed responses on an item in a country are the sum of responses of all students on that item in that country. The expected responses for an item in a country are the sum of the expected response for all students on that item in that country. The expected response on an item is computed as a posterior expected response given the student's response pattern, the item parameters and the population parameters of the country from which the student is sampled. The difference between the sum of the observed and expected responses on an item in a country is an indicator of item bias in that country. Summing up this difference for all the countries and then dividing it by the number of countries is an indicator of the size of the global DIF affecting that item across the set of countries in the analysis. So the larger the value of this difference,

the larger the item bias or DIF across the countries. Likewise a statistic for the difference between observed and expected responses can be obtained for all the items in a scale like HOMEPOS. Next it can be seen which items have a comparatively higher difference between the observed and the expected values. Those are the items which are labelled as DIF items in our analyses. The item difficulty parameters of these DIF items are allowed to vary across countries in the measurement model to fit the data. This variation is country specific and thus the item parameters of the DIF item become country specific.

4.2.2.3 Scoring procedures

After obtaining the item parameters and correcting for DIF, the students can be assigned latent trait scores given the item parameters and their response patterns. In this section the theory behind the scoring procedures is discussed and in the next section the scoring process is described.

Point estimates of the students' latent trait scores are obtained given the estimates item parameters and the response pattern. However this point estimate of the latent trait score has uncertainty associated with it and using it directly in subsequent statistical analyses leads to too narrow confidence intervals. Therefore, both for the dependent and the independent variables, plausible values rather than point estimates are used to take the uncertainty of the estimates into account. Plausible values were first developed for the analyses of NAEP (National Assessment of Educational Progress) data, by Mislevy, Sheehan, Beaton and Johnson (1983), based on Rubin's work on multiple imputations (1978). They are draws from the posterior distribution of a student's ability given his response pattern. Therefore, plausible values provide not only information about the estimate of a variable, but also the uncertainty associated with this estimate.

The two most widely used procedures to estimate person parameters are Maximum Likelihood (ML) estimation and Expected A Posteriori (EAP) estimation. These two estimation procedures resulted out of respectively a frequentist and a Bayesian approach to estimation.

4.2.2.4 WML estimation

A frequentist approach to estimating the parameters of an IRT model is given by the maximum likelihood (ML) method. The likelihood function models the likelihood of a certain response pattern. Maximizing this function with respect to the person parameter results in the Maximum

Likelihood (ML) estimate. However this ML estimate has a bias. This bias can be corrected by attaching a weight to the ML estimates of each person. Correcting for the bias by attaching a weight to the ML estimates results in unbiased weighted maximum likelihood or WML estimates (Warm, 1989).

4.2.2.5 EAP estimation

The Bayesian approach makes use of prior distributions and inferences are based on the posterior distribution. The prior distribution is a beforehand notion about the parameters, for instance about the mean and variance of the population, often based on some theoretical assumption. The posterior distribution incorporates both this prior information and the information from the data. An often noted disadvantage of Bayesian statistics is that the choice of the prior in the parameter estimation procedure is in some way subjective. However, as the sample size increases the weight of the data far outweighs that of the prior (Gelman, Carlin, Stern, & Rubin, 2004).

A point estimate of the latent score can be obtained by using the mode or the expectation of the posterior distribution. We use the latter, which is known as the expected-a-posteriori or EAP estimate (Bock & Mislevy, 1982). The technique capitalizes on an insight of Thomas Bayes which enables us to find the conditional probability of an event theta (e.g. ability) given the conditional probability of event X (e.g. response pattern of student) and the unconditional probabilities of events theta and X. The unconditional probability of theta is the prior information about a person's ability estimate. The EAP estimates are the expected value of the conditional probability distribution of theta.

4.2.2.6 Estimation process

The MML, WLM and EAP estimates were computed using the MIRT software (Glas 2010). Plausible values for person parameters were then drawn for the WML and the EAP approaches. The WML estimates have an asymptotic normal distribution with expectation equal to the WML estimate and variance equal to the square of the standard error. The plausible values were drawn from this normal distribution. For the EAP based plausible value, the draws are made directly from the posterior distribution without asymptotic assumptions.

4.2.3 The Multilevel Regression Model

The dataset used for the analyses was the PISA 2009 International Student dataset. All the analyses were done for equal-sized samples of students from 10 culturally diverse countries. From each of the 10 countries, 1500 students were sampled. The list of these countries for which the analyses were conducted are presented in the tables. The regression analyses were done separately for each country. The regression model used for the analyses was a 2-level random intercepts model (Bryk & Raudenbush, 2001). The multi-level model consisted of a student and school level. The independent variables were the home possession variable at the student level (HOMEPOS) and the Mean of the home possessions variable at the school level (Mean HOMEPOS). The mean of the HOMEPOS variable was obtained by summing the HOMEPOS indices of individuals in a school and then dividing the sum over the number of students in the school. PISA studies have shown that Mean HOMEPOS explains more variance than the individual level HOMEPOS index (as is the case for other facets of socio-economic status variables) and therefore we used a multilevel model that includes the Mean HOMEPOS variable. The model that was run for the multilevel analyses is presented below:

$$\theta_{ij} = \beta_{0j} + \beta_1(HOMEPOS_{ij}) + \varepsilon_{ij} \tag{4.3}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{10}(Mean\_HOMEPOS_j) + \mu_j \tag{4.4}$$

This is a random intercepts model. The random intercept component of the model is given by equation 4.4. The subscript *i* represents a student and subscript *j* denotes the school. Thus the left hand side of equation 3 represents a plausible value of the reading ability of a student *i* in school *j* . These plausible values for reading ability were based on EAP estimates and the measurement model was the PCM. These plausible value draws were fully conditional draws that are conditioned on the background variables (see PISA 2009 Technical Report). Five Plausible values for reading ability were drawn from the posterior distribution based on this fully conditional model. For our analyses, these five plausible values of reading ability and five plausible values for the HOMEPOS variable (which we estimated) were used. Thus five regression analyses were done for five sets of plausible values of the dependent and independent variables, where each of the five values for the HOMEPOS variable is regressed on one of the five values for the dependent variable. The range of the plausible values for the dependent variable (ability estimate) were between 300 and 700 units on the PISA scale approximately. The range of plausible values for the independent variable HOMEPOS ranged between -3 and

3. The multi-level regression analyses were done five times using a different set of plausible values of the dependent and independent variable. The results for the effect size of the independent variables were then averaged by summing the effect size from the five runs and then dividing by five. The standard errors and the significance tests were then adjusted for the variation between the five sets of results. The variance of an estimate is the average of the variances of the estimate from five runs plus the between-runs variance of the five estimates (Rubin, 1978). The standard errors were simply the square root of the estimated variance.

For the HOMEPOS variable, different measurement estimates are used in the analyses. The measurement models used to calibrate the Home Possessions scale are the PCM and the GPCM and the scales scores used for home possessions variable are the WML and EAP based plausible values. Afterwards these estimation exercises are repeated using country specific item parameters to model DIF. The investigated procedures and their labels are listed in Tables 4.1 through 4.4.

4.3 Results

The results of the analyses are presented in Tables 4.1 through 4.4. Table 4.1 and Table 4.2 present the results of regression analyses for the PCM and the GPCM estimates of the HOMEPOS variable without using any country specific item parameters. These different estimates are computed using WML and EAP based plausible values for HOMEPOS. The level 1 results are presented under the heading HOMEPOS and the level 2 effects under the heading Mean HOMEPOS.

The analyses presented in Table 4.1 and Table 4.2 are repeated using country specific item parameters in Table 4.3 and Table 4.4, which present the results for the case where eight country specific item parameters have been used to compensate for DIF items in the HOMEPOS scale.

Table 4.1. Regression coefficients and standard errors (S.E.) for HOMEPOS using the PCM with regular item parameters. Plausible values are used in this table.

	HOMEPOS				Mean HOMEPOS			
Country	WML	S.E.	EAP	S.E.	WML	S.E.	EAP	S.E.
Brazil	8.2	6.8	**17.8**	8.1	**25.4**	10.3	**23.0**	10.3
Germany	**11.6**	4.2	**17.6**	6.3	**44.0**	8.9	**67.0**	13.4
Finland	**18.2**	6.4	**27.6**	12.9	4.2	13.7	14.4	18.2
Indonesia	-4.4	3.2	-5.5	5.0	**26.2**	7.4	**37.8**	8.1
Japan	4.6	7.1	4.3	8.7	**31.4**	13.5	**53.4**	20.3
Mexico	2.4	5.6	1.1	7.4	**17.2**	6.6	**25.2**	8.1
Netherlands	-2.6	6.0	-2.5	9.1	**62.4**	15.1	**118.8**	24.8
Shanghai	6.2	4.6	8.9	7.5	**40.2**	8.9	**57.4**	11.0
Thailand	4.7	4.5	3.9	5.6	**29.0**	6.1	**33.2**	6.3
United States	**13.3**	5.6	**16.0**	7.2	**31.8**	12.4	**46.4**	11.2

The significant effects (at 5%) are highlighted in bold.

Table 4.2. Regression coefficients and standard errors (S.E.) for HOMEPOS using the GPCM with regular item parameters. Plausible values are used in this table.

	HOMEPOS				Mean HOMEPOS			
Country	WML	S.E.	EAP	S.E.	WML	S.E.	EAP	S.E.
Brazil	9.3	4.9	**11.8**	5.6	**10.8**	5.4	**16.0**	6.8
Germany	**7.0**	3.1	**14.2**	4.9	**30.8**	7.6	**46.4**	10.7
Finland	**17.0**	4.9	**25.3**	9.1	0.6	8.4	-1.7	15.7
Indonesia	-1.1	2.1	-2.8	2.9	**14.4**	5.7	**23.7**	5.8
Japan	1.4	5.3	2.8	8.4	**25.3**	10.5	**36.4**	16.7
Mexico	1.8	5.0	2.4	5.6	**13.9**	5.4	**16.8**	6.3
Netherlands	-1.8	4.2	-7.4	9.6	**35.2**	11.5	**86.1**	22.8
Shanghai	4.1	3.3	5.7	4.5	**28.9**	6.8	**40.9**	8.1
Thailand	2.3	2.8	3.4	3.1	**22.5**	4.2	**24.1**	4.2
United States	**7.6**	3.7	**12.7**	5.5	**23.1**	6.5	**35.1**	9.1

The significant effects (at 5%) are highlighted in bold.

Table 4.1 compares the results obtained using the PCM for the WML and EAP estimation procedures. Table 4.2 presents the same results for the GPCM. From the tables it can be seen that the level 1 effects are significant for about half of the countries. The level 2 effects are significant for all countries except Finland. The effects are systematically larger for the PCM. However the effect size for the PCM and the GPCM cannot be compared because they are on different scales. However, the ordering of the countries with respect to the size of the regression coefficients and the significances can be compared and they are very similar using the PCM and the GPCM. Another result that becomes evident is that the size of the regression effects is larger for the EAP based estimates than the WML based estimates for both the PCM and the GPCM. Tables 4.3 and 4.4 present the results obtained using country specific item parameters for the PCM and the GPCM.

Table 4.3. Regression coefficients and standard errors (S.E.) for HOMEPOS using the PCM with eight country specific (C.S.) item parameters. Plausible values are used in this table.

	HOMEPOS				Mean HOMEPOS			
Country	WML	S.E.	EAP	S.E.	WML	S.E.	EAP	S.E.
Brazil	**13.0**	6.0	15.2	8.5	**17.2**	7.3	**24.0**	10.3
Germany	**11.6**	3.8	**15.6**	5.3	**35.0**	8.1	**48.0**	10.2
Finland	**19.2**	5.7	**31.4**	8.3	-1.4	8.9	-7.8	14.4
Indonesia	-2.7	2.7	-3.8	3.8	**22.0**	6.3	**30.4**	7.6
Japan	1.6	4.6	7.0	7.8	**33.6**	11.4	**48.4**	17.5
Mexico	0.2	5.8	3.2	7.2	**21.9**	6.8	**23.6**	8.3
Netherlands	1.7	3.8	0.5	6.6	**40.9**	11.7	**74.0**	13.2
Shanghai	6.6	3.4	**10.0**	4.9	**33.5**	7.0	**45.4**	9.5
Thailand	3.2	3.7	5.9	4.5	**30.9**	5.7	**34.4**	5.9
United States	**10.6**	4.4	**18.1**	6.7	**25.9**	8.6	**36.8**	10.7

The significant effects (at 5%) are highlighted in bold.

Table 4.4. Regression coefficients and standard errors (S.E.) for HOMEPOS using the GPCM with eight country specific (C.S.) item parameters. Plausible values are used in this table.

	HOMEPOS				Mean HOMEPOS			
Country	WML	S.E.	EAP	S.E.	WML	S.E.	EAP	S.E.
Brazil	**11.6**	5.4	12.9	5.9	**13.4**	6.5	**18.0**	7.0
Germany	**9.4**	3.5	**16.2**	4.2	**34.2**	7.2	**47.4**	9.7
Finland	**19.3**	4.8	**28.2**	7.4	-0.8	8.0	0.8	13.4
Indonesia	-2.2	2.2	-2.1	2.8	**18.5**	5.7	**21.1**	5.7
Japan	5.2	4.1	7.4	5.9	**29.2**	10.8	**46.9**	13.3
Mexico	2.1	5.2	1.6	5.4	**16.1**	6.1	**19.2**	6.1
Netherlands	0.5	2.9	0.8	4.9	**33.5**	7.3	**58.3**	11.4
Shanghai	5.2	3.3	**8.9**	4.2	**29.3**	6.3	**39.6**	7.8
Thailand	3.2	3.3	2.5	3.2	**23.3**	4.7	**27.7**	4.5
United States	**7.8**	3.5	**15.0**	5.3	**23.2**	6.6	**30.9**	8.6

The significant effects (at 5%) are highlighted in bold.

For the PCM in Table 4.3, introducing country specific item parameters causes the effect sizes to change compared to Table 4.1. For the GPCM in Table 4.4, introducing country specific item parameters causes a smaller change in the effect sizes (when comparing with Table 4.2) than the change in effect sizes for the PCM. Figures 4.3 and 4.4 show this for the Mean HOMEPOS variable for all the countries (except Finland for which the result was not significant).

The results showed that change for the Netherlands was especially large when using country specific item parameters for EAP estimates with the PCM or the GPCM. This is because the results for DIF showed that the Netherlands was most impacted by DIF in the HOMEPOS scale. A large number of items in the scale displayed sizeable DIF for the Netherlands during item calibration.

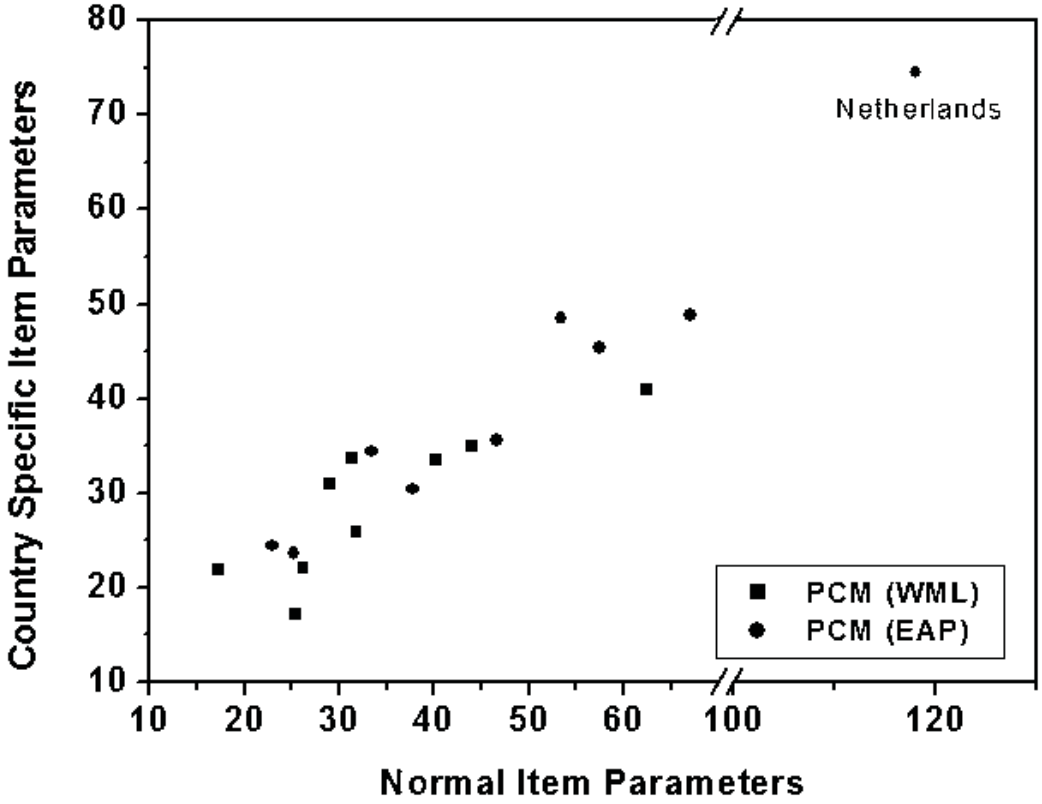

Figure 4.3. Effect on regression parameters when introducing country-specific item parameters for the PCM.

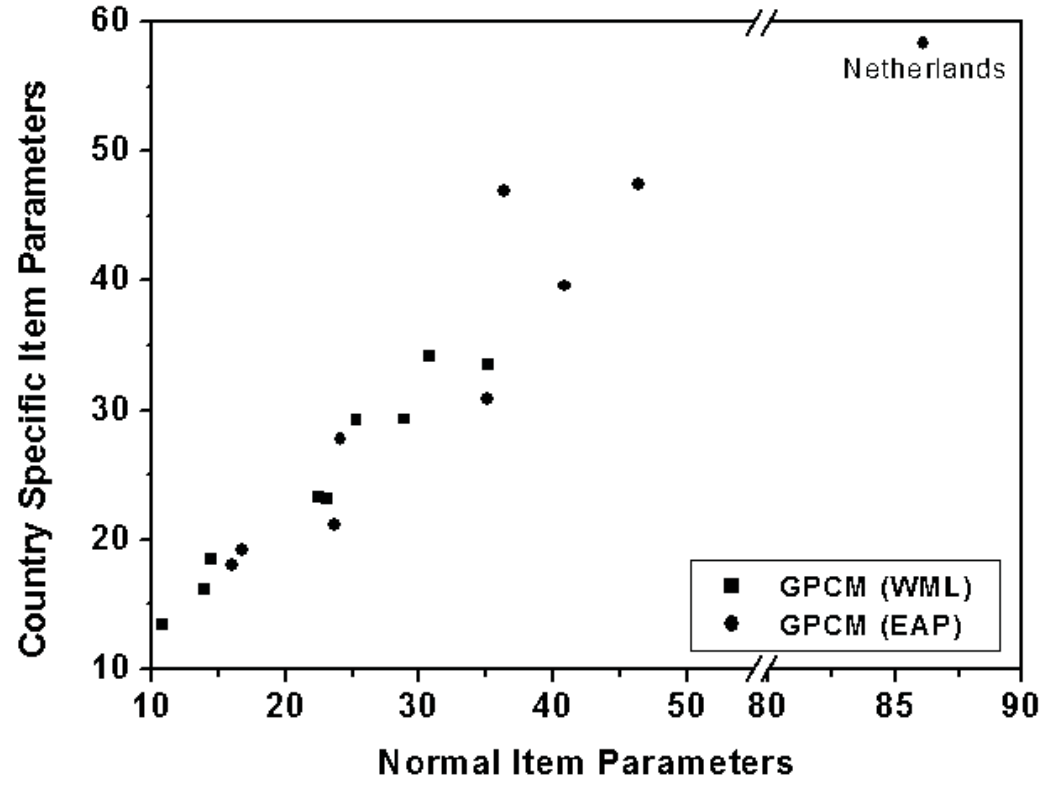

Figure 4.4. Effect on regression parameters when introducing country-specific item parameters for the GPCM

These results suggest that the PCM is more affected by DIF than the GPCM as country specific item parameters which mitigate the impact of DIF affect the PCM results more than the GPCM results.

4.4 Conclusions

The validity of cross national surveys is threatened by the presence of cultural bias especially for a survey like PISA which is conducted across a diverse group of countries around the world. The cultural bias can compromise comparisons across countries which are essential to cross-national surveys. In this study we investigated the impact of cultural bias in the framework of IRT modeling, which is increasingly used for measurement in survey research. We selected an important background scale from PISA 2009 to study the impact of cultural bias or DIF in the framework of IRT. We wished to see the impact of DIF on the regression model as ultimately the inferences from the background scales are made on the regression outcomes. For our analyses we investigated the impact of DIF using different measurement models and scoring methods. We did this for the PCM and the GPCM as the measurement models and the WLM and EAP scoring methods. We modeled DIF in the framework of these different approaches using country specific item parameters and analyzed the results from the subsequent regression analyses.

The first thing that is evident from the analyses is that the country specific item parameters affect the results for the PCM more than the GPCM across the spectrum of analyses. Though the effect sizes for the GPCM and the PCM cannot be directly compared with each other because the two models are estimated on a different scale, however the differences produced when using regular or country specific item parameters for each of the two models can be compared. After using country specific item parameters the change in the effect sizes for the PCM was in most cases larger than the GPCM. As country specific item parameters mitigate the impact of DIF, it can be concluded that the PCM is more vulnerable to DIF than the GPCM.

Another result that becomes evident is that the change in regression coefficients when using normal versus country specific item parameters was larger when using EAP estimates than when using WML estimates of the variable Home Possessions, especially for the PCM. The difference in effect sizes was especially prominent for countries with larger DIF in the Home Possessions variable like the Netherlands.

Concluding the results, we see that DIF in the latent scale under investigation impacts the results of the regression analyses and the use of country specific item parameters helps in mitigating the impact of DIF. The PCM is more affected by DIF than the GPCM especially when using EAP estimates of person parameters, so the use of country specific item parameters has greater need if the PCM is the measurement model. For the GPCM, overall the change in effect sizes was not large when using country specific item parameters but the precision of the estimates is still enhanced. For a country severely affected by DIF, using country specific item parameters with the GPCM using the EAP method also made a substantial difference (see Netherlands in Figure 4.4).

Thus for countries affected by DIF the impact on the regression coefficients cannot be ignored. The effect has been shown to be more for the PCM than the GPCM and generally more for the EAP estimates than the WML estimates. However as the ordering of the countries and the significances (in terms of effect sizes) were quite similar for the PCM and the GPCM using both the WML and EAP estimates which is often of interest in cross national surveys, either method may be used for future analyses provided DIF is adequately accounted for with country specific item parameters.

Chapter 5

Exploration of Order Effects in Test Administration Designs

The impact of order effects in item administration designs in large-scale educational surveys on the calibration of item parameters in various IRT measurement models is studied. As an illustration the PISA 2009 reading dataset is used. In earlier cycles, PISA used the 1 parameter logistic model (1PLM) for calibrating the cognitive data and included a parameter for each booklet in the item administration design. Currently, PISA calibrates item parameters for the cognitive domains using the 2 parameter logistic model (2PLM). The effect of using booklet parameters will be studied for both the 1PLM and 2PLM models. Further, in the current article three alternatives to model the order effect are studied. The first one adds country-by-booklet interaction parameters to the 1PLM and the 2PLM. The second adds position parameters to the 1PLM and 2PLM that represent the response process leading to the order effect more closely than the approach through booklet parameters. The third one adds country-by-position interaction parameters to the previous model.

The PISA 2009 reading domain is recalibrated using the PISA item calibration sample both using Marginal maximum likelihood (MML) and a Bayesian framework, both for the 1PLM and 2PLM, with and without booklet, position and interaction parameters. Model fit is investigated by using residual analysis at the item level in both the MML and Bayesian frameworks. Also the impact of including booklet parameters on the ordering of countries on the cognitive reading scale was investigated.

5.1 Introduction

An IRT model may provide an adequate description of the test data. However, it is essential to test the fit of the model to the data. Especially the model should explain aspects of data that impact inferences made using the model. To improve the fit of the IRT model used for calibrating

a test with possible item positioning effects like in a rotated booklet design, it is possible to include parameters in the IRT model to compensate for the item positioning effects in different test booklets.

A PISA cognitive test booklet contains four clusters of items from the different cognitive domains Reading, Science and Math. The PISA test design is called an incomplete balanced booklet design. It is incomplete because each booklet contains only a subset of the total items in the test. However, the order of appearance of items (in any cognitive test domain) in the rotated booklets is such that the design is balanced and the item parameter estimates that are obtained from scaling should not be influenced by a booklet effect.

Table 5.1 Cluster rotation design used to form standard test booklets for PISA 2009

Booklet ID	Cluster			
1	M1	R1	R3A	M3
2	R1	S1	R4A	R7
3	S1	R3A	M2	S3
4	R3A	R4A	S2	R2
5	R4A	M2	R5	M1
6	R5	R6	R7	R3A
7	R6	M3	S3	R4A
8	R2	M1	S1	R6
9	M2	S2	R6	R1
10	S2	R5	M3	S1
11	M3	R7	R2	M2
12	R7	S3	M1	S2
13	S3	R2	R1	R5

Table 5.1 gives the design used in the present chapter, which is the design used in the 2009 cycle of PISA. The survey items were allocated to thirteen item clusters (seven reading clusters, three mathematics clusters and three science clusters) with each cluster allotted 30 minutes of test time. The items were presented to students in thirteen standard test booklets, with each booklet being composed of four clusters. In Table 5.1, R1 to R7 denote the reading clusters, M1 to M3 denote the mathematics clusters, and S1 to S3 denote the science clusters. Each cluster appears in each of the four positions. In each country, the booklets were randomly administered to the sample of students. However due to the different location of the different clusters within each of the booklets it is expected that there would still be booklet influences on the estimated proficiency distributions. This was first supported by the exploratory analysis done by PISA in the first cycle in 2000. In later cycles, the effect persisted. The variation in the means between booklets was greater than expected. As the booklets were systematically rotated,

it was expected that the only between-booklet variance would be sampling variance. The variations observed between booklets appeared quite stable across countries leaving a picture of systematically easier and harder booklets. These differences would affect the ability estimates of the students who worked on easier or harder booklets and therefore booklet effects needed to be explained and corrected for in an appropriate way.

It was argued that modeling the order effect in terms of item positions in a booklet or at least in terms of cluster positions in a booklet would result in a very complex model. Therefore, for the sake of simplicity in the international scaling, PISA modeled the effect separately for each domain at the booklet level. When estimating the item parameters, booklet effects were included in the measurement model to prevent confounding item difficulties and booklet effects. The booklet parameter, formally defined in the same way as item parameters, reflects *booklet* difficulty. As will be outlined further below, the rationale of adding booklet parameters is that, since the booklets were randomly assigned, the mean ability of the students for the different booklets should be the same. So these means are actually corrected to be the same.

An alternative approach that will be studied below is to take the actual position of the clusters within a booklet into account, and weight the cluster difficulty with its position. When studying the responses and response propensities to subsets of items, this approach is theoretically much more precise than an approach with a global overall booklet effect. Whether this gain in precision is important will be studied below.

A point of concern in both the approach using booklet and position parameters is that it is assumed that the effects are uniform over countries. To study this assumption, interaction effects of countries with booklet and position parameters will be evaluated.

To study the impact of including order effects in the IRT measurement model, a number of analyses were done. Firstly, the models with and without order effects (that is, booklet and positions effects) were estimated to assess the size of the effects. Secondly, using these estimates, the differences between the ordering of the countries under the different models was evaluated. And finally, a number of item fit statistics was computed. The analysis were done in two frameworks, the traditional framework of large-scale educational surveys, that is the framework of the marginal maximum likelihood or MML framework (Bock & Aitkin, 1981) and a new emerging framework, and the Bayesian framework (Albert, 1992).

This chapter is organized as follows. First we describe the method used in this study. Then we describe the PISA 2009 Reading scale and the measurement models and the calibration process used. After that we describe the various methods used for examining model fit. Lastly we present the results and the conclusions.

5.2 Method

In this section, we first present details of how the item calibration is currently done in PISA for the OECD countries. Then we present the methodology used in this section to estimate the measurement model using different IRT models. After that we present different statistical tests that we use to compare model fit for different IRT models.

5.2.1 PISA 2009 Reading Scale

The PISA Reading ability was a latent scale that was measured using an international calibration of a sub-sample of students from the OECD countries. The sub sample of students referred to as an OECD calibration sample consisted of 15,500 students comprising 500 students drawn at random from each of the 31 participating OECD countries. The calibration was done on 101 reading items distributed over 13 booklets. There were 94 dichotomously scored items and 7 polytomously scored items. The polytomously scored items had three response categories. In the original analyses in the PISA project, the 1PLM (Rasch, 1960) for dichotomous items and the Partial Credit Model (PCM, Masters, 1982) for polytomous items were used as the measurement model and a booklet parameter was added to the measurement model to compensate for the booklet effects.

In the analyses described below, we use a similar international calibration sample as used by PISA 2009. Thus the calibration sample consisted of 15,500 students comprising 500 students drawn at random from each of the 31 participating OECD countries. The calibration was done on all 101 reading items distributed over 13 booklets.

5.2.2 Measurement models

In the analysis presented below, both one and two-parameter IRT models were used as scaling models. Both the logistic and the normal-ogive representation of the model were used. The logistic representation is more common, but in a Bayesian framework, the normal ogive

representation proves to be more practical (see, for instance, Albert, 1992), especially for the computation of residuals. The parameter estimates obtained using the logistic and normal ogive representations are hardly distinguishable. For the logistic representation, we define the logistic function

$$\Psi(x) = \frac{\exp(x)}{1+\exp(x)}$$

and for the normal ogive representation we define

$$\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-t^2/2)dt \ .$$

The cognitive scales in the PISA project consist mostly of dichotomously scored items. For dichotomous items, the scores on an item indexed i ($i = 1,2,....,K$) by students indexed n ($n = 1,...,N$) are denoted by $X_{ni}=1$ for a correct response or $X_{ni}=0$ for an incorrect response. To obtain the probability of a correct response in a two-parameter model, insert

$$\eta_{ni} = \alpha_i \theta_n - \beta_i \tag{5.1}$$

into either $\Psi(\eta_{ni})$ or $\Phi(\eta_{ni})$ depending on the representation. A one-parameter model emerges when the discrimination parameter α_i is fixed to 1.

The model with a booklet effect is obtained as

$$\eta_{ni} = \alpha_i \theta_n + \delta_{b(n)} - \beta_i \tag{5.2}$$

where $b(n)$ is the booklet administered to student n, that is, a number between 1 and 13. A restriction on the booklet parameters is needed to identified the model, say, $\delta_{13}=0$. The logistic versions of the model will be referred to as the 1PLMB and 2PLMB. The normal-ogive version of the model will be referred to as the 1PNOB and 2PNOB.

The model with a position effect is obtained as

$$\eta_{ni} = \alpha_i \theta_n + \xi_{b(n),p(i,b(n))} - \beta_i \tag{5.3}$$

where $p(i,b(n))$ is the position of the item i in the booklet $b(n)$ administered to student n. So $p(i,b(n)) = 1,...,4$. Again, restrictions are needed to identified the model, for instance, $\xi_{b(n),1}$ =

0, that is, the position parameter of the first position in each booklet is fixed to zero. The logistic versions of the model will be referred to as the 1PLMP and 2PLMP, the normal-ogive versions are labeled 1PNOP and 2PNOP, respectively.

Specific interactions with countries are obtained by adding a country index to the parameters, that is, $\delta_{c,b(n)}$ and $\xi_{c,b(n),p(i,b(n))}$. The one-parameter logistic versions of these two models will be referred to as the 1PLMCB and 1PLMCP, the two-parameters logistic versions will be referred to as 2PLMCB and 2PLMCP, respectively. Following an analogous logic, normal-ogive versions are labeled 1PNOCB, 1PNOCP, 2PNOCB and 2PNOCP, respectively.

For the MML framework, the MIRT software (Glas, 2010) was used to estimate the item and population parameters (means and variances of the ability distributions per country) of the IRT model for the PISA 2009 reading dataset. In the Bayesian framework the WinBugs software (Lunn, Spiegelhalter, Thomas & Best, 2009) was used to estimate all the item and population parameters in an MCMC framework. In the main study of PISA 2009 the PCM was used as the IRT model for the polytomously score cognitive data. In the present study, however, we use the steps model (Tutz, 1990; Verhelst, Glas & de Vries, 1997) for calibrating the polytomous data both in MIRT and WinBugs software instead of the categorical models like the PCM or the GPCM (Muraki, 1992) as they proved to be computationally too heavy for the WinBugs software given the enormous size of the calibration sample and the large item bank. The steps model considers a polytomous response data as a special case of data emanating from a multistage testing design with dichotomous items where every test consists of one dichotomous item only. So a polytomous item consists of a sequence of item steps and every item corresponds with a so called virtual dichotomous item. A student is only administered the next virtual dichotomous item if a correct response was given to the previous one. Thus the choice of a follow up test is a function of the responses on the previous items. Verhelst, Glas and de Vries (1997) show that the item response curves obtained in the PCM and the step model are very close, so the inferences emanating from the two models are also quite close. For more information on the Steps model see (Tutz, 1990; Verhelst, Glas & de Vries, 1997). Below, abbreviations such as 1PLM, 2PLM, 1PLMB, etc. will refer both to dichotomously and polytomously scored items.

5.2.3 Global tests for model fit

IRT is a special case of the more general framework of latent variable modeling, such as, for instance, implemented in the computer program MPLUS (Muthén & Muthén, 1998–2012). Therefore, traditional test statistics used in latent variable modeling, such as the likelihood ratio statistic and its modifications, the AIC (Akaike, 1974), BIC (Schwarz, 1978), and DIC (Spiegelhalter et al., 2002) are directly applicable to IRT models. For instance, the relative fit of the 1PLM, 2PLM, and 3PLM can be compared using the LR, AIC, and BIC. These statistics give a global indication of (relative) model fit and do not provide detailed information on the location and possible causes of the lack of model fit. Computation of these local model fit indices as opposed to global model fit indices) needs dedicated IRT software, such as the software used for the present thesis. Due to the large sample sizes and the related power problem, significance probabilities are not considered in this thesis. However, relative comparison of the models will be made using the sizes of the fit statistics.

We use four tests for comparing model fit, the likelihood ratio test (LR), the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the deviance information criterion (DIC).

The likelihood ratio test can be used to compare the fit of two models where one of the models (the null model) is a special case of the other (the alternative model). The test is based on the likelihood ratio which indicates how many times more likely the data are under one model than the other. The likelihood ratio is used to compute a critical value to determine which of the two models to select. For computational purposes the logarithm of the likelihood ratio is taken which does not affect the outcome. The likelihood ratio is computed as follows:

$$LR = -2\ln\left(\frac{\ln L(\text{null model})}{\ln L(\text{alternative model})}\right)$$

where ln$L$(null model) is the log-likelihood of the null model and ln$L$(alternative model) is the log likelihood of the alternative model. The alternative model has more parameters and will therefore have a greater likelihood. Whether it fits better and should be selected is determined by the probability of the difference of the likelihoods. The asymptotic distribution of *LR* is a chi-squared distribution with degrees of freedom equal to *df2-df1* where *df2* and *df1* are the number of free parameters of the null and alternative models, respectively.

The Akaike information criterion (AIC) is similar to *LR*, but includes a penalty that is an increasing function of the number of estimates parameters to discourage overfitting. Given a number of models the preferred model is the one with the minimum AIC value. The AIC is given by

$$AIC = 2k - 2\ln L$$

where ln*L* is the log likelihood under the model and *k* is the number of parameters to be estimated.

The Bayesian information criterion (BIC) is closely related to the AIC but the penalty term is larger in the BIC than in the AIC. It is given by:

$$BIC = -2\ln L + k\ln(n)$$

where *n* is the sample size.

Finally, the deviance information criterion (DIC) is a generalization of the AIC and BIC that is particularly useful in the framework of Bayesian estimation using MCMC. Suppose that θ is a vector of model parameters and define the deviance as

$$D(\theta) = -2\ln L + C$$

where *C* is a constant that cancels out when comparing different models. Then

$$DIC = E(D(\theta)) + \text{var}(D(\theta))/2$$

where the expectation and variance are with respect to the distribution of θ, which are computed in the MCMC procedure.

5.2.4 MML estimation of position effects and residual analysis

The estimation equations and expressions for the covariance matrix of the estimates are easily derived using Fisher's identity (Efron, 1977; Louis, 1982; Glas, 1999). The identity plays an important role in the framework of the EM algorithm, which is an algorithm for finding the maximum of a likelihood marginalized over unobserved data. The principle can be summarized as follows. Let $L_o(\lambda)$ be the log-likelihood function of parameters λ given observed data x_o, and let $L_c(\lambda)$ be the log-likelihood function given both observed data x_o and unobserved missing

data x_m. The latter is called the complete data log-likelihood. The interest is in finding expressions for the first-order derivatives of $L_o(\lambda)$, say, the expressions for $L'_o(\lambda)$. Define the first-order derivatives with respect to the complete data log-likelihood as $L'_m(\lambda)$. Then Fisher's identity entails that $L'_o(\lambda)$ is equal to the expectation of $L'_m(\lambda)$ with respect to the posterior distribution of the missing data given the observed data, $p(x_m|x_o;\lambda)$, that is,

$$L'_0(\lambda) \;=\; E(L'_m(\lambda)\,|\,x_0,\lambda) \;=\; \int L'_m(\lambda)\,p(x_m\,|\,x_0,\lambda)\,dx_m$$

To apply this framework to IRT, a very general definition of an IRT model is adopted. Assume an IRT model is defined by the probability of a response pattern x_n, which is a function of student parameters θ_n, and item, location and/or position parameters, which together are denoted by ω. So the IRT model is given by $p(\,x_n \mid \theta_n,\,\omega\,)$. Assume further that the student parameter θ_n has a normal density $N(\theta_n;\,\mu_{c(n)},\,\sigma_{c(n)})$ where $c(n)$ is the country to which student n belongs. The key idea is to view the student parameters θ_n as missing data and all other parameters ω, $\mu_{c(n)}$, and $\sigma^2_{c(n)}$ as structural parameters λ to be estimated. Then the complete data log-likelihood for a student n is

$$L_n = \log p(x_n\,|\,\theta_n,\omega) + \log N(\theta_n;\mu_{c(n)},\sigma^2_{c(n)})\,. \qquad (5.4)$$

The complete-data likelihood equations are easily derived upon recognizing that the complete-data likelihood is an exponential family, and so estimating boils down to equating the sufficient statistic of a parameter to its expectation. So, for instance, first-order complete-data derivative for a person with respect to the mean is given by

$$\frac{\partial L_n}{\partial \mu_c} = \mu_c - \theta_n$$

and the likelihood equation is given by

$$\frac{\partial L_0(\lambda)}{\partial \mu_c} = \sum_{n|c(n)=c} \mu_c - E(\theta_n\,|\,x_n,\lambda) = 0$$

where the expectation is relative to the posterior distribution

$$p(\theta_n\,|\,x_n,\lambda) \propto p(x_n\,|\,\theta_n,\omega)N(\theta_n;\mu_{c(n)},\sigma^2_{c(n)}).$$

This approach can also be used to derive estimation equations for booklet and position parameters and for the definition of residuals. For instance, consider the booklet parameter δ_b. On the person level we have

$$\frac{\partial L_n}{\partial \delta_b} = \sum_{i \in b} X_{ni} - P(\eta_{ni})$$

where $P(\eta_{ni})$ is the probability of a correct response as a function of η_{ni}. Further, the summation $i \in b$ is over all items i in booklet b. Therefore,

$$\frac{\partial L_0(\lambda)}{\partial \delta_b} = \sum_{n|b(n)=b} \sum_{i \in b} X_{ni} - E(P(\eta_{ni}) \mid x_n, \lambda) = 0. \qquad (5.5)$$

The first summation is over all students administers booklet b, while the second summation is over all items in booklet b.

The approach is easily generalized to other models. Consider the parameter $\xi_{b,p}$ in the 2PLMP. The likelihood equation is given by

$$\frac{\partial L_0(\lambda)}{\partial \xi_{b,p}} = \sum_{n|b(n)=b} \sum_{i \in b,p} X_{ni} - E(P(\eta_{ni}) \mid x_n, \lambda) = 0, \qquad (5.6)$$

where the summation $i \in b,p$ is over all items i that are in position p in booklet b.

Note that the estimation equations (5.5) and (5.6) are differences between observations and their posterior expectations. Therefore, the statistical theory outlined above can also be used to construct a residual analysis. Glas (1999) shows that tests for various assumptions underlying IRT models, such as subgroup invariance of parameters, unidimensionality and local independence, can be evaluated using the Lagrange multiplier (*LM*) test (Rao, 1947, Aitchison & Silvey, 1958). This proceeds as follows. Let $\boldsymbol{\eta}_1$ be a vector of the parameters of some IRT model, and let $\boldsymbol{\eta}_2$ be a vector of parameters added to this IRT model to obtain a more general model. Let $\boldsymbol{h}(\boldsymbol{\eta}_1)$ and $\boldsymbol{h}(\boldsymbol{\eta}_2)$ be the first-order derivatives of the log-likelihood function. The parameters $\boldsymbol{\eta}_1$ of the IRT model are estimated by maximum likelihood, so $\boldsymbol{h}(\boldsymbol{\eta}_1) = \mathbf{0}$. The hypothesis $\boldsymbol{\eta}_2 = \mathbf{0}$ can be tested using the statistic

$$LM = \boldsymbol{h}(\boldsymbol{\eta}_2)^t \, \boldsymbol{\Sigma}^{-1} \, \boldsymbol{h}(\boldsymbol{\eta}_2), \tag{5.7}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{h}(\boldsymbol{\eta}_2)$. It can be proved that the LM-statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters in $\boldsymbol{\eta}_2$.

As already noted earlier in this thesis, significance probabilities of formal tests of model fit are not informative in large-scale surveys, because of the power problem, that is, due to the enormous sample size, all tests are significant. However, test statistics can be informative when comparing the relative fit of different models with each other. Below, we will use a test based on residuals derived as follows. The score range is partitioned into G subsets to form subgroups that are homogeneous with respect to θ. In the present application, we will use the proportion of correct responses given by a student as a proxy for θ. Then, for any of the models outlined above, define

$$\eta_{ni}^{*} = \eta_{ni} - \varsigma_{c(n)ig(n)},$$

where $g(n)$ is the subgroup to which student n belongs. Below, we use 3 subgroups, that is, $G = 3$. The subgroups are formed in such a way that they have approximately the same number of students. The parameter $\zeta_{c(n)ig(n)}$ gauges the shift in difficulty of item i in country $c(n)$ in subgroup $g(n)$. The first subgroup is used as a base-line, the shift parameter of subgroup one is fixed to zero, that is, $\zeta_{c(n)i1} = 0$. An *LM* test based on this parametrization for the general model targets the extent to which the response probabilities $P(\eta_{ni})$ are adequate in all three subgroups. The associated first order derivative for an item i in a country c in a subgroup g is

$$\sum_{g=1}^{G} \sum_{n|c(n)=c, g(n)=g} X_{ni} - E(P(\eta_{ni}) \mid x_n, \lambda). \tag{5.8}$$

For every subgroup g, these differences can be and can be inserted into (5.7) to define an *LM* test statistic, either focused on an item, a country or aggregated over all items and countries to obtain an overall test statistic. Further, to obtain interpretable residuals, the absolute differences can be divided by sample size, summed and divided by the number of subgroups. These residuals then give an indication of the average distance between the observed and expected average score in the subgroups.

Table 5.2. Example of observed and expected responses for a booklet within a country

Subgroup		1		2		3		
Item	Country	Obs	Exp	Obs	Exp	Obs	Exp	Res
1	1	0.58	0.59	0.83	0.83	0.94	0.93	0.01
	2	0.61	0.61	0.85	0.86	0.97	0.95	0.01
	3	0.62	0.65	0.86	0.86	0.97	0.94	0.02
2	1	0.17	0.22	0.43	0.47	0.80	0.71	0.06
	2	0.26	0.30	0.64	0.63	0.88	0.85	0.03
	3	0.27	0.31	0.64	0.63	0.87	0.84	0.03
3	1	0.36	0.41	0.70	0.67	0.86	0.84	0.04
	2	0.43	0.46	0.81	0.80	0.95	0.93	0.02
	3	0.45	0.47	0.83	0.80	0.92	0.93	0.02

A small example of the output of the procedure is given in Table 5.2. The example pertains to 3 arbitrary items and three arbitrary countries of the PISA 2009 data set. The columns labeled 'Obs' and 'Exp' give the observed and posterior expected average scores in the three subgroups and the columns labeled 'Res' givens the associated residual. Item 2 in country 1 produces the highest residual, so this is the least fitting combination of an item with a country. Inspection of the differences between the observed and expected values shows that this difference is largely due to the middle scoring group. Aggregating the residuals over items can give an indication of the appropriateness of the IRT model for a country, while, analogously, aggregating the residuals over countries can give an indication of the fit of the IRT model for a specific item.

5.2.5 Bayesian estimation of position effects and latent residuals

In the past two decades, Bayesian statistical approaches have received considerable attention as an alternative for likelihood-based approaches to estimating and testing IRT models. The reason is that MML computations become intractable for highly-dimensional IRT models. Examples of the Bayesian approach to IRT are the articles by Albert (1992, estimation of the 2PLM), Patz and Junker (1999, models for rating data), Bradlow, Wainer, and Wang (1999, testlet model), Janssen, Tuerlinckx, Meulders, and De Boeck (2000, random item parameters), Béguin and

Glas (2001, multidimensional IRT), Fox and Glas (2001, 2003, multilevel IRT), and Geerlings, Glas and van der Linden (2011, item family models).

In a Bayesian approach, the likelihood function (the logarithm of the likelihood function for a person is given in (5.4)) is multiplied by a prior distribution to obtain a posterior distribution given the data. The prior may be based on a priori ideas or information about the parameters, but may also be non-informative. A Markov chain Monte Carlo (MCMC) procedure is used to generate values from the posterior distribution to produce an estimate. In most articles cited above the MCMC chains are be constructed using the Gibbs sampler (Gelfand & Smith, 1990). To implement the Gibbs sampler, the parameter vector is divided into a number of components, and each successive component is sampled from its conditional distribution given sampled values for all other components. This sampling scheme is repeated until the sampled values form stable posterior distributions. However, nowadays the Bayesian estimates of posterior distributions computed from MCMC chains can also be computed with general-purpose software such as WinBugs and OpenBugs (Lunn, Spiegelhalter, Thomas & Best, 2009) and JAGS (Plummer, 2003). The computation presented below were made using WinBugs. The priors where standard normal for the item difficulties, and log-normal with expectation 1.0 and variance 1.0. The country means had a standard normal prior, their variances a gamma prior with expectation 1.0 and variance 5.0. Booklet and position effects had a standard normal prior. Interactions with country effects has a standard normal prior for the mean and a gamma prior with expectation 1.0 and variance 5.0 for their variances. Burn-in iterations were set to 4,000 and 10,000 iterations were made for the actual estimation of the parameter distributions. All point estimates reported are posterior expectations.

Bayesian residuals were estimated for examining model fit. Bayesian residuals are random parameters with unknown values. They can be estimated from the data along with their variance in an MCMC procedure. Model fit can be ascertained after computing summary statistics of the posterior distributions of the residuals. Usually in a residual analysis, residuals are transformed such that they approximately follow a normal distribution but for discreet observations these transformations result in poor approximation by the normal distribution

Albert and Chib (1995) and Johnson and Albert (1999) introduced Bayesian latent residuals as an alternative to the Bayesian residuals. In Bayesian residual analysis the data is augmented such that a regression of the augmented variable on the latent ability is linear and with the same error variance for all abilities. The difference between the latent response and the expected

response is defined as a Bayesian latent residual. A number of identifying assumptions are imposed on the data augmentation scheme due to which each latent residual is standard normally distributed.

In this study, statistics are used that are based on Bayesian latent residuals (Fox, 2010). Applied to the models defined above, Bayesian latent residuals take the form

$$\varepsilon_{ni} = Z_{ni} - \eta_{ni}, \tag{5.9}$$

where η_{ni} is defined in the formulas (5.1), (5.2) and (5.3). These Bayesian latent residuals can be sampled as the mean of computed residual values in each MCMC iteration. The normal-ogive representation is best suited for this purpose. It can be shown (see, for instance, Fox, 2010) that

$$E(\varepsilon_{ni} \mid X_{ni} = 1, \eta_{ni}) = \frac{\phi(\eta_{ni})}{\Phi(\eta_{ni})},$$

and

$$E(\varepsilon_{ni} \mid X_{ni} = 0, \eta_{ni}) = \frac{-\phi(\eta_{ni})}{\Phi(\eta_{ni})},$$

where the nominator on the right-hand side of these two formulas is the standard normal density function and the denominator is the normal-ogive probability of a correct response.

To create residuals that are analogues of the MML residuals, their absolute values are summed over the same indices as the analogous MML residuals, that is, analogous to the summation in (5.8).

5.3 Results

5.3.1 Estimates of order effects

The results from the estimation of the models with booklet parameters are presented in the tables 5.3 and 5.4.

Table 5.3 gives the results for the 1PLM estimated in both MML and Bayesian frameworks. The columns labeled *Booklet Parameter* give the estimated value of the parameters $\delta_{b(n)}$. For the Bayesian estimation procedure, the point estimates are the expected posterior values. The columns labeled *Mean* and *Sd* give the mean and standard deviation over countries of the parameters $\delta_{c,b(n)}$.

Overall, there was a correlation of about 0.85 between the official PISA 2009 booklet effects and those computed using the MML and Bayesian estimates, so the agreement between these estimates was good.

Looking at the standard deviations of the interaction parameters $\delta_{c,b(n)}$, we see that the standard deviations for all the booklets was close. For the country-by-booklet interactions, there were no countries which had outliers for more than one or two booklets. The largest interaction effect was for country 31 (GBR) for booklet 12 and it was 1.12 points on the logit scale. There were no countries with an interaction effect size above 0.90 on more than one booklet. However, as far as booklets were concerned booklets 12 and 5 had the highest interaction effects sizes over different countries. When using a cut of point value of 0.80 on the logit scale the frequency of occurrence for large effect sizes for booklets 12 and 5 were about 34% and 25% of the total. When using a cut of point value of 0.60 points on the logit scale, the frequency of occurrence for semi-large effect sizes for booklets 12 and 5 were approximately 27% and 25 % of the total. So in general, all booklet parameters differed across countries, but there were few interaction effects that could be earmarked as outliers.

Table 5.3. Booklet effects for PISA 2009 reading dataset, the PISA 2009 official values for the one-parameter models

Booklet	PISA 2009 MML	PISA 2009 (rescaled)	MML Estimation Booklet Parameter (1PLMB)	Country-by-booklet parameter (1PLMCB) Mean	Sd	Bayesian Estimation Booklet Parameter (1PLMB)	Country-by-booklet parameter (1PLMCB) Mean	Sd
1	-0.04	-0.30	-0.22	-0.22	0.72	-0.19	-0.09	0.76
2	0.07	-0.19	-0.35	-0.35	0.83	-0.32	-0.21	0.82
3	-0.05	-0.31	-0.32	-0.32	0.74	-0.29	-0.17	0.79
4	0.03	-0.23	-0.19	-0.19	0.73	-0.17	-0.08	0.77
5	-0.18	-0.44	-0.57	-0.57	0.81	-0.56	-0.33	0.81
6	-0.06	-0.32	-0.42	-0.42	0.74	-0.40	-0.25	0.77
7	0.10	-0.16	-0.28	-0.28	0.76	-0.29	-0.21	0.76
8	0.12	-0.14	-0.10	-0.10	0.78	-0.15	-0.10	0.75
9	0.25	-0.01	-0.06	-0.06	0.72	-0.13	-0.09	0.77
10	-0.09	-0.35	-0.20	-0.20	0.73	-0.26	-0.11	0.77
11	-0.03	-0.29	-0.10	-0.10	0.75	-0.22	-0.10	0.76
12	-0.39	-0.65	-0.47	-0.47	0.76	-0.52	-0.43	0.78
13	0.26	0.00	0.00	0.00	0.69	0.00	0.00	0.75

Table 5.4. Booklet effects for PISA 2009 reading dataset, the PISA 2009 official values for the two-parameter models

Booklet	PISA 2009 MML 1PLM	PISA 2009 (rescaled)	MML Estimation Booklet Parameter (2PLMB)	Country-by-booklet parameter (2PLMCB) Mean	Sd	Bayesian Estimation Booklet Parameter (2PLMB)	Country-by-booklet parameter (2PLMCB) Mean	Sd
1	-0.04	-0.30	-0.18	-0.16	0.71	-0.06	-0.11	0.79
2	0.07	-0.19	-0.27	-0.20	0.77	-0.05	-0.17	0.81
3	-0.05	-0.31	-0.26	-0.21	0.71	-0.01	-0.19	0.77
4	0.03	-0.23	-0.17	-0.17	0.71	-0.16	-0.07	0.74
5	-0.18	-0.44	-0.45	-0.45	0.78	-0.14	-0.24	0.82
6	-0.06	-0.32	-0.35	-0.33	0.76	-0.14	-0.27	0.74
7	0.10	-0.16	-0.19	-0.19	0.70	-0.23	-0.35	0.75
8	0.12	-0.14	-0.10	-0.05	0.69	-0.15	-0.01	0.75
9	0.25	-0.01	-0.03	0.00	0.73	-0.01	-0.12	0.75
10	-0.09	-0.35	-0.16	-0.14	0.71	-0.00	-0.12	0.77
11	-0.03	-0.29	-0.11	-0.15	0.68	-0.14	0.04	0.76
12	-0.39	-0.65	-0.34	-0.40	0.75	-0.23	-0.36	0.81
13	0.26	0.00	0.00	0.00	0.73	0.00	0.00	0.73

Table 5.5 and Table 5.6 give MML and Bayesian estimates of the position effects $\xi_{b(n),p(i,b(n))}$, under the one- and two-parameter models, respectively. In both tables, the agreement between the MML and Bayesian estimates is very high. For the 1PLMP in Table 5.5 it can be seen that

the position effects for different booklets do not have the same trend. As represented in the estimation model describer earlier, negative values for the position effects indicate that the items are more difficult when they appear in that position. Or, to put it in another way, negative values can be viewed as lowering the ability parameter and thus lowering the probability of a correct response. The scale has been identified by fixing the position effect for the first position in every booklet to be the same. In Table 5.5 for the 1PLMP, for the Bayesian estimates it can be seen that for booklets 4, 6,7,8,9 and 13 the items appear to be more difficult when they appear at the end of the booklet. For booklet 12 there is no position effect, for booklets 2,3 and 10 and 11 the impact of the position effect is not systematic over the four positions.

Table 5.5. Position effects for PISA 2009 reading dataset, the PISA 2009 official values and results for the one-parameter model 1PLMP

	MML Estimation Position			Bayesian Estimation Position		
Booklet	2	3	4	2	3	4
1	-0.02	-0.29	-0.01	-0.01	-0.30	-0.00
2	-0.17	-0.20	-0.01	-0.15	-0.24	0.00
3	-0.24	-0.01	0.02	-0.21	-0.00	0.01
4	0.09	0.01	-0.53	0.08	0.00	-0.49
5	-0.00	-0.37	-0.02	-0.00	-0.36	0.00
6	0.04	0.05	-0.33	0.04	0.05	-0.30
7	0.00	0.00	-0.56	0.00	0.00	-0.53
8	0.00	0.00	-0.42	0.00	-0.00	-0.40
9	-0.00	-0.14	-0.20	-0.00	-0.12	-0.18
10	-0.36	-0.00	0.00	-0.36	-0.00	0.00
11	0.30	-0.38	0.00	0.31	-0.36	0.00
12	-0.00	0.00	0.00	-0.00	0.00	0.00
13	-0.13	-0.15	-0.68	-0.11	-0.15	-0.67

Table 5.5 shows that the results of the position effects for the 2PLMP are very similar to the results for the 1PLMCP and that the difficult levels for different booklets do not have the same trend vis-à-vis the position effects. It can be seen that for booklets 4, 6,7,8,9 and 13 the items appear to be more difficult when they appear at the end of the booklet. For booklet 12 there is no position effect, for booklets 2,3 and 10 and 11 the impact of the position effect is not systematic over the four positions.

The agreement between the 1PLMP and 2PLMP estimates was high, always above 0.95. So in this respect, adding discrimination parameters to the IRT model made little difference.

Table 5.6. Position effects for PISA 2009 reading dataset, the PISA 2009 official values and results for the two-parameter model 2PLMP

	MML Estimation Position			Bayesian Estimation Position		
Booklet	2	3	4	2	3	4
1	-0.05	-0.33	-0.06	-0.04	-0.32	-0.01
2	-0.19	-0.33	0.05	-0.19	-0.33	0.00
3	-0.20	0.00	-0.05	-0.20	0.00	-0.00
4	0.03	0.00	-0.45	0.03	0.00	-0.45
5	-0.00	-0.31	-0.00	-0.00	-0.34	-0.00
6	0.01	0.07	-0.32	0.01	0.07	-0.31
7	0.00	-0.07	-0.61	0.00	-0.00	-0.58
8	-0.00	0.00	-0.51	-0.00	0.00	-0.49
9	-0.00	-0.18	-0.18	-0.00	-0.16	-0.18
10	-0.39	0.00	0.00	-0.39	0.00	0.00
11	0.23	-0.35	-0.03	0.23	-0.38	-0.00
12	-0.00	0.00	-0.04	-0.00	0.00	0.00
13	-0.09	-0.16	-0.71	-0.09	-0.19	-0.72

Table 5.7 and Table 5.8 give the Bayesian estimates of the means and standard deviations of the country-by-position effects $\xi_{c,b(n),p(i,b(n))}$, under the one- and two-parameter models, respectively. The means and standard deviations are taken over countries. The MML estimates are not shown, because, as above, they are highly similar to the Bayesian estimates.

Table 5.7. Country-by-Position interaction for PISA 2009 reading dataset, one-parameter model 1PLMCP

	Bayesian Estimation Mean Position			SD Position		
Booklet	2	3	4	2	3	4
1	-0.07	-0.15	-0.00	0.71	0.72	1.06
2	-0.11	-0.15	0.00	0.71	0.69	1.05
3	-0.06	0.00	0.00	0.73	1.05	1.06
4	0.00	-0.00	-0.28	0.68	1.05	0.69
5	-0.00	-0.18	0.00	1.05	0.69	1.05
6	0.00	-0.04	-0.14	0.69	0.71	0.70
7	-0.00	0.00	-0.27	1.05	1.05	0.71
8	-0.00	-0.00	-0.23	1.04	1.06	0.68
9	0.00	-0.10	-0.15	1.05	0.71	0.71
10	-0.18	0.00	0.00	0.71	1.05	1.06
11	0.06	-0.22	0.00	0.71	0.71	1.05
12	-0.00	-0.00	0.00	1.05	1.05	1.05
13	-0.07	-0.16	-0.36	0.70	0.71	0.69

From Table 5.7 it can be seen that the mean values for the country-by-position effects for the 1PLCMP correspond nicely with the fixed positon effects in Table 5.5; the correlation is 0.96. However, the magnitude of the estimates is slightly smaller. From the standard deviations we can see that the country-by-position interaction is different across different booklets. For the country-by-position interactions, there were twenty six cases only where the absolute value of the interaction effect was larger than 1.0 points on the logit scale and this occurred for 19 different countries. The maximum frequency of occurrence was two times only for 7 different countries. The number of times this happened for a booklet was four times and this was for booklets 7, 8, 9 and 10. The position location for which this occurred most frequently was position 3 (for which it occurred 11 times). When using a cutoff value of 5.0 points on the logit scale there was only one country (New Zealand) for which this happened more than once (it occurred two times). For booklets 10 and 8, this happened thrice. In terms of the position location it occurred 8 times for position 3, indicating that there are more outliers in position 3.

Table 5.8. Country-by-Position interaction for PISA 2009 reading dataset, two-parameter model 2PLMCP

	Bayesian Estimation Mean Position			SD Position		
Booklet	2	3	4	2	3	4
1	-0.08	-0.16	0.01	0.70	0.71	1.05
2	-0.13	-0.17	0.00	0.71	0.69	1.05
3	-0.05	-0.00	0.00	0.71	1.05	1.06
4	0.01	-0.01	-0.26	0.69	1.05	0.69
5	-0.01	-0.19	0.00	1.05	0.69	1.05
6	0.00	-0.03	-0.13	0.69	0.71	0.69
7	0.00	0.01	-0.25	1.05	1.05	0.71
8	-0.01	-0.01	-0.28	1.05	1.05	0.68
9	0.01	-0.09	-0.11	1.05	0.69	0.68
10	-0.21	0.00	0.00	0.71	1.05	1.05
11	0.08	-0.19	-0.00	0.71	0.70	1.05
12	0.01	-0.00	0.00	1.05	1.05	1.05
13	-0.02	-0.12	-0.36	0.69	0.70	0.68

From Table 5.8 it can be seen that the mean values for the country-by-position effects for the 2PLCMP again correspond nicely with the fixed positon effects in Table 5.6. As with the 1PLMP and 1PLMCP, the correlation was high (0.97) but the magnitude of the estimates was smaller. From the standard deviations we can see that the country-by-position interaction is different across different booklets. There were 28 cases where the absolute value of the

interaction effect was larger than 1.0 points on the logit scale and this occurred for 18 different countries. The maximum frequency of occurrence was 3 times only for a single country. The number of times this happened for a booklet was 6 times and this was for booklet 8. For the position locations it occurred almost as frequently for the three different positions. When using a cutoff value of 5.0 points on the logit scale there were four countries for it happened more than once (it occurred two times). The maximum number of times it happened for a booklet was 3 times for booklet 7. In terms of the position location the maximum occurrence was 7 times which happened at position 4.

5.3.2 Ordering of PISA countries

To evaluate the impact of including booklet and position effects, the rank ordering of the countries' mean latent trait values was estimated using the two measurement models with the booklet effect variations. The rank ordering of countries is of high practical importance for PISA. Especially the media and politicians have a great interest in them. As an example we present the mean latent traits for the 1PLM, 2PLM, 1PLMB and 2PLMB estimated in both the MML and Bayesian frameworks, in the tables 5.9 and 5.10, respectively.

Table 5.9. Rank order and mean scale level of countries on the PISA Reading scale for the 1PLM and 2PLM models with and without booklet parameters for the MML estimates. The column labeled PISA lists the ordering as in the PISA 2009 International report.

	Rank Order					Mean on Latent Scale			
Country	PISA 1PLMB	1PLM	2PLM	1PLMB	2PLMB	1PLM	2PLM	1PLMB	2PLMB
AUS	9	12	12	12	13	.1331	.1172	-.0376	-.0121
AUT	24	25	24	25	24	-.1639	-.1337	-.3379	-.2697
BEL	4	4	4	4	4	.3477	.3001	.1677	.1629
CAN	7	7	7	6	6	.2676	.2367	.0974	.1086
CHL	30	30	30	30	30	-.5428	-.4542	-.6496	-.5345
CZE	13	15	13	15	14	.1013	.1159	-.0703	-.0157
DNK	26	26	26	26	26	-.2216	-.1873	-.4044	-.3343
FIN	2	2	2	2	2	.5210	.4506	.3454	.3185
FRA	12	9	9	9	9	.1906	.1800	.0071	.0387
DEU	16	21	20	20	20	-.0347	-.0214	-.2021	-.1520
GRC	18	16	17	17	17	.0504	.0296	-.1219	-.1023
HUN	10	14	15	14	16	.1048	.0953	-.0693	-.0404
ISL	15	13	14	16	15	.1074	.1029	-.0720	-.0368
IRL	14	11	11	13	12	.1349	.1334	-.0378	0.000
ITA	22	18	16	19	18	.0242	.0319	-.1542	-.1082
JPN	3	3	3	3	3	.4232	.3516	.2451	.2164
KOR	1	1	1	1	1	.6329	.5135	.4515	.3785
LUX	29	27	27	27	27	-.2495	-.1896	-.4342	-.3383
MEX	31	31	31	31	31	-.7574	-.6379	-.8695	-.7236
NLD	5	6	6	7	7	.2759	.2442	.0942	.1051
NZL	6	5	5	5	5	.3242	.2822	.1507	.1498
NOR	11	10	10	11	10	.1528	.1364	-.0157	.0064
POL	8	8	8	8	8	.2108	.2012	.0357	.0681
PRT	23	20	21	21	22	-.0238	-.0215	-.2052	-.1632
SVK	27	28	28	28	28	-.2801	-.2283	-.4547	-.3689
ESP	25	24	25	24	25	-.1572	-.1353	-.3300	-.2711
SWE	20	17	18	18	19	.0286	.0251	-.1456	-.1114
CHE	21	23	23	23	23	-.0515	-.0316	-.2282	-.1691
TUR	28	29	29	29	29	-.3251	-.2762	-.5022	-.4205
GBR	19	22	22	22	21	-.0367	-.0231	-.2141	-.1627
USA	17	19	19	10	11	.0000	.0000	.0000	.0000

It can be seen from Table 5.9 that the rank ordering of the countries with respect to the latent trait means evaluated using the four different measurement models is quite consistent. The rank correlations between country means of latent distributions for various models using the MML estimates will be given in Table 5.11.

Results from the Bayesian analysis in Table 5.10 show that the rank ordering of the countries with respect to the mean of the latent traits is consistent for all measurement models and also consistent with the rank ordering results obtained using the MML estimates.

Table 5.10. Rank order and mean scale level of countries on the PISA Reading scale for the 1PLM and 2PLM type models with and without booklet parameters for Bayesian estimates. The column labeled PISA lists the ordering as in the PISA 2009 International report.

Rank Order						Mean on Latent Scale			
Country	PISA 1PLMB	1PLM	2PLM	1PLMB	2PLMB	1PLM	2PLM	1PLMB	2PLMB
AUS	9	11	12	11	12	.2664	.1706	.1773	.3141
AUT	24	23	24	25	22	.0888	-.0667	-.0007	.0757
BEL	4	4	4	4	4	. 3915	.3364	.3009	.4788
CAN	7	7	7	7	7	.3424	.2814	.2602	.4305
CHL	30	30	30	30	30	-.1395	-.3633	-.2068	-.1926
CZE	13	15	13	15	13	.2449	.1625	.1634	.3117
DNK	26	25	26	26	24	.0531	-.1144	-.0414	.0206
FIN	2	2	2	2	2	.4926	.4801	.4102	.6285
FRA	12	9	9	9	9	.2962	.2216	.2064	.3629
DEU	16	20	20	19	26	.1675	.0381	.0837	.0000
GRC	18	16	16	16	17	.2082	.0891	.1232	.2304
HUN	10	14	15	13	15	.2520	.1518	.1670	.2938
ISL	15	13	14	14	14	.2556	.1600	.1637	.2985
IRL	14	12	11	12	11	.2634	.1844	.1761	.3292
ITA	22	17	17	17	16	.1987	.0859	.1102	.2311
JPN	3	3	3	3	3	.4372	.3724	.3436	.5197
KOR	1	1	1	1	1	.5703	.5416	.4791	.6887
LUX	29	26	27	27	25	.0350	-.1217	-.0544	.0111
MEX	31	31	31	31	31	-.2692	-.5393	-.3399	-.3725
NLD	5	6	6	6	6	.3512	.2877	.2610	.4306
NZL	6	5	5	5	5	.3801	.3253	.2893	.4748
NOR	11	10	10	10	10	.2779	.1862	.1957	.3370
POL	8	8	8	8	8	.3080	.2489	.2206	.3952
PRT	23	19	19	20	20	.1693	.0398	.0807	.1792
SVK	27	27	28	28	28	.0214	-.1572	-.0677	-.0197
ESP	25	24	25	24	23	.0857	-.0672	-.0011	.0754
SWE	20	18	18	18	18	.1961	.0825	.1084	.2277
CHE	21	22	22	22	21	.1584	.0316	.0709	.1781
TUR	28	29	29	29	29	-.0111	-.1998	-.0996	-.0684
GBR	19	21	21	21	19	.1624	.0351	.0718	.1814
USA	17	28	23	23	27	.0000	.0000	.0000	.0000

The rank correlations between country means of latent distributions for various models using the Bayesian estimates are given in Table 5.11. The rank correlations obtained using MML are not displayed, because they are virtually identical. From the tables it can be see that the correlations between the means are lower for the two-parameter models, especially when the correlation involves a model with country-by-booklet parameters. The conclusion is somewhat unclear. The two-parameter models have more parameters than the one-parameter models, so the results obtained using the latter models have a higher credibility. However, the 2PLMCB correlates lower with the other two-parameter rankings, which are virtually identical. Though the practical implication of the difference may be small, that is, the correlations are in the 0.80-0.90 range, it is still of interest whether the 2PLMCB produces a more precise ranking than the

others. It might be the case that the sheer number of added parameters deters the precision of the outcomes. The evaluation of model fit presented in the next sections is meant to investigate this question.

Table 5.11. Rank correlations between country means of latent distributions computed using the Bayesian method.

	1PLM	1PLMB	1PLMCB	1PLMP	1PLMCP
1PLM	1				
1PLMB	0.99	1			
1PLMCB	0.99	0.99	1		
1PLMP	0.99	0.99	0.99	1	
1PLMCP	0.96	0.96	0.94	0.97	1
	2PLM	2PLMB	2PLMCB	2PLMP	2PLMCP
2PLM	1				
2PLMB	0.99	1			
2PLMCB	0.84	0.85	1		
2PLMP	0.99	0.99	0.83	1	
2PLMCP	0.96	0.96	0.85	0.96	1
	1PLM	1PLMB	1PLMCB	1PLMP	1PLMCP
2PLM	0.99	0.99	0.99	0.99	0.96
2PLMB	0.99	0.99	0.99	0.99	0.96
2PLMCB	0.81	0.82	0.82	0.79	0.74
2PLMP	0.99	0.99	0.99	0.99	0.96
2PLMCP	0.95	0.95	0.93	0.95	0.97

5.3.3 Residual analyses

MML residuals were computed using formula (5.4) with 3 subgroups formed according to score level. Residuals at the country level were computed by summing over the items, while residuals at the item level were computed by summing over countries.

The results for the country level residuals are given in Table 5.12, for the 1PLM , the 2PLM, the 1PLMB and the 1PLMP. Note that there was little difference between the 1PLM and the 2PLM. Chili fitted best, while the Netherlands had the worst model fit. However, a difference between 0.08 and 0.15 is not very large. Adding booklet parameters generally reduced the residuals, while adding position parameters reduced the residuals further. So in this analysis, the position parameters did a slightly better job than the booklet parameters. Again, the residuals under the 1PLMB and 2PLMB (not shown), and the 1PLMP and 2PLMP (not shown) were virtually identical. So here, the more parsimonious one-parameter models are preferable. Also the model with country-by-booklet and country-by-position effects did not result in a notable improvement.

The results for the item-oriented residuals for the 1PLM and 1PLMP are given in Table 5.13. For reasons of space, the fit of the other models is not displayed, but the pattern was the same as for the country-oriented residuals. So the 1PLMP was the most parsimonious best fitting model in this analysis. The residuals for all the items were very similar, also within the countries. However, in some specific countries, for the items with large residuals, introduction of the 1PLMP did impact the size of the residuals for that item but such cases were very few. While analyzing the residuals per country it emerged that large residual values are spread across countries for the various items, however there were some countries like Turkey and Mexico that contributed regularly to high residual values for certain items.

Table 5.12. Average absolute differences between observed and expected values over all items for each OECD country for four measurement models

Country	1PLM	2PLM	1PLMB	1PLMP
AUS	0.12	0.12	0.09	0.07
AUT	0.11	0.11	0.08	0.07
BEL	0.12	0.12	0.09	0.05
CAN	0.11	0.12	0.10	0.07
CHL	0.08	0.09	0.07	0.07
CZE	0.11	0.11	0.12	0.12
DNK	0.12	0.12	0.09	0.09
FIN	0.12	0.12	0.10	0.08
FRA	0.11	0.11	0.08	0.07
DEU	0.12	0.13	0.10	0.08
GRC	0.11	0.11	0.09	0.07
HUN	0.12	0.12	0.08	0.05
ISL	0.11	0.12	0.11	0.09
IRL	0.12	0.13	0.10	0.11
ITA	0.11	0.11	0.11	0.10
JPN	0.14	0.14	0.11	0.09
KOR	0.14	0.14	0.13	0.11
LUX	0.12	0.12	0.09	0.07
MEX	0.10	0.10	0.11	0.07
NLD	0.15	0.15	0.13	0.11
NZL	0.11	0.11	0.11	0.09
NOR	0.11	0.11	0.11	0.09
POL	0.12	0.12	0.11	0.10
PRT	0.11	0.11	0.08	0.04
SVK	0.12	0.12	0.12	0.08
ESP	0.11	0.11	0.13	0.09
SWE	0.12	0.13	0.11	0.08
CHE	0.11	0.11	0.10	0.06
TUR	0.14	0.14	0.15	0.14
GBR	0.12	0.12	0.10	0.07
USA	0.11	0.12	0.10	0.06

Table 5.13. Average absolute differences between observed and expected values for each item over all countries, all booklets, all subgroups.

	Residual			Residual			Residual	
Item	1PLM	1PLMP		1PLM	1PLMP	Item	1PLM	1PLMP
1	0.10	0.12	37	0.12	0.12	73	0.11	0.11
2	0.14	0.05	38	0.14	0.06	74	0.12	0.13
3	0.13	0.09	39	0.12	0.10	75	0.10	0.06
4	0.09	0.05	40	0.09	0.08	76	0.13	0.12
5	0.08	0.10	41	0.13	0.06	77	0.13	0.07
6	0.10	0.11	42	0.17	0.09	78	0.11	0.11
7	0.10	0.04	43	0.13	0.11	79	0.13	0.10
8	0.09	0.12	44	0.11	0.08	80	0.10	0.07
9	0.12	0.07	45	0.10	0.08	81	0.13	0.06
10	0.15	0.09	46	0.12	0.09	82	0.13	0.12
11	0.09	0.09	47	0.13	0.12	83	0.05	0.11
12	0.11	0.08	48	0.10	0.12	84	0.09	0.08
13	0.11	0.06	49	0.13	0.07	85	0.09	0.16
14	0.13	0.09	50	0.11	0.12	86	0.12	0.09
15	0.11	0.10	51	0.12	0.09	87	0.14	0.08
16	0.16	0.11	52	0.13	0.08	88	0.14	0.09
17	0.11	0.08	53	0.10	0.07	89	0.12	0.10
18	0.10	0.09	54	0.09	0.10	90	0.12	0.08
19	0.14	0.10	55	0.12	0.10	91	0.14	0.09
20	0.13	0.11	56	0.07	0.09	92	0.12	0.10
21	0.10	0.10	57	0.15	0.11	93	0.07	0.11
22	0.11	0.07	58	0.17	0.13	94	0.10	0.09
23	0.12	0.08	59	0.08	0.10	95	0.12	0.09
24	0.14	0.09	60	0.14	0.08	96	0.17	0.12
25	0.08	0.09	61	0.09	0.07	97	0.11	0.09
26	0.13	0.06	62	0.10	0.06	98	0.15	0.09
27	0.14	0.10	63	0.08	0.08	99	0.13	0.10
28	0.12	0.09	64	0.13	0.11	100	0.17	0.07
29	0.10	0.08	65	0.07	0.09	101	0.13	0.10
30	0.10	0.07	66	0.13	0.12	102	0.07	0.07
31	0.10	0.08	67	0.11	0.05	103	0.14	0.10
32	0.12	0.05	68	0.12	0.08	104	0.17	0.09
33	0.12	0.13	69	0.09	0.07	105	0.10	0.08
34	0.13	0.07	70	0.14	0.07	106	0.14	0.08
35	0.11	0.05	71	0.09	0.11	107	0.12	0.08
36	0.12	0.09	72	0.11	0.07	108	0.08	0.10

5.3.4 Global model fit

The Table 5.14 below lists the results from the three tests for model fit the different IRT measurement models described earlier.

Table 5.14. Global fit to IRT models

Null Model	Alternative	Df	LR	ΔAIC	ΔBIC	ΔLM	ΔDIC
1PLM	2PLM	107	4396*	4182	3948	1116*	4130
1PLM	1PLMB	12	3644*	3620	3594	357*	3622
1PLMB	1PLMCB	360	198	-522	-1311	266	122
1PLM	1PLMP	39	5042*	4964	4879	1219*	4918
1PLMP	1PLMCP	1170	224	-2116	-4679	282	199
2PLM	2PLMB	12	3522*	3498	3472	238*	3498
2PLMB	2PLMCB	360	124	-596	-1385	274	109
2PLM	2PLMP	39	3734*	3656	3571	1239*	3645
2PLMP	2PLMCP	1170	88	-2428	-4991	247	90
1PLMP	2PLMP	107	3088*	2874	2640	1181*	2878

The significant effects at 1% are flagged by a *.

As noted on several other places in this thesis, formal tests of model fit and the associated significance probabilities are not very informative due to the large sample size leading to excessive power. That is, even the smallest insignificant model violation leads to a rejection of the model. Still, the statistics that go with the tests are generally very well founded, and comparing the values of the statistics can provide insight into the relative fit of the various models. The first two columns of Table 5.14 give the null models and alternative models that are compared. The column labeled ‘Df’ gives the difference in the number of parameters between the two models. The column labeled ‘LR’ gives the value of the likelihood-ratio statistic. The columns labeled ‘ΔAIC’ and ‘ΔBIC’ give the difference between the values of the AIC and BIC of the two models, respectively. The column labeled ‘ΔLM’ gives the difference between the two LM statistics computed for the two models. It must be noted that strictly speaking, the asymptotic distribution with the degrees of freedom in column 3 does not hold, because, while the two LM statistics of the two models do have an asymptotic chi-square distribution, their difference does not, because the two statistics are computed on the same data and therefore dependent. LR, ΔAIC, ΔBIC and ΔLM were computed using MML estimates. Finally, the column labeled ‘ΔDIC’ gives the difference between the values of the DIC of the two models, computes using the Bayesian framework implemented in WinBugs. The stars in the table indicate significant outcomes. The AIC and BIC do not provide a test of a model in the sense of testing a null hypothesis; that is, they tell nothing about the quality of the model in an absolute sense. They present a basis for comparing the relative fit of models, where the

model with the highest value is the model of choice. Also the DIC must be interpreted in this way.

Inspection of the first row of Table 5.14 shows that the 2PLM fist the data better than the 1PLM. The rows 2, 4, 6 and 8 show that adding booklet or position parameters leads to a significant improvement according to the LR test; the effect is most prominent when adding position parameters. The same pattern emerges with the ΔLM. Adding interaction effects with countries does not lead to a significant improvement. The ΔAIC and ΔBIC, which penalize the number of parameter and the sample size, even give negative outcomes, which means that the simpler model is preferred.

The last row gives a comparison between the 1PLMP and 2PLMP; the latter model is preferred by all computed indices. So in the analysis of global model fit, the 2PLMP is the model of choice.

5.4 Conclusions

A large scale survey like PISA is a complex undertaking with many practical challenges that need to be addressed. One of the main decisions taken when the PISA study began more than a decade ago was to use IRT as the measurement tool. Besides its theoretical basis, IRT gave flexibility in managing practical issues that a large scale survey entails. IRT separates person and item parameters and thus, for instance, allows the use of an incomplete booklet design like in PISA which allows for a larger number of items to be included in a test. In an IRT analysis a number of different models can be fitted to the data to ascertain the right choice of model for the particular research question. Though an IRT model may provide an adequate description of the test data, it is essential to test the fit of the model to the data. Especially the model should explain aspects of data that impact any inferences made using the model. PISA made the choice to use the 1PLM with a booklet effect for analyzing the cognitive tests. In this chapter, it was our aim to compare the fit of various IRT measurement models to the reading test and see if adding booklet effect to the PISA scaling model was still necessary.

We analyzed various aspects of model-data fit. The results of the various tests for global model fit indicated that the 2PLM fit better than the 1PLM and that the models with booklet and position effects fitted even better, where position effects had a small edge over booklet effects. Models where interaction effects with countries were added did not result in a better model fit. A residual analysis at the item and country level showed that the 1PLM and 2PLM could hardly

be distinguished, while adding booklet and positon parameters had a quite small positive effect. Finally, the effect of model choice on the ordering of countries was completely negligible. So here the 1PLM proved very robust. However, for secondary analyses relating the cognitive outcomes to all sorts of background variables, the effects of lack of model fit is as yet unclear, so there the best fitting model is preferable, which is the 2PLMP.

Finally, there was little difference between the MML and Bayesian procedures, in terms of parameter estimates, global model fit and residual analysis. So a possible change in orientation from likelihood-based to Bayesian methods will not produce remarkable changes in the conclusions of large-scale educational surveys.

Chapter 6

Comparison of Different Approaches to Estimation of Regression Models with Latent Variables

Because of the complexity of using raw data in analyses for many secondary practitioners of educational research, large scale datasets like PISA provide estimates of latent background variables and plausible values for outcome variables that can be directly used in a regression analysis. In this research we compare the results from a regression analysis when using raw data for all the variables in the model with a regression analysis where separately estimated latent estimates of the background variables and plausible values for the outcome variable are used. In the first case, the parameters of the IRT measurement model for the variables and the structural regression model are estimated concurrently. In the second, case a two-step procedure is employed where the measurement model and the regression model are estimated separately. These analyses were done in a fully Bayesian framework for the outcome variable Reading ability and some selected latent covariates/background variables for a set of diverse countries from the PISA 2009 dataset.

6.1 Introduction

International large scale surveys give rise to a large amount of data and what is often of interest to secondary practitioners of educational research is to examine interesting relationships in the data through a regression analysis. Several methods are available to properly estimate the parameters of the regression model on latent outcome variables. One method is to estimate the parameters of the IRT measurement model (item parameters) and the structural model (the regression parameters) concurrently. The disadvantage of such a method is the numerical complexity of the method and the need to use specialized software. An alternative is to separately estimate the values of the latent variables for every respondent and to do the analysis of variance or regression on these estimates. The problem with this approach is that the

variables are not direct observations but estimates with an estimation error. The estimation error is particularly large for outcome variables in large scale assessments because of large amounts of missing data due to a booklet structure in a typical large scale assessment design. In such cases, standard estimation methods such as weighted maximum likelihood (WML, Warm, 1989) and expected-a-posteriori or EAP (Bock & Mislevy, 1982) which are point estimates for individual students result in biased estimates of regression coefficients and their standard errors. A solution to this problem is to use multiple draws from the posterior distributions of the latent variables, known as plausible values. The variance of these draws accounts for the uncertainty in the estimates to provide unbiased estimates. Plausible values were first developed for the analyses of NAEP (National Assessment of Educational Progress) data, by Mislevy and colleagues (Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992; Mislevy, Beaton, Kaplan, & Sheehan, 1992) based on Rubin's work on multiple imputations (Rubin, 1978, 1987). In this research, we compare the regression parameter estimates obtained from a concurrent estimation procedure using raw data only with a two-step procedure where the IRT measurement model and the regression model are estimated separately, that is, we aim to compare the results obtained when using raw data for the outcome variable with plausible values for the outcome variable. Furthermore, we also compare the regression results when using raw data for the covariates with point estimates of the latent covariates estimated separately. These analyses are done in a fully Bayesian framework using the WinBugs software, with Reading ability as the outcome variable and two selected latent background variables as covariates. The procedure is described in detail in the method section below.

6.2 Method

This section gives an overview of how the analyses were conducted. Firstly, we give a brief description of the cognitive reading data and the covariates that are used from the PISA 2009 dataset. Then describe the methods used in our analysis and the estimation procedure that was employed is described. Lastly the results and conclusions based on these results are presented.

6.2.1 Measures

Reading ability (outcome variable)

The PISA Reading ability was a latent scale that was measured using an international calibration of a sub-sample of students from the OECD countries. The sub sample of students

referred to as an OECD calibration sample consisted of 15 500 students comprising 500 students drawn at random from each of the 31 participating OECD countries. The calibration was done on 101 reading items distributed over 13 booklets. There were 94 dichotomous items and 7 polytomous items. The Steps model (Verhelst, Glas and de Vries, 1997) was used as the measurement model.

The background variables (covariates)
A number of background variables or covariates were used in our analyses which have been identified in PISA as important predictors of student performance. The variables that are used in the current analysis are: Joy Reading, Control Strategies in Reading, Socio-economic-status and the mean of the socio-economic-status at the school level.

Joy Reading is a latent scale that consists of 11 items. The scale consists of item like 'do you feel happy when you get a book as a present' and 'how often do you borrow books for reading from the library' etc. The scale Control Strategies is a latent scale that consists of 4 items. The scale consists of items like if the students checks if he has understood the important points etc. Both these scales explain variance in student reading performance across countries. These are the two latent scales for which raw data and latent estimates are used in the regression model for comparisons. The other covariate in the model is the student socio-economic variable called ESCS. This is not a latent variable but it is used in the model because socio-economic status is an important predictor of student performance. It is a composite index derived from a principal component analysis of three sub-indices; the possessions at home, the highest educational level of the parents and the highest occupational status of the parents. The values of ESCS in the data range from about -5 to 3.5. We also use Mean ESCS as a variable in the analysis. Mean ESCS is the average ESCS of all students in a school. The mean of the ESCS variable was obtained by summing the ESCS indices of individuals in a school and then dividing the sum over the number of students in the school. PISA studies have shown that Mean ESCS explains more variance than the individual level ESCS index in most countries and therefore we include the school level variable Mean ESCS in the regression model.

6.2.2 The regression model

All the analyses were done for equal-sized samples of 500 students each from 10 culturally diverse countries from the PISA 2009 international dataset. The students were sampled randomly using the PISA sample weights. The list of these countries for which the analyses

were conducted are presented in the tables. The regression analyses were done separately for each country. The covariates used in the analyses have been described in the section before. The model (Bryk and Raudenbausch, 2002) that was run for the analyses is given by

$$\theta_{ijc} = \alpha_{0c} + \alpha_{1c}(Joy\_reading_{ic}) + \alpha_{2c}(Control\_strategies_{ic}) + \alpha_{3c}(SES_{ic}) + \beta_{1c}(Mean\_SES_{jc}) + \mu_{jc} + \varepsilon_{ic} \quad (6.1)$$

where θ_{ijc} is a measure of the cognitive reading ability of student i in school j in country c, and where for the model for the random slopes is given by

$\alpha_{1c} = \gamma_{01} + \mu_{1j}$,
$\alpha_{2c} = \gamma_{02} + \mu_{2j}$,
$\alpha_{3c} = \gamma_{03} + \mu_{3j}$.

The analyses were carried out in a Bayesian framework using the WinBugs (Lunn, Thomas & Spiegelhalter, 2000) software. There were four different types of regression models studied. In model type 1, the parameters of the model are estimated concurrently using raw data for both the outcome variable reading ability and also the two covariates joy reading and control strategies. In model type 2, we use raw data for the outcome variable but separately estimated values for the two latent covariates. In model type 3, we use 5 plausible values for the outcome variable reading ability drawn from its posterior distribution obtained in model 1 and raw data for the two covariates. In model type 4, we use 5 plausible values for the outcome variable reading ability drawn from its posterior distribution obtained in model 1 and separately estimated values for the two latent covariates.

The four variants of the multi-level regression model described above are estimated for the 1PLM (Rasch, 1960) and the 2PLM (Lord, 1968, 1980) with and without booklet effects in the IRT measurement models (for the outcome variable only) and also using both fixed slopes and random slopes for the covariates in the model. Therefore we have a total of 4*2*2*2=32 set of results. These results are presented in tables below under the respective table headings. Tables 6.1- 6.16 present the results for the fixed slopes models and Tables 6.17- 6.32 present the results for the random slopes models.

6.3 Results

Tables 6.1 to 6.8 show the results obtained for the 4 model types described earlier (with fixed slopes) for the 1PLM as the measurement model (for scaling both the outcome variable and the

two covariates) with and without booklet effects. Tables 6.9 to 6.16 present similar results for the fixed slopes regression models with the 2PLM as the measurement model (for scaling both the outcome variable and the two covariates) with and without booklet effects. Tables 6.17 to 6.24 present the results for the random slopes regression models with the 1PLM as the measurement model (for scaling both the outcome variable and the two covariates) with and without booklet effects. Tables 6.25 to 6.32 present results for the random slopes regression with the 2PLM as the measurement model (for scaling both the outcome variable and the two covariates) with and without booklet effects.

6.3.1 1PLM results for fixed slopes models (tables 6.1 to 6.8)

Tables 6.1 to 6.8 show the results obtained for the 4 model types described earlier (with fixed slopes) for the 1PLM as the measurement model (for scaling both the outcome variable and the two covariates) with and without booklet effects. Table 6.1 presents the results for model 1 where we use raw data for both the outcome variable and the covariates and Table 6.2 for the same model but with booklet effects (in the outcome variable) added to the 1PLM.

Table 6.1. Results for the regression coefficients and the residual variance using raw data for both the outcome (reading ability) and the covariates (alpha1 & alpha2) using the 1PLM as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0,15	0,08	**0,26**	0,04	**0,17**	0,05	**0,11**	0,05	**1,01**	0,11	0,51	0,11
CHL	-0,17	0,09	**0,20**	0,04	**0,14**	0,04	0,07	0,05	**0,42**	0,07	0,46	0,09
FIN	**0,46**	0,09	**0,37**	0,04	0,09	0,05	**0,24**	0,06	0,14	0,14	0,57	0,07
DEU	**-0,21**	0,08	**0,23**	0,03	**0,18**	0,04	**0,20**	0,05	**0,96**	0,11	0,41	0,15
JPN	**0,49**	0,07	**0,26**	0,04	**0,15**	0,05	0,14	0,07	**0,96**	0,15	0,66	0,16
KOR	**0,75**	0,07	**0,23**	0,04	**0,17**	0,04	**0,11**	0,05	**0,50**	0,11	0,36	0,12
NLD	0,16	0,08	**0,21**	0,03	**0,10**	0,04	0,07	0,05	**0,92**	0,13	0,26	0,31
POL	**0,35**	0,08	**0,38**	0,04	**0,18**	0,05	**0,23**	0,06	0,18	0,11	0,75	0,05
TUR	**0,53**	0,10	**0,21**	0,04	**0,11**	0,05	**0,17**	0,04	**0,60**	0,07	0,47	0,10
USA	0,02	0,08	**0,29**	0,04	0,09	0,05	**0,15**	0,06	**0,72**	0,11	0,78	0,12

The significant effects (at 5%) are highlighted in bold.

Table 6.2. Results for the regression coefficients and the residual variance using raw data for both the outcome (reading ability) and the covariates (alpha1 & alpha2) using the 1PLM as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0,09	0,10	**0,25**	0,04	**0,16**	0,04	**0,11**	0,05	**1,01**	0,11	0,50	0,11
CHL	-0,17	0,10	**0,20**	0,04	**0,14**	0,04	0,08	0,06	**0,43**	0,07	0,43	0,11
FIN	**0,41**	0,10	**0,38**	0,04	0,09	0,05	**0,24**	0,05	0,12	0,15	0,56	0,07
DEU	**-0,26**	0,10	**0,23**	0,04	**0,18**	0,04	**0,20**	0,05	**0,95**	0,10	0,41	0,16
JPN	**0,42**	0,09	**0,27**	0,04	**0,15**	0,05	0,14	0,07	**0,99**	0,15	0,63	0,15
KOR	**0,68**	0,09	**0,23**	0,04	**0,17**	0,04	**0,12**	0,05	**0,50**	0,11	0,36	0,11
NLD	0,10	0,11	**0,22**	0,03	**0,09**	0,04	0,06	0,05	**0,94**	0,13	0,24	0,33
POL	**0,29**	0,10	**0,38**	0,04	**0,19**	0,05	**0,23**	0,05	0,18	0,11	0,75	0,05
TUR	**0,46**	0,11	**0,20**	0,04	**0,12**	0,05	**0,17**	0,04	**0,60**	0,07	0,47	0,09
USA	0,04	0,09	**0,29**	0,04	0,09	0,05	**0,15**	0,06	**0,73**	0,11	0,76	0,12

The significant effects (at 5%) are highlighted in bold.

Table 6.3 presents the results for model 2 where we use raw data for the outcome variable but independently estimated latent estimates for the two covariates and Table 6.4 results for the same model but with booklet effects (in the outcome variable) added to the 1PLM.

Table 6.3. Results for the regression coefficients and the residual variance using raw data for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 1PLM as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,19**	0,08	**0,26**	0,04	**0,16**	0,04	**0,11**	0,05	**1,00**	0,10	0,53	0,11
CHL	-0,13	0,08	**0,20**	0,04	**0,14**	0,04	0,07	0,05	**0,42**	0,07	0,48	0,09
FIN	**0,50**	0,09	**0,37**	0,04	**0,09**	0,04	**0,25**	0,06	0,14	0,14	0,61	0,06
DEU	**-0,17**	0,08	**0,23**	0,03	**0,18**	0,04	**0,20**	0,05	**0,97**	0,10	0,45	0,15
JPN	**0,53**	0,07	**0,26**	0,04	**0,15**	0,04	0,14	0,07	**0,97**	0,15	0,68	0,16
KOR	**0,79**	0,07	**0,23**	0,04	**0,17**	0,04	**0,11**	0,05	**0,50**	0,11	0,39	0,12
NLD	0,15	0,09	**0,21**	0,03	**0,09**	0,04	0,07	0,05	**0,93**	0,13	0,27	0,32
POL	**0,39**	0,08	**0,38**	0,04	**0,18**	0,05	**0,23**	0,06	0,19	0,11	0,79	0,06
TUR	**0,57**	0,11	**0,20**	0,04	0,10	0,05	**0,17**	0,04	**0,60**	0,07	0,48	0,10
USA	0,06	0,08	**0,29**	0,04	0,09	0,05	**0,15**	0,06	**0,73**	0,11	0,80	0,12

The significant effects (at 5%) are highlighted in bold.

Table 6.4. Results for the regression coefficients and the residual variance using raw data for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 1PLM as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0,04	0,09	**0,25**	0,04	**0,16**	0,05	**0,11**	0,05	**1,00**	0,11	0,52	0,11
CHL	**-0,22**	0,09	**0,20**	0,04	**0,13**	0,04	0,08	0,05	**0,43**	0,07	0,47	0,10
FIN	**0,36**	0,10	**0,37**	0,04	0,09	0,05	**0,25**	0,06	0,11	0,14	0,61	0,06
DEU	**-0,31**	0,09	**0,23**	0,04	**0,18**	0,04	**0,20**	0,05	**0,96**	0,11	0,44	0,16
JPN	**0,38**	0,08	**0,26**	0,04	**0,15**	0,04	**0,15**	0,06	**1,00**	0,15	0,65	0,15
KOR	**0,64**	0,08	**0,23**	0,04	**0,17**	0,04	**0,12**	0,05	**0,50**	0,11	0,39	0,12
NLD	0,06	0,10	**0,22**	0,03	**0,09**	0,04	0,06	0,05	**0,93**	0,13	0,26	0,31
POL	**0,24**	0,09	**0,38**	0,04	**0,18**	0,05	**0,23**	0,05	0,18	0,11	0,80	0,06
TUR	**0,42**	0,11	**0,20**	0,04	**0,11**	0,04	**0,17**	0,04	**0,60**	0,07	0,48	0,10
USA	-0,09	0,09	**0,28**	0,04	0,09	0,05	**0,15**	0,06	**0,73**	0,11	0,78	0,12

The significant effects (at 5%) are highlighted in bold.

Table 6.5 presents the results for model 3 where we use 5 plausible values for the outcome variable reading ability (drawn from the posterior distribution obtained in results from model 1) and raw data for the two covariates. Table 6.6 presents the results for the same model but with booklet effects (in the outcome variable) added to the 1PLM.

Table 6.5. Results for the regression coefficients and the residual variance using 5 plausible values for the outcome (reading ability) and raw data for the covariates (alpha1 & alpha2) using the 1PLM model as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,17**	0,07	**0,26**	0,04	**0,15**	0,05	**0,11**	0,05	**1,00**	0,10	0,55	0,11
CHL	-0,11	0,07	**0,20**	0,04	**0,13**	0,05	0,09	0,05	**0,41**	0,07	0,48	0,08
FIN	**0,42**	0,07	**0,36**	0,04	0,10	0,05	**0,25**	0,06	0,13	0,14	0,57	0,06
DEU	**-0,24**	0,07	**0,22**	0,04	**0,19**	0,05	**0,20**	0,05	**0,96**	0,11	0,42	0,17
JPN	**0,39**	0,06	**0,26**	0,04	**0,17**	0,04	0,14	0,07	**0,96**	0,15	0,66	0,14
KOR	**0,70**	0,05	**0,23**	0,04	**0,17**	0,04	**0,12**	0,05	**0,52**	0,10	0,37	0,10
NLD	**0,16**	0,08	**0,23**	0,04	**0,10**	0,04	0,08	0,05	**0,93**	0,12	0,27	0,30
POL	**0,35**	0,07	**0,38**	0,04	**0,18**	0,05	**0,22**	0,06	0,18	0,11	0,76	0,05
TUR	**0,53**	0,09	**0,21**	0,05	0,10	0,05	**0,16**	0,04	**0,62**	0,07	0,47	0,09
USA	0,01	0,07	**0,29**	0,05	**0,12**	0,05	**0,14**	0,06	**0,73**	0,11	0,77	0,11

The significant effects (at 5%) are highlighted in bold.

Table 6.6. Results for the regression coefficients and the residual variance using 5 plausible values for the outcome (reading ability) and raw data for the covariates (alpha1 & alpha2) using the 1PLM model as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0.18**	0.07	**0.26**	0.04	**0.16**	0.04	**0.11**	0.05	**1.00**	0.10	0,56	0.11
CHL	-0.12	0.07	**0.19**	0.05	**0.13**	0.05	0.08	0.05	**0.43**	0.06	0,49	0.09
FIN	**0.46**	0.06	**0.37**	0.04	0.09	0.05	**0.23**	0.06	0.12	0.11	0,59	0.06
DEU	**-0.20**	0.07	**0.23**	0.04	**0.19**	0.05	**0.19**	0.05	**0.97**	0.12	0,41	0.16
JPN	**0.49**	0.08	**0.24**	0.05	**0.16**	0.05	**0.14**	0.06	**0.97**	0.14	0,67	0.14
KOR	**0.72**	0.08	**0.23**	0.04	**0.18**	0.04	**0.11**	0.04	**0.51**	0.10	0,37	0.11
NLD	**0.14**	0.06	**0.20**	0.04	**0.09**	0.04	0.08	0.04	**0.96**	0.11	0,27	0.29
POL	**0.39**	0.06	**0.38**	0.04	**0.18**	0.05	**0.22**	0.05	0.18	0.11	0,77	0.05
TUR	**0.53**	0.08	**0.21**	0.04	**0.11**	0.05	**0.17**	0.05	**0.60**	0.07	0,46	0.09
USA	0.04	0.08	**0.29**	0.05	0.10	0.06	**0.14**	0.06	**0.72**	0.12	0,79	0.12

The significant effects (at 5%) are highlighted in bold.

Table 6.7 presents the results for model 4 where we use 5 plausible values for the outcome variable reading ability (drawn from the posterior distribution obtained in results from model 1) and independently estimated latent estimates for the two covariates. Table 6.8 presents the results for the same model but with booklet effects (in the outcome variable) added to the 1PLM.

Table 6.7. Results for the regression coefficients and the residual variance using 5 plausible values for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 1PLM model as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,19**	0,07	**0,26**	0,04	**0,15**	0,05	**0,11**	0,05	**0,99**	0,10	0,57	0,10
CHL	**-0,14**	0,06	**0,20**	0,04	**0,13**	0,04	0,09	0,05	**0,41**	0,07	0,50	0,08
FIN	**0,45**	0,07	**0,36**	0,04	0,10	0,05	**0,25**	0,06	0,13	0,14	0,60	0,06
DEU	**-0,24**	0,09	**0,22**	0,04	**0,19**	0,05	**0,20**	0,05	**0,96**	0,10	0,45	0,17
JPN	**0,48**	0,06	**0,26**	0,04	**0,17**	0,05	0,14	0,07	**0,96**	0,15	0,68	0,15
KOR	**0,75**	0,06	**0,23**	0,04	**0,17**	0,04	**0,12**	0,05	**0,51**	0,11	0,39	0,10
NLD	**0,17**	0,08	**0,23**	0,04	**0,09**	0,04	0,09	0,05	**0,95**	0,12	0,29	0,31
POL	**0,34**	0,07	**0,37**	0,04	**0,18**	0,05	**0,22**	0,06	0,18	0,11	0,80	0,06
TUR	**0,55**	0,09	**0,21**	0,05	0,10	0,05	**0,16**	0,04	**0,62**	0,07	0,48	0,09
USA	0,02	0,07	**0,28**	0,05	**0,11**	0,05	**0,14**	0,06	**0,73**	0,11	0,80	0,11

The significant effects (at 5%) are highlighted in bold.

Table 6.8. Results for the regression coefficients and the residual variance using 5 plausible values for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 1PLM model as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,18**	0,07	**0,26**	0,04	**0,14**	0,04	**0,10**	0,05	1,00	0,10	0,58	0,10
CHL	**-0,17**	0,06	**0,21**	0,04	**0,14**	0,04	0,09	0,06	0,42	0,07	0,52	0,09
FIN	**0,42**	0,06	**0,35**	0,05	0,10	0,05	**0,24**	0,06	0,13	0,14	0,62	0,06
DEU	**-0,24**	0,08	**0,22**	0,04	**0,18**	0,05	**0,20**	0,05	**0,97**	0,10	0,45	0,18
JPN	**0,47**	0,09	**0,26**	0,05	**0,17**	0,05	**0,14**	0,06	**0,96**	0,15	0,69	0,14
KOR	**0,77**	0,07	**0,23**	0,04	**0,18**	0,05	**0,14**	0,05	**0,50**	0,11	0,38	0,10
NLD	**0,18**	0,08	**0,22**	0,05	**0,09**	0,04	0,09	0,05	**0,93**	0,12	0,29	0,31
POL	**0,33**	0,07	**0,36**	0,04	**0,17**	0,04	**0,23**	0,06	0,16	0,11	0,81	0,07
TUR	**0,56**	0,07	**0,21**	0,05	**0,11**	0,05	**0,17**	0,03	**0,62**	0,07	0,46	0,09
USA	0,05	0,08	**0,28**	0,05	0,10	0,05	0,14	0,07	**0,74**	0,12	0,79	0,12

The significant effects (at 5%) are highlighted in bold.

The results obtained for the 1PLM in the tables 6.1 to 6.8 above are consistent for all the 4 regression models described earlier except for the following few cases. When using plausible values for the outcome variable instead of raw data, some border line results became significant; the size of the effect became fractionally larger to make them significant. These effects are noticed for the covariate alpha3 for Chile and the covariate alpha2 for USA. Lastly, adding booklet effects to the measurement model for the outcome variable did not have an impact on the coefficients of the regression model.

6.3.2 2PLM results for fixed slopes models (tables 6.9 to 6.16)

Tables 6.9 to 6.16 show the results obtained for the 4 models described earlier for the 2PLM as the measurement model (for scaling both the outcome variable and the two covariates) with and without booklet effects. Table 6.9 presents the results for model 1 where we use raw data for both the outcome variable and the covariates and Table 6.10 the results for the same model but with booklet effects (in the outcome variable) added to the 2PLM.

Table 6.9. Results for the regression coefficients and the residual variance using raw data for both the outcome (reading ability) and the covariates (alpha1 & alpha2) using the 2PLM as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,29**	0,08	**0,30**	0,05	**0,22**	0,06	0,10	0,06	**1,08**	0,11	0,59	0,11
CHL	-0,14	0,08	**0,19**	0,05	**0,19**	0,05	0,07	0,05	**0,43**	0,07	0,48	0,10
FIN	**0,55**	0,10	**0,42**	0,05	**0,15**	0,06	**0,27**	0,06	0,14	0,15	0,64	0,07
DEU	-0,06	0,08	**0,29**	0,04	**0,23**	0,05	**0,22**	0,05	**0,99**	0,11	0,47	0,15
JPN	**0,57**	0,09	**0,35**	0,05	**0,20**	0,06	**0,17**	0,07	**1,00**	0,16	0,75	0,17
KOR	**0,85**	0,08	**0,30**	0,05	**0,22**	0,05	0,11	0,06	**0,54**	0,11	0,42	0,12
NLD	**0,21**	0,10	**0,26**	0,04	**0,13**	0,05	0,08	0,05	**0,99**	0,13	0,28	0,35
POL	**0,50**	0,08	**0,50**	0,05	**0,19**	0,06	**0,26**	0,06	0,16	0,11	0,81	0,06
TUR	**0,55**	0,11	**0,25**	0,06	**0,12**	0,05	**0,17**	0,04	**0,60**	0,07	0,50	0,09
USA	0,08	0,08	**0,34**	0,06	0,12	0,06	**0,16**	0,07	**0,73**	0,12	0,85	0,15

The significant effects (at 5%) are highlighted in bold.

Table 6.10. Results for the regression coefficients and the residual variance using raw data for both the outcome (reading ability) and the covariates (alpha1 & alpha2) using the 2PLM as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,21**	0,08	**0,29**	0,05	**0,21**	0,06	**0,11**	0,05	**1,07**	0,11	0,57	0,12
CHL	-0,17	0,08	**0,19**	0,05	**0,19**	0,05	0,08	0,05	**0,43**	0,07	0,45	0,10
FIN	**0,48**	0,09	**0,42**	0,05	**0,15**	0,06	**0,27**	0,06	0,12	0,15	0,63	0,07
DEU	-0,13	0,08	**0,29**	0,04	**0,23**	0,06	**0,22**	0,05	**0,99**	0,11	0,47	0,16
JPN	**0,49**	0,08	**0,35**	0,05	**0,20**	0,06	**0,17**	0,07	**1,03**	0,15	0,73	0,15
KOR	**0,77**	0,07	**0,30**	0,05	**0,22**	0,05	**0,12**	0,05	**0,53**	0,11	0,41	0,12
NLD	0,14	0,08	**0,27**	0,05	**0,12**	0,05	0,07	0,05	**0,98**	0,13	0,27	0,35
POL	**0,43**	0,08	**0,50**	0,05	**0,19**	0,06	**0,26**	0,06	0,17	0,11	0,80	0,07
TUR	**0,47**	0,10	**0,24**	0,06	**0,14**	0,06	**0,17**	0,04	**0,60**	0,07	0,49	0,08
USA	0,00	0,08	**0,34**	0,06	0,11	0,06	**0,16**	0,07	**0,73**	0,11	0,84	0,13

The significant effects (at 5%) are highlighted in bold.

Table 6.11 presents the results for model 2 where we use raw data for the outcome variable but separately estimated latent estimates for the two covariates. Table 6.12 presents the results for the same model but with booklet effects (in the outcome variable) added to the 2PLM.

Table 6.11. Results for the regression coefficients and the residual variance using raw data for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,33**	0,07	**0,27**	0,05	**0,22**	0,05	0,10	0,05	**1,09**	0,11	0,61	0,12
CHL	-0,09	0,08	**0,19**	0,05	**0,19**	0,05	0,07	0,05	**0,43**	0,07	0,49	0,10
FIN	**0,60**	0,09	**0,38**	0,05	**0,16**	0,06	**0,27**	0,06	0,13	0,15	0,67	0,07
DEU	-0,01	0,07	**0,25**	0,04	**0,24**	0,05	**0,22**	0,05	**1,01**	0,11	0,49	0,18
JPN	**0,63**	0,08	**0,35**	0,05	**0,20**	0,06	**0,17**	0,07	**0,99**	0,16	0,77	0,17
KOR	**0,90**	0,07	**0,30**	0,05	**0,21**	0,05	**0,12**	0,05	**0,53**	0,11	0,44	0,12
NLD	**0,27**	0,09	**0,26**	0,04	**0,12**	0,05	0,08	0,05	**0,98**	0,14	0,29	0,35
POL	**0,56**	0,07	**0,51**	0,05	**0,13**	0,06	**0,27**	0,06	0,16	0,11	0,86	0,06
TUR	**0,61**	0,10	**0,25**	0,06	0,12	0,06	**0,17**	0,04	**0,60**	0,08	0,51	0,09
USA	0,13	0,08	**0,33**	0,06	0,12	0,06	**0,16**	0,07	**0,74**	0,12	0,88	0,12

The significant effects (at 5%) are highlighted in bold.

Table 6.12. Results for the regression coefficients and the residual variance using raw data for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,24**	0,08	**0,27**	0,05	**0,21**	0,06	**0,11**	0,05	**1,09**	0,10	0,61	0,12
CHL	-0,14	0,08	**0,19**	0,05	**0,19**	0,05	0,09	0,05	**0,43**	0,08	0,48	0,10
FIN	**0,53**	0,09	**0,39**	0,05	**0,16**	0,06	**0,27**	0,06	0,11	0,15	0,68	0,07
DEU	-0,09	0,08	**0,26**	0,04	**0,23**	0,06	**0,22**	0,05	**1,03**	0,11	0,50	0,17
JPN	**0,53**	0,08	**0,36**	0,05	**0,20**	0,05	**0,16**	0,07	**1,03**	0,16	0,76	0,16
KOR	**0,81**	0,07	**0,30**	0,05	**0,21**	0,05	0,12	0,06	**0,54**	0,11	0,45	0,12
NLD	**0,18**	0,08	**0,27**	0,04	**0,11**	0,05	0,07	0,05	**0,99**	0,13	0,29	0,35
POL	**0,48**	0,08	**0,52**	0,05	**0,13**	0,06	**0,27**	0,06	0,17	0,11	0,87	0,06
TUR	**0,52**	0,10	**0,24**	0,06	**0,14**	0,05	**0,17**	0,04	**0,61**	0,09	0,51	0,08
USA	0,04	0,08	**0,33**	0,05	0,11	0,06	**0,15**	0,06	**0,75**	0,12	0,87	0,13

The significant effects (at 5%) are highlighted in bold.

Table 6.13 presents the results for model 3 where we use 5 plausible values for the outcome variable reading ability (drawn from the posterior distribution obtained in results from model 1) and raw data for the two covariates. Table 6.14 presents the results for the same model but with booklet effects (in the outcome variable) added to the 2PLM.

Table 6.13. Results for the regression coefficients and the residual variance using 5 plausible values for the outcome (reading ability) and raw data for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,32**	0,06	**0,30**	0,05	**0,20**	0,06	0,10	0,05	**1,06**	0,10	0,63	0,11
CHL	**-0,13**	0,06	**0,19**	0,05	**0,18**	0,05	0,10	0,05	**0,41**	0,08	0,50	0,08
FIN	**0,54**	0,08	**0,41**	0,05	**0,17**	0,05	**0,27**	0,06	0,15	0,14	0,64	0,07
DEU	-0,06	0,06	**0,27**	0,04	**0,24**	0,05	**0,22**	0,05	**1,00**	0,11	0,48	0,17
JPN	**0,58**	0,07	**0,35**	0,04	**0,22**	0,06	**0,16**	0,07	**1,00**	0,14	0,76	0,15
KOR	**0,86**	0,06	**0,30**	0,04	**0,22**	0,05	**0,13**	0,05	**0,55**	0,10	0,43	0,11
NLD	**0,23**	0,08	**0,28**	0,04	**0,13**	0,05	0,09	0,05	**0,99**	0,12	0,30	0,34
POL	**0,49**	0,06	**0,49**	0,05	**0,19**	0,06	**0,25**	0,06	0,17	0,11	0,81	0,07
TUR	**0,57**	0,09	**0,26**	0,06	**0,12**	0,05	**0,16**	0,04	**0,63**	0,07	0,50	0,09
USA	0,07	0,06	**0,34**	0,06	**0,15**	0,06	**0,15**	0,07	**0,74**	0,12	0,86	0,13

The significant effects (at 5%) are highlighted in bold.

Table 6.14. Results for the regression coefficients and the residual variance using 5 plausible values for the outcome (reading ability) and raw data for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,23**	0,06	**0,30**	0,04	**0,23**	0,05	**0,13**	0,05	**1,07**	0,11	0,58	0,12
CHL	**-0,13**	0,06	**0,20**	0,05	**0,18**	0,05	0,08	0,04	**0,46**	0,06	0,46	0,10
FIN	**0,49**	0,08	**0,40**	0,04	**0,15**	0,06	**0,27**	0,05	0,17	0,14	0,64	0,06
DEU	-0,09	0,06	**0,28**	0,04	**0,25**	0,06	**0,21**	0,05	**1,06**	0,11	0,48	0,14
JPN	**0,55**	0,07	**0,35**	0,05	**0,22**	0,06	**0,16**	0,06	**1,05**	0,14	0,74	0,13
KOR	**0,82**	0,06	**0,31**	0,05	**0,22**	0,04	**0,12**	0,05	**0,54**	0,12	0,42	0,11
NLD	**0,15**	0,08	**0,26**	0,05	**0,12**	0,05	0,04	0,04	**1,01**	0,13	0,29	0,33
POL	**0,45**	0,06	**0,52**	0,05	**0,17**	0,06	**0,25**	0,06	0,16	0,11	0,83	0,06
TUR	**0,51**	0,09	**0,24**	0,05	**0,15**	0,05	**0,17**	0,04	0,60	0,07	0,51	0,08
USA	0,04	0,06	**0,34**	0,05	**0,12**	0,05	**0,16**	0,06	0,72	0,12	0,87	0,11

The significant effects (at 5%) are highlighted in bold.

Table 6.15 presents the results for model 4 where we use 5 plausible values for the outcome variable reading ability (drawn from the posterior distribution obtained in model 1) and separately estimated latent estimates for the two covariates. Table 6.16 presents the results for the same model but with booklet effects (in the outcome variable) added to the 2PLM.

Table 6.15. Results for the regression coefficients and the residual variance using 5 plausible values for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,31**	0,06	**0,26**	0,04	**0,21**	0,06	0,10	0,05	**1,08**	0,10	0,66	0,11
CHL	**-0,13**	0,06	**0,19**	0,05	**0,18**	0,05	**0,10**	0,04	**0,42**	0,08	0,52	0,09
FIN	**0,55**	0,08	**0,37**	0,05	**0,18**	0,05	**0,27**	0,05	0,13	0,14	0,68	0,07
DEU	-0,06	0,06	**0,23**	0,04	**0,25**	0,05	**0,22**	0,04	**1,02**	0,11	0,51	0,18
JPN	**0,58**	0,07	**0,35**	0,04	**0,22**	0,06	**0,16**	0,06	**0,99**	0,15	0,78	0,15
KOR	**0,86**	0,06	**0,29**	0,04	**0,21**	0,05	**0,13**	0,05	**0,54**	0,10	0,45	0,11
NLD	**0,23**	0,08	**0,28**	0,04	**0,12**	0,05	**0,10**	0,04	**1,02**	0,12	0,31	0,34
POL	**0,50**	0,06	**0,51**	0,05	**0,13**	0,05	**0,26**	0,06	0,17	0,11	0,86	0,07
TUR	**0,58**	0,09	**0,25**	0,06	**0,12**	0,05	**0,16**	0,03	**0,63**	0,07	0,52	0,09
USA	0,08	0,07	**0,33**	0,05	**0,15**	0,06	**0,15**	0,06	**0,75**	0,12	0,87	0,14

The significant effects (at 5%) are highlighted in bold.

Table 6.16. Results for the regression coefficients and the residual variance using 5 plausible values for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,23**	0,06	**0,28**	0,04	**0,23**	0,05	**0,12**	0,05	**1,08**	0,10	0,61	0,12
CHL	**-0,13**	0,06	**0,19**	0,05	**0,18**	0,05	0,08	0,04	**0,46**	0,06	0,48	0,10
FIN	**0,51**	0,08	**0,38**	0,04	**0,15**	0,06	**0,27**	0,05	0,15	0,14	0,66	0,07
DEU	-0,09	0,06	**0,25**	0,04	**0,25**	0,06	**0,21**	0,05	**1,08**	0,10	0,51	0,15
JPN	**0,56**	0,07	**0,35**	0,05	**0,21**	0,06	**0,16**	0,06	**1,04**	0,15	0,77	0,14
KOR	**0,82**	0,06	**0,31**	0,05	**0,22**	0,04	**0,13**	0,05	**0,53**	0,12	0,44	0,11
NLD	**0,15**	0,08	**0,26**	0,05	**0,12**	0,05	**0,09**	0,04	**1,02**	0,13	0,30	0,33
POL	**0,46**	0,06	**0,53**	0,05	**0,12**	0,05	**0,25**	0,06	0,16	0,11	0,87	0,06
TUR	**0,52**	0,09	**0,24**	0,05	**0,15**	0,05	**0,17**	0,04	**0,61**	0,07	0,52	0,08
USA	0,05	0,06	**0,34**	0,05	**0,12**	0,05	**0,16**	0,06	**0,72**	0,12	0,88	0,11

The significant effects (at 5%) are highlighted in bold.

The results obtained for the 2PLM in the tables above are generally consistent for the 4 regression models described earlier. However the coefficients of the covariate Joy Reading tends to be slightly lower for Belgium, Finland and Germany when using raw data for the covariate Joy Reading in the regression model instead of its latent estimate.

When using plausible values for the outcome variable instead of raw data, some border line cases became significant; the size of the effect became fractionally larger to make them

significant. These effects are noticed for the covariate alpha3 for Chile and the covariate alpha2 for USA (as was also the case for the IPLM models earlier).

Adding booklet effects to the measurement model for the outcome variable did not have an impact on the coefficients of the regression model.

Comparing the results from the fixed slopes regression when using the 1PLM and the 2PLM we see that the standard errors are larger when using the 2PLM. They are smaller for the results obtained using the 1PLM.

6.3.3 The 1PLM results for the Random Slopes models (tables 6.17 to 6.24)

Tables 6.17 to 6.24 show the results obtained for the 4 models described above (with random slopes) for the 1PLM as the measurement model (for scaling both the outcome variable and the two covariates) with and without booklet effects. Table 6.17 presents the results for model 1 where we use raw data for both the outcome variable and the covariates and Table 6.18 the same model but with booklet effects added to the 1PLM.

Table 6.17. Results for the regression coefficients and the residual variance for the random slopes model using raw data for both the outcome (reading ability) and th covariates (alpha1 & alpha2) using the 1PLM as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0.24**	0.08	**0.25**	0.04	**0.17**	0.06	0.10	0.06	**1.04**	0.12	0.42	0.11
CHL	-0.12	0.08	**0.21**	0.05	**0.14**	0.05	0.07	0.05	**0.41**	0.07	0.36	0.06
FIN	**0.56**	0.09	**0.38**	0.05	0.09	0.05	**0.24**	0.06	0.13	0.14	0.37	0.06
DEU	-0.13	0.08	**0.24**	0.04	**0.20**	0.05	**0.18**	0.05	**0.99**	0.12	0.35	0.14
JPN	**0.59**	0.07	**0.27**	0.04	**0.18**	0.05	0.13	0.07	**0.96**	0.16	0.49	0.17
KOR	**0.83**	0.07	**0.25**	0.06	**0.19**	0.05	0.12	0.06	**0.48**	0.12	0.27	0.11
NLD	**0.23**	0.09	**0.23**	0.04	**0.11**	0.05	0.08	0.06	**0.95**	0.12	0.17	0.28
POL	**0.46**	0.08	**0.40**	0.04	**0.19**	0.06	**0.25**	0.07	0.17	0.12	0.57	0.05
USA	0.10	0.07	**0.31**	0.05	0.10	0.05	**0.15**	0.07	**0.74**	0.12	0.70	0.09

The significant effects (at 5%) are highlighted in bold.

Table 6.18. Results for the regression coefficients and the residual variance for the random slopes model using raw data for both the outcome (reading ability) and the covariates (alpha1 & alpha2) using the 1PLM as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0.08	0.09	**0.24**	0.04	**0.18**	0.05	0,11	0.06	**1.04**	0.11	0.41	0.10
CHL	**-0.21**	0.10	**0.20**	0.05	**0.14**	0.05	0,09	0.06	**0.40**	0.08	0.33	0.06
FIN	**0.43**	0.10	**0.39**	0.04	0.09	0.06	**0,26**	0.06	0.09	0.15	0.37	0.06
DEU	**-0.25**	0.09	**0.23**	0.04	**0.19**	0.05	**0,19**	0.05	**0.99**	0.11	0.35	0.14
JPN	**0.45**	0.08	**0.28**	0.04	**0.17**	0.05	0,15	0.08	**0.99**	0.15	0.49	0.15
KOR	**0.69**	0.09	**0.25**	0.05	**0.19**	0.04	**0,13**	0.06	**0.48**	0.12	0.27	0.11
NLD	0.12	0.10	**0.24**	0.04	**0.10**	0.05	0,07	0.05	**0.93**	0.14	0.16	0.28
POL	**0.31**	0.09	**0.39**	0.04	**0.20**	0.06	**0,25**	0.07	0.17	0.13	0.59	0.05
USA	-0.05	0.09	**0.30**	0.05	**0.11**	0.05	**0,15**	0.07	**0.74**	0.12	0.70	0.09

The significant effects (at 5%) are highlighted in bold.

Table 6.19 presents the results for the model where we use raw data for the outcome variable but independently estimated latent estimates for the two covariates and Table 6.20 the same model but with booklet effects added to the 1PLM.

Table 6.19. Results for the regression coefficients and the residual variance for the random slopes model using raw data for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 1PLM as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0,16	0,09	**0,24**	0,04	**0,18**	0,05	0,11	0,06	**1,03**	0,11	0,47	0,10
CHL	**-0,18**	0,08	**0,20**	0,05	**0,14**	0,04	0,08	0,05	**0,40**	0,07	0,39	0,06
FIN	**0,48**	0,09	**0,39**	0,05	0,10	0,05	**0,25**	0,06	0,11	0,15	0,44	0,06
DEU	**-0,20**	0,08	**0,23**	0,04	**0,20**	0,05	**0,21**	0,06	**0,97**	0,11	0,40	0,13
JPN	**0,53**	0,08	**0,27**	0,04	**0,16**	0,05	**0,15**	0,07	**0,95**	0,16	0,56	0,16
KOR	**0,76**	0,08	**0,26**	0,05	**0,19**	0,04	0,12	0,06	**0,47**	0,12	0,31	0,11
NLD	0,18	0,09	**0,25**	0,04	**0,10**	0,04	0,08	0,05	**0,93**	0,13	0,20	0,29
POL	**0,39**	0,09	**0,40**	0,04	**0,18**	0,06	**0,24**	0,07	0,18	0,12	0,67	0,06
USA	0,04	0,08	**0,32**	0,05	0,09	0,05	0,16	0,08	**0,73**	0,12	0,76	0,10

The significant effects (at 5%) are highlighted in bold.

Table 6.20. Results for the regression coefficients and the residual variance for the random slopes model using raw data for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 1PLM as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0,08	0,09	**0,23**	0,04	**0,19**	0,05	0,12	0,06	**1,02**	0,11	0,44	0,10
CHL	**-0,21**	0,10	**0,19**	0,05	**0,14**	0,05	0,09	0,06	**0,39**	0,08	0,38	0,06
FIN	**0,43**	0,10	**0,40**	0,04	0,09	0,05	**0,24**	0,06	0,10	0,15	0,45	0,06
DEU	**-0,25**	0,09	**0,24**	0,04	**0,19**	0,04	**0,19**	0,05	**0,98**	0,11	0,40	0,14
JPN	**0,46**	0,09	**0,28**	0,04	**0,16**	0,05	0,15	0,08	**0,99**	0,16	0,53	0,15
KOR	**0,69**	0,09	**0,26**	0,05	**0,19**	0,04	**0,13**	0,06	**0,48**	0,12	0,31	0,11
NLD	0,12	0,09	**0,24**	0,04	0,09	0,05	0,07	0,05	**0,93**	0,13	0,20	0,28
POL	**0,31**	0,10	**0,39**	0,04	0,20	0,06	**0,24**	0,07	0,18	0,12	0,68	0,05
USA	-0,03	0,10	**0,31**	0,05	0,09	0,05	**0,15**	0,06	**0,74**	0,12	0,75	0,10

The significant effects (at 5%) are highlighted in bold.

Table 6.21 presents the results for the model where we use 5 plausible values drawn for the outcome variable reading ability, drawn from the posterior distribution obtained in model 1 and raw data for the two covariates Table 6.22 the same model but with booklet effects added to the 1PLM.

Table 6.21. Results for the regression coefficients and the residual variance for the random slopes model using 5 plausible values for the outcome (reading ability) and raw data for the covariates (alpha1 & alpha2) using the 1PLM model as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0,16	0,09	**0,25**	0,04	**0,21**	0,07	0,12	0,06	**0,98**	0,11	0,44	0,08
CHL	**-0,11**	0,01	**0,19**	0,04	**0,14**	0,04	**0,09**	0,04	**0,40**	0,06	0,40	0,05
FIN	**0,40**	0,16	**0,38**	0,04	0,10	0,05	**0,27**	0,05	0,08	0,13	0,39	0,05
DEU	**-0,13**	0,01	**0,22**	0,03	**0,20**	0,04	**0,18**	0,05	**0,98**	0,10	0,37	0,13
JPN	**0,43**	0,15	**0,26**	0,04	**0,19**	0,05	0,14	0,07	**0,95**	0,14	0,51	0,14
KOR	**0,62**	0,19	**0,25**	0,04	**0,19**	0,04	**0,13**	0,05	**0,49**	0,10	0,29	0,10
NLD	0,15	0,10	**0,25**	0,03	**0,12**	0,04	**0,08**	0,05	**0,93**	0,12	0,19	0,26
POL	**0,32**	0,12	**0,40**	0,04	**0,18**	0,05	**0,25**	0,06	0,16	0,11	0,58	0,05
USA	0,02	0,05	**0,31**	0,04	**0,12**	0,05	**0,13**	0,06	0,76	0,11	0,72	0,09

The significant effects (at 5%) are highlighted in bold.

Table 6.22. Results for the regression coefficients and the residual variance using 5 plausible values for the outcome (reading ability) and raw data for the covariates (alpha1 & alpha2) using the 1PLM model as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0,14	0,08	**0,23**	0,04	**0,20**	0,04	0,12	0,06	**1,01**	0,11	0,43	0,08
CHL	**-0,17**	0,07	**0,20**	0,04	**0,14**	0,05	0,08	0,05	**0,41**	0,07	0,39	0,05
FIN	**0,44**	0,12	**0,39**	0,04	0,10	0,05	**0,26**	0,05	0,11	0,14	0,38	0,06
DEU	**-0,19**	0,04	**0,23**	0,05	**0,21**	0,04	**0,20**	0,06	**0,99**	0,10	0,36	0,13
JPN	**0,44**	0,14	**0,26**	0,04	**0,17**	0,05	**0,15**	0,06	**0,95**	0,14	0,50	0,13
KOR	**0,68**	0,18	**0,26**	0,05	**0,19**	0,04	**0,12**	0,05	**0,48**	0,11	0,28	0,10
NLD	0,16	0,09	**0,24**	0,04	**0,10**	0,05	0,08	0,05	**0,93**	0,12	0,18	0,25
POL	**0,34**	0,13	**0,41**	0,05	**0,17**	0,04	**0,25**	0,06	0,18	0,12	0,57	0,05
USA	0,03	0,07	**0,30**	0,04	**0,11**	0,05	**0,14**	0,06	**0,75**	0,11	0,72	0,09

The significant effects (at 5%) are highlighted in bold.

Table 6.23 presents the results for the model where we use 5 plausible values drawn for the outcome variable reading ability from the posterior distribution obtained in model 1 and independently estimated latent estimates for the two covariates. Table 6.24 the results for the same model but with booklet effects added to the 1PLM.

Table 6.23. Results for the regression coefficients and the residual variance for the random slopes model using 5 plausible values for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 1PLM model as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0,18	0,15	**0,28**	0,04	**0,20**	0,05	0,12	0,06	**0,99**	0,12	0,45	0,08
CHL	-0,17	0,11	**0,18**	0,05	**0,13**	0,04	0,14	0,07	**0,36**	0,16	0,42	0.05
FIN	**0,48**	0,06	**0,38**	0,04	**0,11**	0,04	**0,26**	0,05	0,08	0,13	0,41	0,05
DEU	**-0,19**	0,06	**0,22**	0,03	**0,20**	0,04	**0,18**	0,05	**0,98**	0,10	0,41	0,13
JPN	**0,52**	0,05	**0,27**	0,04	**0,18**	0,04	**0,13**	0,07	**0,96**	0,14	0,56	0,14
KOR	**0,76**	0,05	**0,25**	0,04	**0,18**	0,04	**0,13**	0,05	**0,49**	0,11	0,33	0,10
NLD	0,14	0,10	**0,24**	0,04	**0,13**	0,04	**0,09**	0,05	**0,94**	0,11	0,20	0,26
POL	**0,37**	0,06	**0,40**	0,04	**0,17**	0,05	**0,24**	0,06	0,17	0,11	0,67	0,05
USA	0,01	0,06	**0,30**	0,04	**0,11**	0,05	**0,13**	0,06	**0,76**	0,11	0,77	0,09

The significant effects (at 5%) are highlighted in bold.

Table 6.24. Results for the regression coefficients and the residual variance using 5 plausible values for the outcome (reading ability) and latent estimates for covariates (alpha1 & alpha2) using the 1PLM model as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0,18	0,14	**0,27**	0,04	**0,21**	0,05	0,12	0,06	**1,02**	0,11	0,44	0,10
CHL	-0,20	0,12	**0,19**	0,05	**0,14**	0,04	0,13	0,07	**0,39**	0,08	0,40	0,06
FIN	**0,47**	0,06	**0,40**	0,05	0,10	0,05	**0,24**	0,06	0,10	0,15	0,40	0,06
DEU	**-0,23**	0,07	**0,23**	0,03	**0,19**	0,04	**0,19**	0,05	**0,98**	0,11	0,42	0,14
JPN	**0,49**	0,06	**0,27**	0,04	**0,17**	0,04	**0,15**	0,07	**0,96**	0,16	0,54	0,15
KOR	**0,67**	0,06	**0,26**	0,05	**0,19**	0,05	**0,13**	0,06	**0,48**	0,12	0,33	0,11
NLD	0,11	0,08	**0,24**	0,04	**0,12**	0,05	0,07	0,05	**0,93**	0,13	0,21	0,28
POL	**0,34**	0,06	**0,39**	0,04	**0,17**	0,06	**0,24**	0,07	0,18	0,11	0,66	0,05
USA	-0,03	0,07	**0,31**	0,05	**0,11**	0,05	**0,15**	0,06	**0,74**	0,12	0,78	0,10

The significant effects (at 5%) are highlighted in bold.

The results obtained for the 1PLM in the tables above are quite consistent for all the 4 regression models described earlier. When using plausible values for the outcome variable instead of raw data, some border line results became significant; the size of the effect became fractionally larger to make them significant. These effects are noticed for the covariate alpha3 for Chile and the covariate alpha2 for USA for the models without booklet effects. Besides these 2 particular cases, adding booklet effects to the measurement model for the outcome variable did not have an impact on the coefficients of the regression model.

6.3.4 The 2PLM results for the random slopes models (tables 6.25 to 6.32)

Tables 6.25 to 6.32 show the results obtained for the 4 models described earlier for the 2PLM as the measurement model (for scaling both the outcome variable and the two covariates) with and without booklet effects. Table 6.25 presents the results for model 1 where we use raw data for both the outcome variable and the covariates and Table 6.26 the same model but with booklet effects added to the 2PLM.

Table 6.25. Results for the regression coefficients and the residual variance for the random slopes model using raw data for both the outcome (reading ability) and the covariates (alpha1 & alpha2) using the 2PLM as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,34**	0,09	**0,34**	0,06	**0,24**	0,08	0,12	0,06	**1,15**	0,12	0,56	0,15
CHL	-0,14	0,09	**0,21**	0,06	**0,22**	0,06	0,09	0,06	**0,45**	0,08	0,44	0,07
FIN	**0,61**	0,10	**0,48**	0,06	**0,17**	0,07	**0,28**	0,07	0,15	0,16	0,55	0,07
DEU	-0,03	0,09	**0,32**	0,05	**0,27**	0,06	**0,23**	0,05	**1,07**	0,12	0,47	0,16
JPN	**0,68**	0,10	**0,39**	0,06	**0,23**	0,07	**0,19**	0,10	**1,04**	0,17	0,70	0,19
KOR	**0,94**	0,09	**0,36**	0,07	**0,24**	0,06	**0,13**	0,06	**0,54**	0,13	0,40	0,13
NLD	**0,27**	0,10	**0,31**	0,05	**0,14**	0,06	0,11	0,06	**1,02**	0,15	0,22	0,39
POL	**0,57**	0,09	**0,56**	0,06	**0,20**	0,08	**0,28**	0,07	0,17	0,12	0,74	0,06
USA	0,11	0,09	**0,41**	0,08	0,13	0,07	**0,19**	0,07	**0,78**	0,13	0,91	0,13

The significant effects (at 5%) are highlighted in bold.

Table 6.26. Results for the regression coefficients and the residual variance for the random slopes model using raw data for both the outcome (reading ability) and the covariates (alpha1 & alpha2) using the 2PLM as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,26**	0,10	**0,31**	0,07	**0,24**	0,07	**0,13**	0,07	**1,16**	0,13	0,56	0,13
CHL	-0,18	0,10	**0,21**	0,07	**0,21**	0,06	0,09	0,06	**0,46**	0,07	0,40	0,09
FIN	**0,55**	0,11	**0,49**	0,06	**0,16**	0,06	**0,27**	0,06	0,13	0,16	0,56	0,07
DEU	-0,09	0,09	**0,33**	0,06	**0,28**	0,06	**0,25**	0,07	**1,05**	0,13	0,47	0,16
JPN	**0,58**	0,10	**0,40**	0,06	**0,22**	0,07	**0,20**	0,09	**1,06**	0,17	0,69	0,17
KOR	**0,87**	0,09	**0,34**	0,05	**0,25**	0,05	**0,15**	0,07	**0,54**	0,12	0,40	0,12
NLD	0,19	0,10	**0,31**	0,05	0,14	0,07	0,09	0,06	**1,05**	0,13	0,22	0,40
POL	**0,50**	0,10	**0,57**	0,07	**0,20**	0,08	**0,28**	0,07	0,16	0,13	0,72	0,06
USA	0,04	0,09	**0,41**	0,06	0,11	0,08	**0,20**	0,08	**0,76**	0,13	0,89	0,14

The significant effects (at 5%) are highlighted in bold.

Table 6.27 presents the results for the model where we use raw data for the outcome variable but independently estimated latent estimates for the two covariates and Table 6.28 the same model but with booklet effects added to the 2PLM.

Table 6.27. Results for the regression coefficients and the residual variance for the random slopes model using raw data for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	0,40	0,10	**0,31**	0,06	**0,25**	0,07	0,12	0,06	**1,17**	0,13	0,65	0,13
CHL	-0,10	0,10	**0,22**	0,07	**0,22**	0,06	0,09	0,06	**0,45**	0,08	0,50	0,08
FIN	**0,69**	0,11	**0,47**	0,06	**0,15**	0,07	**0,29**	0,07	0,15	0,16	0,62	0,07
DEU	0,03	0,10	**0,31**	0,05	**0,27**	0,07	**0,24**	0,06	**1,09**	0,12	0,54	0,18
JPN	**0,74**	0,10	**0,40**	0,06	**0,22**	0,07	0,18	0,09	**1,05**	0,17	0,77	0,17
KOR	**1,00**	0,09	**0,35**	0,06	**0,24**	0,06	0,13	0,07	**0,55**	0,12	0,45	0,13
NLD	**0,32**	0,10	**0,31**	0,05	0,12	0,06	0,12	0,06	**1,04**	0,14	0,26	0,40
POL	**0,64**	0,09	**0,58**	0,07	0,15	0,08	**0,29**	0,08	0,16	0,13	0,88	0,06
USA	0,18	0,09	**0,40**	0,07	0,12	0,07	**0,18**	0,07	**0,79**	0,13	0,97	0,14

The significant effects (at 5%) are highlighted in bold.

Table 6.28. Results for the regression coefficients and the residual variance for the random slopes model using raw data for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,25**	0,09	**0,30**	0,06	**0,25**	0,06	0,12	0,07	**1,17**	0,12	0,62	0,12
CHL	**-0,20**	0,09	**0,21**	0,07	**0,22**	0,06	0,10	0,05	**0,45**	0,07	0,47	0,08
FIN	**0,56**	0,10	**0,48**	0,06	**0,16**	0,06	**0,29**	0,07	0,11	0,17	0,63	0,08
DEU	-0,10	0,08	**0,31**	0,05	**0,25**	0,06	**0,25**	0,07	**1,08**	0,13	0,56	0,17
JPN	**0,60**	0,08	**0,39**	0,06	**0,22**	0,06	**0,19**	0,09	**1,07**	0,17	0,75	0,17
KOR	**0,85**	0,08	**0,37**	0,07	**0,23**	0,06	0,14	0,07	**0,55**	0,13	0,44	0,13
NLD	**0,28**	0,10	**0,32**	0,06	0,12	0,06	0,10	0,06	**1,06**	0,13	0,26	0,39
POL	**0,50**	0,08	**0,58**	0,06	**0,15**	0,07	**0,30**	0,08	0,14	0,13	0,89	0,06
USA	0,03	0,09	**0,37**	0,06	0,14	0,07	**0,19**	0,07	**0,78**	0,12	0,96	0,13

The significant effects (at 5%) are highlighted in bold.

Table 6.29 presents the results for the model where we use 5 plausible values drawn for the outcome variable reading ability, drawn from the posterior distribution obtained in model 1 and raw data for the two covariates Table 6.30 the same model but with booklet effects added to the 2PLM.

Table 6.29. Results for the regression coefficients and the residual variance for the random slopes model using 5 plausible values for the outcome (reading ability) and raw data for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,38**	0,07	**0,33**	0,06	**0,22**	0,07	0,12	0,06	**1,14**	0,11	0,60	0,12
CHL	**-0,13**	0,07	**0,21**	0,07	**0,20**	0,06	0,12	0,06	**0,43**	0,08	0,48	0,07
FIN	**0,61**	0,09	**0,48**	0,06	**0,17**	0,06	**0,30**	0,07	0,14	0,15	0,57	0,08
DEU	-0,02	0,07	**0,31**	0,05	**0,27**	0,06	**0,23**	0,06	**1,06**	0,13	0,49	0,17
JPN	**0,68**	0,08	**0,39**	0,06	**0,25**	0,07	**0,19**	0,08	**1,03**	0,16	0,73	0,15
KOR	**0,95**	0,07	**0,34**	0,06	**0,25**	0,06	**0,14**	0,06	**0,57**	0,12	0,42	0,10
NLD	**0,29**	0,08	**0,32**	0,05	**0,14**	0,06	0,12	0,06	**1,04**	0,13	0,25	0,37
POL	**0,55**	0,07	**0,57**	0,06	**0,19**	0,08	**0,28**	0,07	0,15	0,12	0,74	0,07
USA	0,11	0,07	**0,40**	0,07	**0,16**	0,07	**0,18**	0,08	**0,78**	0,13	0,92	0,13

The significant effects (at 5%) are highlighted in bold.

Table 6.30. Results for the regression coefficients and the residual variance for the random slopes model using 5 plausible values for the outcome (reading ability) and raw data for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,23**	0,07	**0,30**	0,05	**0,27**	0,07	**0,11**	0,05	**1,18**	0,11	0,57	0,12
CHL	**-0,20**	0,07	**0,21**	0,06	**0,19**	0,06	**0,12**	0,05	**0,44**	0,08	0,47	0,07
FIN	**0,51**	0,09	**0,46**	0,06	**0,16**	0,07	**0,32**	0,07	0,10	0,16	0,61	0,06
DEU	-0,12	0,07	**0,31**	0,05	**0,27**	0,06	**0,23**	0,06	**1,08**	0,12	0,52	0,15
JPN	**0,59**	0,07	**0,40**	0,06	**0,23**	0,07	**0,20**	0,08	**1,08**	0,16	0,72	0,16
KOR	**0,86**	0,07	**0,36**	0,06	**0,26**	0,05	**0,14**	0,06	**0,56**	0,12	0,40	0,13
NLD	**0,18**	0,09	**0,33**	0,05	**0,13**	0,06	0,11	0,06	**1,05**	0,13	0,25	0,38
POL	**0,46**	0,07	**0,55**	0,06	**0,16**	0,07	**0,31**	0,08	0,14	0,12	0,77	0,06
USA	0,01	0,07	**0,42**	0,07	0,13	0,07	**0,18**	0,08	**0,76**	0,13	0,90	0,14

The significant effects (at 5%) are highlighted in bold.

Table 6.31 presents the results for the model where we use 5 plausible values drawn for the outcome variable reading ability from the posterior distribution obtained in model 1 and independently estimated latent estimates for the two covariates. Table 6.32 presents the results for the same model but with booklet effects added to the 2PLM.

Table 6.31. Results for the regression coefficients and the residual variance for the random slopes model using 5 plausible values for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,39**	0,07	**0,31**	0,05	**0,23**	0,07	0,12	0,06	**1,15**	0,11	0,65	0,14
CHL	-0,10	0,15	**0,19**	0,08	**0,19**	0,07	0,16	0,08	**0,36**	0,13	0,56	0,07
FIN	**0,66**	0,08	**0,46**	0,06	**0,18**	0,06	**0,29**	0,06	0,14	0,15	0,64	0,07
DEU	0,01	0,06	**0,28**	0,05	**0,28**	0,06	**0,24**	0,06	**1,08**	0,13	0,55	0,19
JPN	**0,71**	0,06	**0,39**	0,06	**0,25**	0,07	**0,17**	0,08	**1,04**	0,15	0,79	0,15
KOR	**0,98**	0,07	**0,35**	0,06	**0,24**	0,06	0,14	0,06	**0,57**	0,12	0,47	0,11
NLD	**0,31**	0,07	**0,33**	0,05	**0,13**	0,06	0,11	0,06	**1,04**	0,13	0,29	0,38
POL	**0,60**	0,07	**0,57**	0,06	**0,16**	0,07	**0,27**	0,07	0,17	0,13	0,89	0,06
USA	**0,04**	0,06	**0,38**	0,06	**0,15**	0,07	**0,18**	0,07	**0,78**	0,13	0,97	0,13

The significant effects (at 5%) are highlighted in bold.

Table 6.32. Results for the regression coefficients and the residual variance for the random slopes model using 5 plausible values for the outcome (reading ability) and latent estimates for the covariates (alpha1 & alpha2) using the 2PLM model as the measurement model with booklet effects

	α_0	s.d	α_1	s.d	α_2	s.d	α_3	s.d	β	s.d	σ_c	τ_{jc}
BEL	**0,24**	0,07	**0,28**	0,05	**0,27**	0,06	0,11	0,06	**1,19**	0,11	0,62	0,12
CHL	-0,21	0,14	**0,19**	0,08	**0,17**	0,07	0,17	0,09	**0,37**	0,13	0,54	0.07
FIN	**0,54**	0,09	**0,45**	0,06	**0,17**	0,06	**0,32**	0,07	0,10	0,16	0,67	0,07
DEU	-0,11	0,06	**0,29**	0,05	**0,27**	0,06	**0,24**	0,06	**1,10**	0,12	0,57	0,15
JPN	**0,61**	0,07	**0,39**	0,06	**0,22**	0,06	**0,18**	0,08	**1,09**	0,15	0,78	0,15
KOR	**0,88**	0,07	**0,36**	0,06	**0,25**	0,05	**0,14**	0,06	**0,56**	0,12	0,44	0,12
NLD	**0,20**	0,08	**0,33**	0,05	0,12	0,06	0,11	0,06	**1,05**	0,13	0,28	0,38
POL	**0,49**	0,07	**0,55**	0,06	0,14	0,07	**0,30**	0,08	0,15	0,12	0,89	0,06
USA	0,03	0,07	**0,40**	0,06	0,13	0,07	**0,19**	0,08	**0,77**	0,13	0,95	0,14

The significant effects (at 5%) are highlighted in bold.

The results obtained for the 2PLM in the tables above are quite consistent for all the 4 regression models described earlier. When using plausible values for the outcome variable instead of raw data, some border line results became significant; the size of the effect became fractionally larger to make them significant. These effects are noticed for the covariate alpha3 for Chile and the covariate alpha2 for USA for the models without booklet effects. Besides these 2 particular cases, adding booklet effects to the measurement model for the outcome variable did not have an impact on the coefficients of the regression model.

Adding booklet effects to the measurement model for the outcome variable did not have an impact on the coefficients of the regression model.

Comparing the results from the random slopes regression when using the 1PLM and the 2PLM as the measurement models, we see that the standard errors for the covariates are larger when using the 2PLM. They are smaller for the results obtained using the 1PLM.

6.4 Conclusions

In this research we studied the differences between a regression analysis on survey data when using a concurrent estimation of the measurement and structural models versus a two-step procedure where the IRT measurement model and the structural regression model are estimated separately. In the one step procedure we used raw data for the outcome variable and the covariates. In the two step procedure we used latent estimates of covariates and plausible values for the outcome variable. Datasets like PISA provide plausible values for the outcome variable and latent estimates of covariates for secondary practitioners of research and in this paper we have studied the accuracy of these in comparison with the concurrent estimation using raw data. We did these analyses in a fully Bayesian framework using reading ability as the outcome variable and some selected covariates. Four kinds of regression models were computed where we had a combination of latent estimates and raw data on covariates/background variables and raw data and plausible values on the outcome variable. These four regression models were run for a regression analysis with both fixed and random slopes for the covariates. For the measurement model, the analysis were done for the both the 1PLM and the 2PLM with and without booklet effects for the outcome variable reading ability.

The results obtained for the models with fixed slopes and the corresponding models with random slopes were similar and are as follows. The results obtained for the 1PLM and the 2PLM were consistent for all the 4 regression models described earlier except for the following two or three cases; when using plausible values for the outcome variable instead of raw data, some border line results became significant; the size of the effect became fractionally larger to make them significant. These effects are noticed for the covariate ‘Control Strategies’ for Chile and the covariate ‘Joy in Reading’ for USA. Otherwise the results were overall consistent.

Adding booklet effects to the measurement model for the outcome variable did not have an impact on the coefficients of the regression model for all the 4 types of models. This holds for both the 1PLM and the 2PLM measurement models with fixed or random slopes regression.

In view of these findings, results obtained using plausible values and latent covariates (as provided in large scale data sets like PISA) were comparable with the results obtained using a concurrent estimation with raw data for the various kinds of IRT measurement and structural regression models. Thus plausible values for outcomes variables and maximum likelihood estimates of latent covariates (as provided in the PISA datatset) can be directly used by secondary researchers to obtain accurate estimates of regression coeffecients for various IRT models and multilevel regression models with fixed and random slopes.

Chapter 7

Exploring the relation between socio-economic status and reading achievement in PISA 2009 through an Intercepts-and-Slopes-as-Outcomes paradigm

In some countries more than others, factors like parental socio-economic status (SES) can cause inequalities in educational achievement. Here we show how the mechanisms leading to such inequalities can be scrutinized by involving background variables which impact the relation between SES and achievement. We use the Intercept-and-Slopes-as-Outcomes paradigm which recognizes that the outcomes of schooling systems are not only characterized by average achievement (the intercept) but also by the achievement-SES regression slope. We show how background variables moderate the relationship between SES and achievement. As an illustration, we examine the relationship between reading achievement and SES, and how this is moderated by school funding and curriculum. This is done for several countries that participated in the PISA 2009 cycle.

7.1 Introduction

Advancing equality of opportunity is widely seen as a significant function of public goods such as higher education. Fair access to education and training, and flexibility in the way it can be accessed can help to unlock the opportunities that will allow the widest pool of talent to enter and progress within society. Equity in educational achievement is a crucial determinant of the extent of equality of opportunity and intergenerational mobility achievement by societies (Schutz, Ursprung & Woßmann, 2008). Since Coleman's (1966) landmark study on Equality of Educational Opportunity, socioeconomic status has been seen as a strong predictor of student achievement. Coleman asserted that the influence of student background was greater than anything that goes on within schools. Subsequent research in the sociology of education offers

conclusive evidence of a positive relationship between family socio-economic status (SES) and the academic achievement of students (Sirin, 2005; White, 1982). Though the factor of race or ethnicity is closely associated with that of poverty as a predictor of achievement, Harkreader and Weathersby (1998) found its influence much less than economic factors. The relationship between family SES and academic achievement is referred to in the literature as a socioeconomic gradient because it is gradual and increases across the range of SES (Adler et al., 1994; Willms, 2002), or as a socio-economic gap because it implies a gap in academic achievement between students of high and low SES families. Scholars have shown that a socio-economic gap in the early school years has lasting consequences. Particularly, as low SES children get older their situation tends to worsen. In the longer term, they are less likely to enter the labour market successfully or pursue post-secondary education (Alexander, Entwisle, & Olson, 2007; Cabrera & La Nasa, 2001; Kerckhoff, Raudenbush, & Glennie, 2001; Raudenbush & Kasim, 1998). That educational and labour opportunities are unequally distributed among individuals of varying SES poses concerns and challenges in societies that value equal opportunity irrespective of socio-economic background. Therefore numerous studies have been undertaken to explain and understand the processes that configure socio-economic gradients.

The issue of socio-economic status and its relationship to student achievement is more complex than Coleman's (1966) report first intimated. Chall (1996) analyzed a combination of National Assessment of Educational Progress (NAEP) reading results, Scholastic Aptitude Tests scores over time, and a synthesis of research on reading beginning from 1910 to 1996. She also concluded that there are large differences between higher and lower socio-economic status children but that the differences were smaller among younger children and increased in the higher grades. Other researchers have examined the underlying family processes that mediate the relationship between SES and educational outcomes (Chao & Willms, 2002; Guo & Harris, 2000; Hanson, McLanahan, & Thomson, 1997; Lareau, 2002; Yeung, Linver, & Brooks-Gunn,2002); the extent to which socio-economic gaps in academic achievement are consistent across subject areas (Ma, 2000); the school practices that can effectively reduce achievement inequalities across SES groups (Cohen, 1982; Rutter & Maughan, 2002; Scheerens, 1992); the impact of the school organizational model (Coleman & Hoffer, 1987; Coleman et al., 1981, 1982); the role of tracking (Ansalone, 2010, Loveless, 1999; Oaks, 1985; Schofield, 2010; Van de Werfhorst & Mijs, 2010). In this study we focus on two important facets of educational organization, school type and tracking, that are known to moderate the relationship between

SES and achievement and are malleable characteristics of the educational system from a policy perspective at the system level.

We first outline the theory and past research on the role of the school sector in academic achievement. The influence of school type on students' results has been a major topic in the educational field for quite a long time. Much research has been conducted on the relative effects of school type on academic achievement. Such studies were influenced by the relative merits of public and private sectors in education, and had implications for the allocation of government resources, that is, if the private sector can educate children more effectively (and for less money), then it is difficult to justify the exclusive hold that public schools have on public funding for education (Lubienski, 2001, 2003). Public schools are viewed as input-oriented organizations, accountable to bureaucracies, not to consumers, so they lack structural incentives to innovate, improve, or respond to demands for quality from the groups that they serve (Chubb & Moe, 1990; Coleman, 1997). Private schools on the other hand are considered to be free of much of the bureaucracy and regulation that inhibit performance in the public sector, but they are not shielded from competition as public schools are. They must demonstrate greater effectiveness in terms of their outputs in order to attract families willing to pay tuition. So, while private schools tend to draw more advantaged families that can afford the added costs, if such schools can be shown to achieve superior results with the same types of students who attend public schools, then there is a stronger argument for policies that promote the private sector. This would not only be a more efficient and effective use of public resources in educating the public, but there is a serious equity concern about trapping poorer families in the underperforming public sector. However not all researchers agree about academic advantages that private schools ought to convey based on the above reasoning. In their opinion public schools are better resourced than independent private schools on average and public school teachers and administrators are required to receive a certain level of training. If such factors are linked to higher student achievement, then we might expect that public schools could perform relatively well. Over the last decades researchers have collected actual evidence to put such theoretical arguments about school type to the test. Dan Goldhaber (1996) examined a subsample from the NELS data of over 3,000 students each in mathematics and reading. After controlling for the fact that the private school students come from more affluent and educated families, he found no achievement advantage in private schools. On the other hand a report by James Coleman and his colleagues found a notable private school effect, inherent advantages for schools in the private sector that resulted in greater academic achievement even after

controlling for differences in student populations (Coleman & Hoffer, 1987; Coleman et al., 1981, 1982). In the UK studies about public/private schools do show that the school type does make a difference on students' achievement (Archer, 1984). The admissions system strongly favors privately educated applicants, particularly in accessing top universities. According to the Independent Schools Council, the independent sector educates around 6.5% of the total number of school children in the UK, however it accounts for over 48% of the entrants to the UK's 30 most selective universities. Furthermore in the UK the most socio-economically advantaged 20% of the young people are up to six times more likely to enter higher education than the most socio-economically disadvantaged 20%. These differences are even more skewed in the most selective disciplines (Brennan & Osborne, 2008, p.180). In the USA a government report on mathematics results from the 2003 National Assessment of Educational Progress (NAEP) highlighted this finding: "Public-school students scored lower on average than non-public-school students at both grades 4 and 8". This finding was not considered new as private school students in the United States have typically scored higher than public school students on standardized tests, confirming the perception among the US public and policymakers that private schools are inherently more effective than public schools.

However, the real question for researchers and policymakers is whether differences in test scores between the private schools and public schools is primarily due to differences in the student populations served by these different sectors. According to Lubienski & Lubienski (2005), there are reasons to expect that private schools would have higher achievement than public schools, even after accounting for differences in their student population. Private schools are free of much of the bureaucracy that plagues schools in the public sector, and are able to focus on a core academic curriculum. Furthermore, in private schools, parents are positioned to select a school based on academic quality, and to choose another if their school fails to meet expectations. Additionally, private school parents, through the act of choosing, demonstrate a commitment to their children's education, a characteristic that goes beyond typical SES measures and is associated with higher student achievement. In view of such factors, it seems reasonable to expect that private schools have many incentives to excel.

Next we outline the theory and past research on the role of tracking students in different curriculums in academic achievement. Tracking or student grouping has also been a much investigated issue in educational circles. Advocates of tracking believe that tracking benefits students' academic performance and detractors on the other hand believe that tracking would increase the achievement gap between tracked students. A number of studies can be found on

the impact of tracking on student performance (Ansalone, 2010, Loveless, 1999; Oaks, 1985; Schofield, 2010; Van de Werfhorst & Mijs, 2010). The key issues of tracking is whether this widely adopted educational practice has an impact on students achievement, which in turn may lead to widened inequality among students tracked into different ability groups or curriculum types. The arguments about tracking often rest on a perceived trade-off between equity and efficiency. Some discussions of tracking are mainly concerned with placements between different types of schools and others with placements into different tracks within schools but the arguments are basically the same. The main argument behind tracking is that homogeneous classrooms permit a focused curriculum and appropriately paced instruction that leads to the maximum learning by all students. The arguments against tracking largely revolve around the concerns that the lower groups will be systematically disadvantaged by slower learning environments that leave them far behind the skills of those in the upper groups. Furthermore, a number of studies on school influences argue that because school practices like tracking are not neutral in their treatment of students of varying socio-economic backgrounds, they tend to produce a widening gap between students in higher and lower tracks. For example, researchers have suggested that the disproportionate assignment of low SES students to lower school tracks (Kerckhoff, 1993; Pallas, Entwisle, Alexander, & Stiuka, 1994) lead to increasing inequalities between high SES and low SES students over time. Close investigation of trends in high school student results in the USA indicates that curriculum differentiation has had a negative effect on the education of many young adults, particularly working-class and African-American students (Mirel & Angus, 1994). Research on tracking in the US and UK has consistently shown that tracking was positive to the students who were assigned to high ability tracks and was negative to the students assigned to low ability tracks (Gamoran, 1992). Reviewers of empirical studies of tracking or ability grouping have pointed out that one possible mechanism for tracking to be positive to students' performance is for ability grouping to be accompanied with curriculum differentiation (Kulik & Kulik, 1992; Schofield, 2010; Slavin, 1990). Tracking without providing students with different levels of difficulty or providing them with different programs or instruction would not produce significant impact on student's performance. The data used in this study is about curriculum based tracking and thus enables the possibility to test this theory.

Hence the literature reviewed suggests that socio-economic background is a key factor in inequality in educational achievement, not only to start with, but results in further barriers over time at later stages of education. Tracking is one manner in which this occurs as there tends to be a disproportionate assignment of low SES students to lower school tracks in many countries

leading to increasing inequalities between high SES and low SES students over time. With respect to the impact of school type on academic inequalities, the private schools perform better than public schools in most countries where performance results have been published. Some researchers are of the view that the difference in the background characteristics of students can explain much of the performance difference between the two school sectors while there are others who feel that even after taking into account the background characteristics of the enrolled student populations, there are reasons to expect that private schools would have higher achievement than public schools.

The present study examines the academic achievement gap between high SES and low SES students across different countries. Focusing on reading academic performance, we seek to establish how the achievement gap associated with SES is moderated by the type of school and curriculum orientation. We investigate the relationship between these background variables and educational achievement using the Intercepts-and-Slopes-as-Outcomes model of Bryk and Raudenbausch (2002). This approach recognizes that the outcomes of schooling systems are not only characterized by average achievement (the intercept) but also by the achievement-SES regression slope. Thus we can gain insights regarding equity, not only on intercepts (average achievement of schools) but also on slopes (the differences within schools).

In our analyses, we use a step-wise approach for modeling the data to get insights into occurring patterns. We start with the ANOVA model and gradually build up the analyses to the more detailed Intercepts-and-Slopes-as-Outcomes model. This allows us to obtain a broad picture of what can be distilled from these analyses regarding equity. We use school type and study orientation as moderating variables on the relation between SES and achievement but the method can be applied to other variables also. We use the PISA 2009 dataset for our analyses. The PISA (Program for International Student Assessment) educational survey of the OECD aims to evaluate students' skills and aptitude for lifelong learning in the areas of reading literacy, mathematics, science and problem solving. The target population of PISA consists of 15 year-old students just before the school leaving age. The first PISA cycle started in the year 2000 and is repeated every three years with one of the three mentioned domains being the focus in each cycle. Besides collecting responses on cognitive tests, PISA also collects data on background characteristics of the students and information about the schools through the so-called context questionnaires. PISA 2009 has explored the impact of SES and Mean SES on student achievement using multi-level models in the International Report but by only using

Intercepts-as-Outcomes and without incorporating the moderating variables we selected for our study. This study aims to shed new light on equity related outcomes using the PISA dataset.

School Type

In many countries that participate in the PISA project, the public, private and semi-private schools exist alongside each other. Because of the struggle between the state and churches on the ownership and financing of schooling, these three types of schools have emerged in most of the European countries. Because the tuition fee of private schools is often higher, private schools may be pressed to deliver better results to justify the higher fee. The schools can be categorized based on two main aspects. The first is that who has the right to decide the school's organization and the curriculum. The second is that how the schools are funded. In relation to the first aspect, the schools are categorized into two types: the public school and the private schools. If the public agencies takes the decisions on the school organization and take the responsibility to raise the funds, then it is called public education (Coleman & Hoffer, 1987). However, if the churches, other religious institutions or commercial organizations established the school, they are called private schools. Within the private sector, the private schools can be sorted as either government-dependent or government-independent schools based on where they get funding from. The private-independent schools raise their funding mainly by means of pupil fees, donations, sponsoring, and parental fund-raising (Corten & Dronkers, 2006).

Study orientation

PISA also provides information on the orientation of the study program that students are enrolled in. In past decades, the European Union has made distinctions between educational path ways of higher education. In general, they distinguished between the vocational education and the academic or general education. The curriculum of the two kinds of education is not the same because they prepare pupils for different types of occupations. For the vocational education, most of the students enter the labor market directly after their graduation.

Research questions

We investigate the relations between the socioeconomic status and reading performance in private and public schools. We also study the effect of the orientation of the study program a student is enrolled in. Using the PISA 2009 data, we addressed the following research questions:

1. Is the relation between reading achievement and socioeconomic status different in public and private schools?
2. Is the relation between reading achievement and socioeconomic status different in general and vocational schools?

7.2 Method

7.2.1 Sample

In this study, the sample comprised of students in 8 countries/regions who participated in PISA 2009. The 8 countries/regions selected are: Austria, Belgium, Finland, Indonesia, Ireland, Japan, the Netherlands and Chinese Taipei. This set of countries was chosen because it covers a wide variety of educational systems across the globe. The sampling process for selecting students within these countries can be found in the technical report of PISA 2009 (see PISA, OECD website). For our analyses the entire sampled dataset for each country was used.

7.2.2 Measures

Socio-economic Status (independent variable)

The socio-economic variable in PISA is called the ESCS. It is a composite index derived from a principal component analysis of three sub-indices; the possessions at home (HOMEPOS), the highest educational level of the parents (PARED) and the highest occupational status of the parents (HISEI). We also use Mean ESCS as a variable in the analyses. Mean ESCS is the average ESCS of all students in a school. The values of ESCS in the data range from about -5 to 3.5. Please note however that in the remainder of the text the term "ESCS" will be replaced by the term "SES", even though it refers to the ESCS variable in the PISA dataset.

School Type (independent variable)

The PISA school questionnaire contains a question about the school type. In the PISA database this variable is denoted as SCHLTYPE. There are three types of schools in the data: public schools=1, private government dependent=2 and private-independent=3. For our analyses, we combine the two different kinds of private schools under one heading of private schools. One of the reasons for doing this is that in many countries there is only one kind of private school either private independent or private government funded and combining them for our analyses

allows us to make a comparison between countries besides making a clear distinction between private and public types. Thus in the regression of achievement on School type, public schools is coded as 1 and private schools is coded as 0.

Study Orientation (independent variable)

The PISA student questionnaire contains a question about the study program orientation. In the PISA database this variable is called ISCEDO and indicates whether the programs curricular content is general, pre vocational or vocational. For our analyses, we combine pre-vocational and vocational orientation under one heading of vocational orientation. One of the reasons for doing this is that in many countries there is only one kind of orientation, pre vocational or vocational and combining them for our analyses allows us to make a comparison between countries besides making a clear distinction between vocational and general types. Thus in the regression of achievement on program orientation, Vocational is coded as 1, General is coded as 0.

Reading Literacy (dependent variable)

Reading Literacy was a latent scale that was measured by 131 items distributed over 13 booklets. An Item Response Theory (IRT) model (Lord & Novick, 1968) was used to estimate the students' proficiency scores on a common scale. Plausible values were used to account for the measurement error. Plausible values were first developed for the analyses of NAEP (National Assessment of Educational Progress) data, based on Rubin's work on multiple imputation (1978).

Plausible Values

Plausible values are student proficiency estimates. Mathematically, we can describe the process as follows: Given an item response pattern $\boldsymbol{x}$, and ability θ, let $f(\boldsymbol{x}|\theta)$ be the probability of the response pattern under the IRT model. We assume that θ has a normal distribution $g(\theta|Y) \sim N(YB, \sigma^2)$ where θ is regressed on all covariates measured through the regression model $E(\theta)=YB$. (In our terminology, we often call $f(\boldsymbol{x}|\theta)$ the item response model, and $g(\theta|Y)$ the population model). It can be shown that the posterior distribution, $h(\theta|\boldsymbol{x},\boldsymbol{Y})$, is given by

$$h(\theta \mid x, Y) \propto f(x \mid \theta) g(\theta \mid Y) \ . \qquad (7.1)$$

That is, if a student's item response pattern is **x**, then the student's posterior θ-distribution is given by $h(\theta|\boldsymbol{x},\boldsymbol{Y})$. Plausible values for a student with item response pattern **x** are random draws from the probability distribution with density $h(\theta|\boldsymbol{x},\boldsymbol{Y})$. Therefore, plausible values provide not only information about a student's "proficiency estimate", but also the uncertainty associated with this estimate. All the regression analyses were done 5 times with different plausible values for reading literacy for each regression. The results for the effect size of the regression coefficients presented in this study are the average of the results obtained from the five runs. The standard errors and the significance tests were also adjusted for the variation between the five sets of results.

7.2.3 Data Analysis

Our regression models contain two levels: student level and school level. All regression analyses were done separately for each country. The regression model used for the analyses was a 2-level random intercepts model (Bryk & Raudenbush, 2002). In our analyses, we used four steps. First, an ANOVA model without predictors is built to evaluate within and between groups variance components. Second, the effects of school level predictor mean SES is included in the model. In the third model, the effects of student level predictor SES (group centered) is added. Fourth, both the effects of level 1 and level 2 predictors are added, that is, mean SES, School type, Study orientation and group-centered SES.

The one-way ANOVA model

To assess the effects of the predictor variables, the ANOVA model is used as a baseline. In the ANOVA model, only the outcome variable, reading achievement, is included. In this model, we define Y_{ij} as the reading performance for student *i* in school *j*, γ_{00} is the grand mean of reading performance across the population of schools, r_{ij} is the student level error and u_{0j} is the school level effect. So, the level 1 model is:

$$Y_{ij} = \beta_{0j} + r_{ij}, \tag{7.2}$$

where β_{0j} is school *j*'s mean reading achievement. Further we assume $r_{ij} \sim (0, \sigma^2)$ and refer to σ^2 as level 1 variance.

At level 2 (school level), each school's mean reading achievement is represented as a function of the grand mean plus a random error, that is,

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{7.3}$$

where we assume $u_{0j} \sim (0,\ \tau^2_{00})$, and refer to τ^2_{00} as the level 2 variance.

With this model, the within school variance σ^2 and the between school variance τ_{00} of the reading achievement can be estimated.

Means-as-outcomes regression

The school average reading performance can be predicted by group characteristics. This model is motivated by the question whether school differences in achievement can be explained by school mean SES differences. Individual reading achievement scores are viewed as varying around their school means. The level 1 model remains the same as

$$Y_{ij} = \beta_{0j} + r_{ij} \ .$$

The level 2 model now includes the mean SES of the students in a school.

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(mean\_SES)_j + u_{0j} \ , \tag{7.4}$$

where γ_{00} is the intercept, γ_{01} is the effect of mean SES on the school average reading achievement, β_{0j}, u_{0j} the deviation of school *j*'s mean from the grand mean and we assume $u_{0j} \sim (0,\ \tau^2_{00})$.

With this model, we can estimate the proportion of variance of school performance explained by mean SES. Furthermore, we can test if the reading performance still varies significantly across schools if the mean SES is controlled for.

The random-coefficient model

Next we focus on the within school effects of SES rather than between school effects as described above. The within school effects are represented by the within school regression of achievement on individual SES, also called the within school slope of SES. Regarding the regression of reading achievement on SES for every school, it was investigated whether there is a distinction in the different schools in the slopes and the intercepts. If so, how much do the

regression effects vary from school to school? And what is the correlation between the intercepts and the slopes? So we use the random-coefficient regression model to investigate this.

In the Random coefficient model, the group (classroom) centered predictor SES is added to level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij} \quad . \tag{7.5}$$

As in the model above, we assume $r_{ij} \sim (0, \sigma^2)$, however, σ^2 is the residual variance at the student level after controlling for the effect of student SES.

The level 2 model is

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{7.6}$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \tag{7.7}$$

with variance-covariance matrix for the random effects:

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} \right)$$

where

γ_{00} is the average intercept across the population of schools,
u_{0j} is the unique increment to the intercept associated with school j,
γ_{10} is the average SES-reading achievement regression slope across the schools,
u_{1j} is the unique increment to the slope associated with school j,
τ_{00} is the residual variance in the intercept,
τ_{11} is the residual variance in the slopes,
$\tau_{01} = \tau_{10}$ is the covariance between intercepts and slopes.

Using this model, we can test the hypothesis whether there are differences in SES slopes among the schools. We can also analyze the public schools and the private schools separately with this model. Suppose we want to explore the different relationship between SES and reading achievement in private and public schools. This is exemplified in Figure 7.1. The hypothetical

figure shows that the overall slope of SES and reading achievement is steeper in private schools than in public schools. It can also be seen that in the private schools with increasing average school reading performance, the relationship between SES and reading achievement becomes stronger and that it is the other way around in public schools. This indicates that there is a positive correlation between intercepts and slopes in private schools and a negative correlation between intercepts and slopes in public schools.

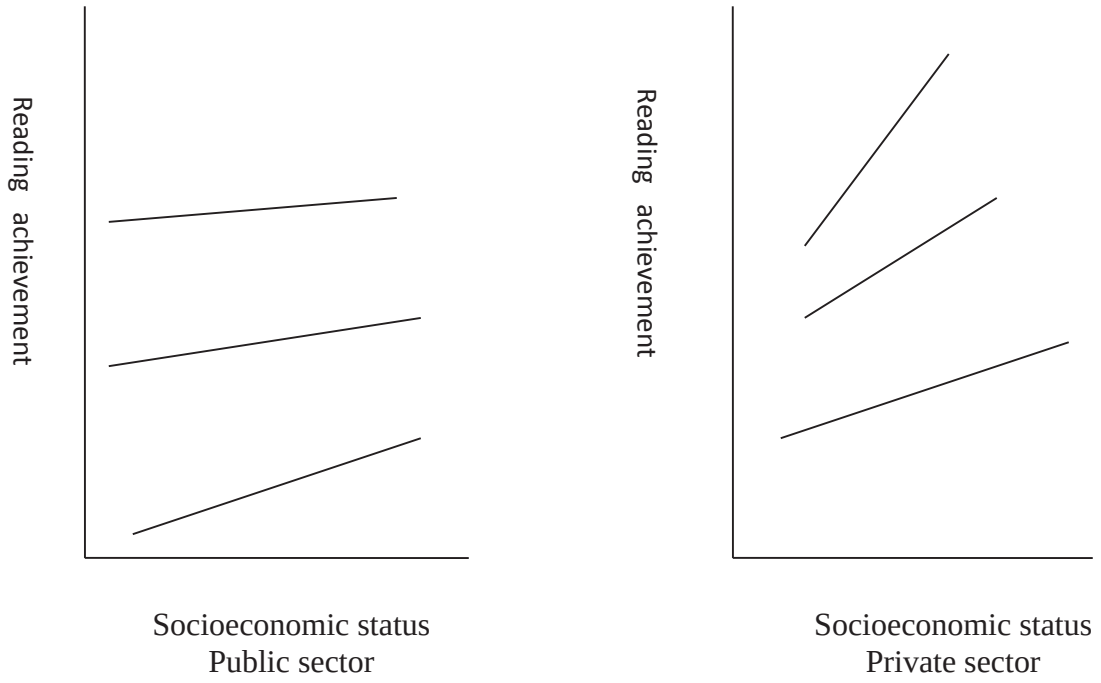

Figure 7.1. Hypothetical regression of reading achievement as a function of SES within private and public sectors.

The intercept- and slopes-as-outcomes model

In order to explore whether School type and Program orientation have different reading achievement as well as whether the strength of association between SES and reading achievement are the same, the intercept- and slopes-as-outcomes model is also used. The school level model is elaborated with new predictors School type, Study orientation and Mean SES. The school level model is also elaborated with new predictors School type and Study orientation. This leads to the model

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(mean\_SES)_j + \gamma_{02}(SCHTYPE)_j + \gamma_{03}(Orientation) + u_{0j} \qquad (7.8)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(mean\_SES)_j + \gamma_{12}(SCHTYPE)_j + \gamma_{13}(Orientation) + u_{1j} \quad (7.9)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}\right).$$

7.3 Results

Tables 7.1 and 7.2 present descriptive statistics about the schooling systems, the enrollment of students in them and the distribution of SES across the school types and study orientation types.

Table 7.1. Percentages of students by school sector and orientation of the study across countries in the dataset

	Public Schools		Private Schools		General Programs		Vocational Programs	
Country	Number of schools	%	Number of schools	%	Number of students	%	Number of students	%
Austria	234	85.7	39	14.3	1917	29.1	4666	70.9
Belgium	89	34.6	168	65.4	4556	53.6	3945	46.4
Finland	191	94.1	12	5.9	5804	99.9	6	0.1
Indonesia	85	47.2	95	52.8	4387	85.4	749	14.6
Ireland	57	44.5	71	45.5	3871	98.4	63	1.6
Japan	135	73.4	49	26.6	4580	75.2	1508	24.8
Netherlands	69	39.9	104	60.1	3527	74.1	1232	25.9
Taipei	97	64.7	53	35.3	3539	60.7	2292	39.3

Table 7.2. The SES means by school sectors and study orientation across countries and the standard deviation of SES within school sectors and study orientation programs

Country	School Sector	Number of schools	Mean	S.D.	Study Orientation	Number of students	Mean	S.D.
Austria	Public	234	0.03	0.81	General	1917	0.43	0.94
	Private	39	0.45	0.79	Vocational	4666	0.05	0.72
Belgium	Public	89	-0.01	0.89	General	4556	0.52	0.88
	Private	168	0.31	0.92	Vocational	3945	-0.15	0.81
Finland	Public	191	0.40	0.77	N.A.	N.A.	N.A.	N.A.
	Private	12	0.61	0.78	N.A.	N.A.	N.A.	N.A.
Indonesia	Public	85	-1.46	1.10	General	4387	-1.50	1.10
	Private	95	-1.61	1.08	Vocational	749	-1.68	0.92
Ireland	Public	57	-0.12	0.81	N.A.	N.A.	N.A.	N.A.
	Private	71	0.16	0.85	N.A.	N.A.	N.A.	N.A.
Japan	Public	135	-0.06	0.73	General	4580	-0.09	0.70
	Private	49	0.15	0.69	Vocational	1508	0.22	0.67
Netherlands	Public	69	0.35	0.83	General	3527	0.45	0.81
	Private	104	0.28	0.85	Vocational	1232	-0.11	0.79
Taipei	Public	97	-0.27	0.83	General	3539	-0.17	0.83
	Private	53	-0.35	0.81	Vocational	2292	-0.50	0.76

Table 7.3 shows the One-Way ANOVA results. The proportion of variance explained by the school difference is given by the so called intra-class correlation index *p*, which is displayed in the last column. For Austria, Belgium, Japan and the Netherlands, there is more variation in the reading achievement is at school level, while in Finland, Indonesia, Ireland and Taipei there is more variance within schools.

Table 7.3. Between and within school variance with the ANOVA model

Country	γ_{00}	σ^2	τ_{00}	p
Austria	453.8	4421.2	5813.4	0.56
Belgium	496.7	4781.4	6617.5	0.58
Finland	530.2	6915.2	646.9	0.08
Indonesia	398.9	2253.2	2088.7	0.48
Ireland	494.1	6841.0	2260.5	0.24
Japan	517.3	5033.8	5122.3	0.51
The Netherlands	514.9	2842.2	4740.4	0.62
Taipei	495.8	4795.4	2671.4	0.35

γ_{00} is the grand mean of reading achievement
σ^2 is the within school variance
τ_{00} is the between school variance
ρ is the intra-class correlation

The school differences explain most of the variance in the Netherlands, where 62% of the variance in reading achievement is between schools. In contrast, in Finland only 8% of the variance can be explained by the school differences.

Table 7.4. Results from means-as-outcomes regression

Country	γ_{00}	γ_{01}	S.E.	σ^2	τ_{00}	V	ρ
Austria	**456**	**104.6**	2.2	**4428**	**2503**	0.57	0.36
Belgium	**480**	**125.0**	1.7	**4783**	**1765**	0.73	0.26
Finland	**518**	**29.6**	3.6	**6915**	**562**	0.13	0.07
Indonesia	**456**	**36.8**	1.2	**2253**	**1393**	0.33	0.38
Ireland	**492**	**73.4**	3.2	**6870**	**1136**	0.50	0.14
Japan	**523**	**151.0**	2.9	**5034**	**2017**	0.61	0.28
The Netherlands	**481**	**113.0**	2.3	**2843**	**2151**	0.55	0.43
Taipei	**528**	**100.2**	2.5	**4831**	**1019**	0.62	0.17

The significant effects (at 5%) are highlighted in bold.

γ_{01} is the effect of mean SES
σ^2 is the within school variance
τ_{00} is the between school variance
V is the proportion of between school variance explained by the school Mean SES
ρ is the conditional intra-class correlation

Table 7.4 presents the results of the means-as-outcomes regression. In the means-as-outcomes model, the student reading achievement scores are viewed as varying around their school means. By comparing the τ_{00} estimates across the ANOVA and the means-as-outcomes model, we can develop an index of the proportion of reduction in the between school variance τ_{00} that is explained by the Mean SES. We refer to it as V in Table 7.4. The estimated ρ is now a conditional intra-class correlation that measures the degree of the dependence among observations within schools that are of the same Mean SES.

For Belgium, Mean SES plays the most important role as it explains 73% of the between school variance in reading achievement. For Finland and Indonesia, only 13% and 33% of the true between school variance in reading is accounted for by Mean SES. In all of the countries the intra-class correlation is reduced after accounting for the effect of Mean SES of the schools.

Table 7.5. The fixed and random effects in the Random Coefficient model

Country	γ_{00}	*S.E.*	γ_{10}	S.E.	τ_{00}	*S.E.*	τ_{11}	S.E.	τ_{01}	σ^2
Austria	**454.2**	1.4	**10.9**	1.4	**5801.1**	526.2	**109.8**	40.1	**-0.14**	**4290.4**
Belgium	**497.0**	1.2	**13.1**	1.1	**6447.2**	571.1	**109.2**	32.3	**0.18**	**4564.8**
Finland	**530.8**	1.7	**29.9**	1.5	**654.4**	90.4	76.1	46.2	N.A.	**6354.4**
Indonesia	**398.0**	0.8	**2.9**	0.8	**2082.6**	223.4	7.2	25.1	N.A.	**2241.8**
Ireland	**495.8**	1.8	**28.1**	1.8	**2127.2**	294.0	28.3	49.1	N.A.	**6341.1**
Japan	**518.2**	1.6	**4.5**	1.5	**5060.6**	541.2	60.1	45.2	N.A.	**4917.3**
The Netherlands	**515.2**	1.1	**6.7**	1.1	**4717.6**	497.5	8.0	21.2	N.A.	**2772.2**
Taipei	**496.0**	1.4	**15.5**	1.5	**2668.6**	321.3	**108.1**	41.3	**-0.15**	**4571.2**

The significant effects (at 5%) are highlighted in bold. For countries with N.A., there was insufficient variation in the slopes to determine a correlation.

γ_{00} is the average of school means on reading achievement across the population of schools
γ_{10} is the average SES-reading achievement regression slope across the school
τ_{00} is the unconditional variance in the student level intercepts
τ_{11} is the unconditional variance in the student level slopes
τ_{01} is the correlation (standardized covariance) between intercepts and slopes

Table 7.5 presents the Random coefficient model (see equation 5). It can be seen from Table 7.5 that in all the countries, the average of the SES-achievement slopes is positive. That indicates that, on average, the students SES is positively related to reading achievement within

the schools. From the results, we can also see the size of the within school effects of SES on achievement and the level 1 variance in reading achievement explained by SES. The slopes γ_{10} indicate the effect size and the variance comparison quantifies the difference. Overall, the within school SES does not account for much of the variation in any of the countries. The SES in Finland explained 8.1 % and for Ireland explained 7.3% of reading achievement variation within the schools, which are the largest values from all the countries in our analyses. The within school variance explained by SES is computed by taking the difference in the residual student level variance between the Random-coefficient model and the ANOVA model. For Indonesia, Japan and the Netherlands the within school slopes of reading achievement on SES are quite flat and the SES accounts for little within school variation in reading achievement.

Table 7.5 also provides the estimates of the variances of the random effects and test of the hypothesis that these variances of the random effects are null. For all the countries that we tested, highly significant differences exist among the different school means as all τ_{00} differ significantly from zero. For Austria, Belgium and Taipei, we reject the null hypothesis, in this case that $\tau_{11}=0$, and infer that the relationship between SES and reading achievement within schools does indeed vary significantly across the population of schools. The model also produces the correlation between the intercepts and the slopes. For Austria and Taipei there is a negative relationship between the intercepts and slopes; in other words, in schools with higher reading achievement for mean SES, the within school slopes for SES tend to be less steep. For Belgium the correlation is positive

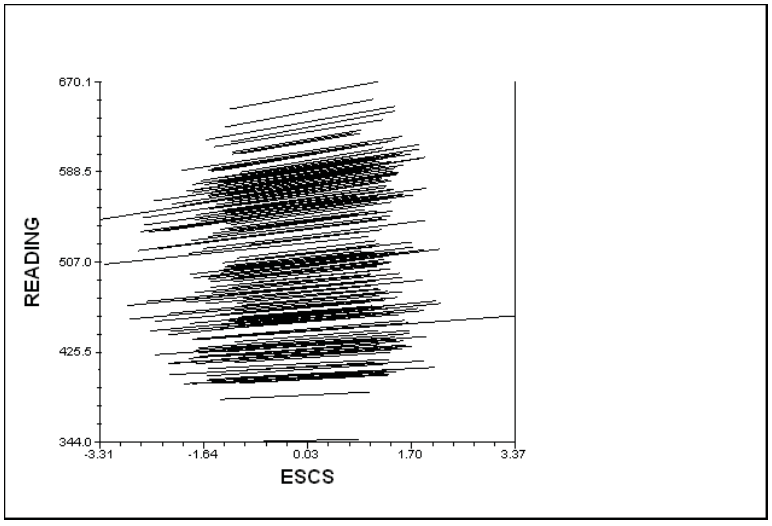

Figure 7.2. Within school regression of reading achievement on SES for the Netherlands

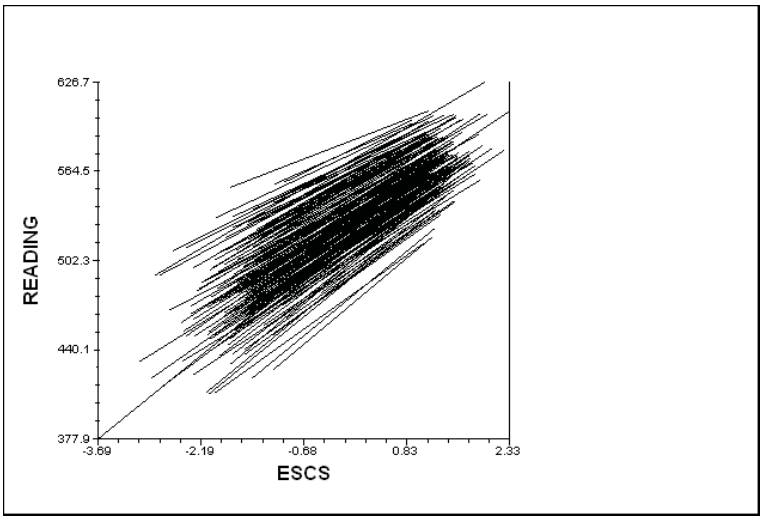

Figure 7.3. Within school regression of reading achievement on SES for Finland

such that for schools with higher reading achievement for mean SES the within school slopes for SES tend to be more steep. Thus in Austria and Taipei the choice of school makes less impact on the results of students with high SES while at the other end of the spectrum, in Belgium, the choice of school has the largest impact on achievements of students with high SES. In rest of the countries there is not enough variation in the slopes to estimate a correlation.

Next we study the outcomes of the random coefficient model separately for the public and private school types that are given in Table 7.6.

Table 7.6. The fixed effects from the Random coefficient model in public and private schools

	Public Schools				Private Schools			
Country	γ_{00}	γ_{10}	S.E.	τ_{01}	γ_{00}	γ_{10}	S.E.	τ_{01}
Austria	**448.4**	**12.0**	1.5	**-0.21**	**483.4**	6.3	4.2	N.A.
Belgium	**463.8**	**17.0**	2.1	**0.30**	**512.8**	**10.9**	1.3	**0.10**
Finland	**530.8**	**30.4**	1.6	N.A.	**532.6**	**29.4**	7.2	N.A.
Indonesia	**409.8**	**3.1**	1.1	N.A.	**388.4**	**3.0**	1.2	N.A.
Ireland	**475.4**	**31.0**	3.1	N.A.	**508.2**	**25.6**	2.2	N.A.
Japan	**520.2**	**3.41**	1.7	N.A.	**513.8**	**8.4**	3.3	**0.18**
The Netherlands	**518.8**	2.9	1.7	N.A.	**510.8**	**8.2**	1.4	N.A.
Taipei	**511.2**	**20.8**	1.8	**-0.22**	**472.4**	**3.9**	1.9	**-0.10**

The significant effects (at 5%) are highlighted in bold. For countries with N.A., there was insufficient variation in the slopes to determine a correlation.

τ_{01} is the correlation (standardized covariance) between intercepts and slopes

γ_{00} is the average of school means on reading achievement across the population of schools

γ_{10} is the average SES-reading achievement regression slope across the schools

Note that the intercept, which is a manifestation of the average reading level, is higher in public schools for Indonesia, Japan, the Netherlands and Taipei. This is contrary to the popular perception that private schools convey an academic advantage. As far as differences in the within school SES slopes are concerned, they are generally larger for the public schools and in the case of Taipei there is a large difference of about 5 times in the average within school SES slopes between public and private schools, with public schools having larger SES slopes.

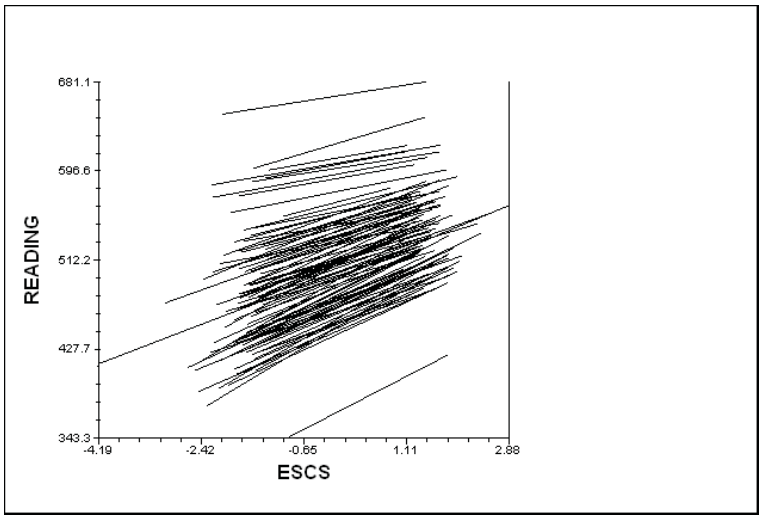

Figure 7.7. Within school regression of reading achievement on SES for Public schools in Taipei

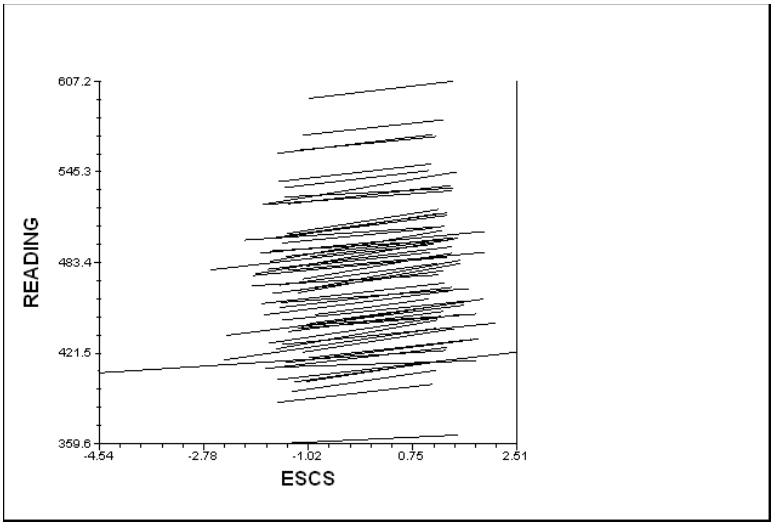

Figure 7.5. Within school regression of reading achievement on SES for Private schools in Taipei

Besides analyzing public and private sector schools separately, we also study schools with different orientations. We analyze two kinds of schools, with a 'general' orientation and with a 'vocational' orientation. The results are shown in Table 7.7.

Table 7.7. The fixed effect from the Random coefficient model for students in programs with a 'General' or 'Vocational' orientation

	General orientation				Vocational orientation			
Country	γ_{00}	γ_{10}	S.E.	τ_{01}	γ_{00}	γ_{10}	S.E.	τ_{01}
Austria	**463**	**9.3**	2.1	**0.16**	**448**	**11.0**	1.7	**0.08**
Belgium	**533**	**12.6**	1.3	**0.22**	**446**	**7.4**	1.5	N.A.
Indonesia	**401**	**2.8**	0.9	N.A.	**402**	3.2	2.0	N.A.
Japan	**515**	**3.8**	1.8	**0.07**	**526**	**8.9**	3.2	N.A.
Netherlands	**531**	**7.0**	1.2	N.A.	**431**	1.9	2.0	N.A.
Taipei	**512**	**17.9**	1.8	**-0.20**	**471**	**7.1**	1.9	N.A.

The significant effects (at 5%) are highlighted in bold. Results for Finland and Ireland are excluded for vocational orientation because of insufficient sample size. For countries with N.A., there was insufficient variation in the slopes to determine a correlation.

τ_{01} is the correlation (standardized covariance) between intercepts and slopes

γ_{00} is the average of school means on reading achievement across the population of schools

γ_{10} is the average SES-reading achievement regression slope across the schools

Note that the intercept, which is an indication of the average reading level, is higher in schools with a general orientation, except for Indonesia and Japan where the differences are insignificant. The differences are especially large for Belgium and the Netherlands.

So far we have analyzed the differences in the scores of school types and program orientations using the actual scores that did not take into account the socio-economic differences in their respective student populations. In Table 8 we present the average scores for public and private schools and for classrooms with a general and vocational program orientation before and after taking into account the differences in the socio-economic backgrounds of their respective student populations.

Table 7.8. Mean student scores for public and private school types and also separately for schools with General and Vocational program orientation.

	Public Schools		Private Schools		General Programs		Vocational Programs	
Country	Original Score	After Accounting	Original Score	After Accounting	Original Score	After Accounting	Original Score	After Accounting
Austria	468	465	500	457	519	474	454	461
Belgium	475	478	526	488	559	499	448	466
Finland	530	520	534	517	N.A.	N.A	N.A.	N.A.
Indonesia	412	465	391	460	402	457	404	466
Ireland	477	490	509	499	N.A.	N.A	N.A.	N.A.
Japan	524	535	514	495	519	533	528	509
Netherlands	521	482	511	483	547	498	429	437
Taipei	511	540	478	513	514	534	472	523

Results for Finland and Ireland are excluded for vocational orientation because of insufficient sample size.

From the results in Table 7.8 above it can be seen that for all the countries the average scores differences between different school types and program orientations change significantly after taking into account the socio-economic status of their student populations. For Austria, Belgium and Ireland there was a large difference in the scores of private schools above those of public schools but after taking into account the SES differences, the achievement differences almost disappear and in Austria the public schools even score fractionally higher. In Indonesia and the Netherlands the public schools had higher averages but after accounting for SES differences the public and private school scores become almost equal. In Taipei the public schools performed better than private schools before accounting for SES. After accounting for SES the average scores of both public and private schools increase by approximately the same margin. In Japan public schools had performed slightly better and after taking SES differences into account the gap between public over private schools only increased.

The change between student scores in vocational and general programs was even greater than the change observed between private and public schools after taking SES differences into account. The average scores of vocational schools in many countries improved tremendously after taking the SES differences into account. In Austria, Belgium, Netherlands and Taipei the average score differences between general and vocational tracks were huge. After taking SES differences into account the gap decreased significantly in all the four countries. In Austria and Taipei it virtually disappeared while in Belgium and the Netherlands it decreased by over one half. In Indonesia and Japan the actual scores were similar for general and vocational students. After accounting for SES, the differences in Indonesia remained the same while in Japan the students in general programs performed better than those in the vocational programs.

Table 7.9. Fixed effects of results from intercept-and-slopes-as-outcomes model

Country	γ_{01}	S.E.	γ_{02}	S.E.	γ_{03}	S.E.	γ_{11}	S.E.	γ_{12}	S.E.	γ_{13}	S.E.	τ_{01}
Austria	**108.6**	6.3	19.3	9.7	-2.0	6.8	1.8	3.2	5.9	4.8	-0.6	3.3	**-0.12**
Belgium	**89.3**	4.9	**-19.1**	5.3	**-57.0**	2.4	4.5	2.4	**8.3**	2.5	-4.1	2.1	**0.20**
Finland	**29.1**	6.5	4.4	8.8	___	___	-2.9	5.6	-0.1	7.7	___	___	N.A.
Indonesia	**35.9**	4.0	**18.2**	5.6	**17.0**	8.1	1.6	1.3	0.2	1.6	2.0	7.9	N.A.
Ireland	**74.0**	7.2	-10.9	6.6	___	___	-9.0	4.6	3.1	3.8	___	___	N.A.
Japan	**165.3**	8.7	**39.8**	7.1	-9.2	15.5	2.3	4.6	-5.6	3.7	2.0	7.9	N.A.
Netherlands	**90.8**	7.7	-3.2	7.1	**-45.0**	3.0	0.1	2.3	-3.9	2.2	-2.5	2.6	N.A.
Taipei	**89.5**	5.9	**19.5**	4.9	**-23.0**	3.8	-2.3	3.6	**14.3**	3.0	-4.5	3.2	**-0.17**

The significant effects (at 5%) are highlighted in bold. Results for Finland and Ireland are excluded for study orientation because of insufficient sample size. For countries with N.A., there was insufficient variation in the slopes to determine a correlation.

γ_{01} is the coefficient of Mean_SES in the intercept
γ_{02} is the coefficient of school type in the intercept (private schools= 0, public schools= 1)
γ_{03} is the coefficient of program orientation in the intercept (General= 0, vocational= 1)
γ_{11} is the coefficient of the Mean_SES in the slope
γ_{12} is the coefficient of school type in the slope(private schools= 0, public schools= 1)
γ_{13} is the coefficient of program orientation in the slope. (General= 0, vocational= 1)

Table 7.9 presents the results from the intercepts and slopes as outcomes model (see equations 7.6 and 7.7). We see from the table that all γ_{01} are significant. It yields that Mean SES is positively related to school mean reading achievement except after controlling for school type and study orientation. For Indonesia, Belgium and Taipei there was an independent effect of both orientation and school type on school average scores after controlling for the school Mean SES. For Belgium, after accounting for Mean SES, public schools are disadvantageous and if they have a vocational orientation it conveys a further disadvantage. For Indonesia after accounting for Mean SES, public schools are advantageous and if they have a vocational orientation it conveys a further advantage. Likewise for Taipei, public schools are advantageous and if they have a general orientation it conveys a further advantage. These results are new in the context of PISA outcomes. With regard to the within school SES slopes, there is a positive effect of public schools on the within school slopes in Belgium and Taipei. The differences in within school slopes for other countries are not impacted by any of the three background variables we employed in our analysis.

7.4 Conclusions

In this study we applied the Intercepts-and-slopes-as-outcomes method to examine the effects on educational achievement of socio-economic status of individuals and characteristics of schooling systems they are enrolled in. Previous literature suggested that such aspects can constitute initial and continuing barriers to educational achievement and we studied this in the context of the PISA dataset of students at the school leaving age. We investigated how different variables or educational settings like the type of school and the study orientation moderated the relationship between SES and achievement. PISA 2009 explored the impact of SES on student achievement using multi-level models but using only intercepts as outcomes and without incorporating the moderating variables we selected for this study. This study shed new light on equity related outcomes in the context of PISA outcomes also.

We started the analysis with the ANOVA and Means-as-Outcomes regression models. Results indicated that a sizeable portion of the total variance in student performance was at the school level or, in other words, there were sizeable differences in performance between schools (or intra class correlation) in all countries except Finland. This between school variance decreased sizably for all the countries (except Finland) after accounting for the Mean SES of the schools. This indicated that the difference in the average socio-economic characteristics of the school's students was a very important predictor of in-equity across schools.

Next we proceeded with the Random-effects model where we introduced a within school SES slope and a random intercept for the schools. The percentage of within school variance explained by individuals SES was much less than the percentage of between school variance explained by school Mean SES. The within school effects of SES on student performance were noticeable only for Finland and Ireland both of which had a low intra class correlation. One of the aims of this research was to study the correlation between the intercepts and slopes to see if the schools with a higher intercept also had higher slopes or vice versa. However there was noticeable variation in SES slopes across schools in only a few countries. For the remaining countries there was hardly any variation in SES slopes across schools, that is, the relationship between SES and performance was quite similar across schools in these countries and correlations could not be computed for these countries. Countries where there was a positive correlation between intercepts and slopes could be said to have lesser equity while countries with a negative correlation could be said to have more equity.

Our analysis led us to the recognition of socioeconomic status as a strong predictor of student achievement. This had been observed in past studies also. However using hierarchal models allowed us to separate the effects of socio-economic status within schools and between schools. We observed that in nearly all the countries there was a tendency of segregation of students in schools on the basis of SES and that explained a high proportion of between school performance differences. The within school differences of SES on performance were overshadowed by the between school differences in SES. Next we studied the how the type of school and study orientation moderated the between school differences resulting from the socio-economic characteristics of the schools.

Analyzing the private and publicly financed schools separately, we observed that in Austria, Belgium and Ireland there was a large difference in the mean scores of private schools over the public schools. However after taking the SES differences into account, the achievement differences between private and public schools virtually disappear and in Austria the public schools scored even score fractionally higher. In the remaining countries the public schools had performed better or equal to the private schools and that remained so even after taking SES differences of the student populations into account. Some earlier researchers had suggested that the advantage of private schools over public schools was largely due to the higher socio-economic characteristics of the student populations that are likely to be found in the private schools. Others had suggested that there were reasons to expect that private schools would have higher achievement than public schools, even after accounting for differences in their student population as private schools enjoy structural incentives over public schools such as parental selection, minimal bureaucracy and transfer possibilities in case the school fails to meet parental expectations. The notable findings of this study regarding the remarkable performance of public schools vis-à-vis private schools are significant, not just statistically, but in terms of their policy implications. In about half of the countries we analyzed, the public schools already performed better than private schools even before accounting for SES differences (and it remained so even after accounting for the SES differences). In the remaining countries where public schools had initially fared poorer, after accounting for the SES differences, the performance of the public schools became equal to or even better than that of the private schools. Thus the opinion that the higher scores of the private schools can be attributed to the higher SES students that private schools attract was confirmed in all the countries where private schools had performed better. While in other countries even the higher SES intake of the private schools did not convey an

academic advantage over public schools. In light of these results, the arguments about the inherent private schools advantage cited earlier can be dispelled.

Next we analyzed the differences between students tracked in different curriculums, namely with a general or vocational orientation. The key issue of tracking was whether this widely adopted educational practice would have impact on student's achievement, which in turn may lead to widened inequality among students tracked into different ability groups or types of programs. In this study we analyzed tracking vis-à-vis curriculum differentiation in the form of general or vocational study programs. While the advocates of tracking believe that tracking benefits students' academic performance, the detractors believe that tracking increases the achievement gap between tracked students. The main argument in favour of tracking was that homogeneous classrooms permit a focused curriculum and appropriately paced instruction that leads to the maximum learning by all students. The arguments against tracking were largely based on concerns that the lower groups will be systematically disadvantaged by slower learning environments that leave far behind the skills of those in the upper groups. Overall in the dataset that we analyzed, an average student in a general program performed better than his/her peers in the vocational programs and in countries like the Netherlands, Belgium, Austria and Taipei the differences were huge. However the average scores of vocational schools in these countries improved significantly after taking the SES differences into account. In the remaining countries, Indonesia and Japan, for which we had valid data, there were minor differences across general and vocational tracks and they didn't change much after accounting for SES. These findings lend support to one of the outcomes in previous literature that there tends to be a disproportionate assignment of low SES students to lower school tracks. However the main argument often cited in favor of tracking, that homogeneous classrooms lead to the maximum learning by all students did not hold in majority of the countries as there were large differences in average scores in general and vocational programs. The main argument against tracking that curriculum differentiation benefits initially higher achieving students at the cost of initially lower achieving students also did not hold across all the countries as in Austria and Taipei the differences became very narrow after accounting for SES alone. On the other hand in the Netherlands and Belgium substantial achievement differences still remained between the two tracks even after accounting for SES, suggesting that curriculum differentiation has not narrowed the gap between higher and lower tracks and supporting the argument that the slower learning environments of the (lower) vocational track were contributing to lower achievement. Though the results from some of these countries support the conclusion that ability grouping

with curriculum differentiation undermines the achievement of initially lower achieving students, research on this question has yet to completely solve difficult methodological issues, like how controlling variables can impact estimates of these effects. So these conclusions should be treated as tentative and reflecting the current state of knowledge, however open to revision as methodological advances allow more precise estimates of effects. What can be said with more certainty is that tracking is causing a disproportionate segregation of lower SES students in vocational streams and this contributes significantly to the visible achievement gap between high and low tracks.

We also analyzed SES and both the moderating variables together in the Intercepts-and-Slopes-as-Outcomes model. It transpired that school Mean SES was strongly related to school mean reading achievement in almost all the countries even after controlling for school type and study orientation. This was an important result that the average SES level of a school that a student attends is still the most important predictor of academic achievement after accounting for both the type of schools students attended and the curriculum they followed. For students in Indonesia, Belgium and Taipei there was an independent effect of both study orientation and school type on the intercepts after controlling for the school Mean SES. The within school SES slopes were only affected by the school type variable in Belgium and Taipei. The average within school slopes were not significantly different across vocational of general programs in any of the countries.

Coleman had asserted that the influence of student background was greater than anything that goes on within schools. Others had argued that the issue of economic status and its relationship to student achievement was more complex and it increased with student age because of school related factors and other policies like curriculum differentiation. Certain school factors like school climate and classroom discipline predictably affect student achievement and the size of their impact may vary for higher or lower SES students, but these factors are intrinsic to school level control and are not malleable variables at the system level. What is malleable from a system level policy perspective is that whether curriculum differentiation should be implemented or not and also which organizational model of schools is more effective and should be copied; should governments keeping pouring resources into the public school model or should a shift be made towards the private school model. The Intercepts-and-Slopes-as-Outcomes study enabled us to contextualize the impact of SES and these moderating variables on achievement at both the student and school level across different countries. The overall results of our study are in conformity with Coleman's assertion that student background was

the single most important factor in determining academic performance. The differences in average performances across different school organizational types were explained by differences in their students' socio-economic backgrounds. For curriculum orientation types, in some countries the differences virtually disappeared after accounting for SES, while in other countries differences still remained across curriculums types even after accounting for student background, indicating that there was a negative independent effect of curriculum differentiation on the lower tracks in such countries. However due to the non-availability of longitudinal data on student academic performance since they first underwent tracking (and also the possible impact of other controlling variables) the exact dynamics or progression of this effect over the years could not be analyzed. On the other hand, differences in school organizational models, between private and public school types, could be fully explained by the student background composition of schools. Thus, the presumed superiority of private-style organizational models, the private-school advantage was not supported by these results on reading achievement in any of the different countries we analyzed. The data suggest significant reasons to be suspicious of claims of general failure in the public schools, and raise substantial questions regarding a basic premise of the current generation of school reforms based on mechanisms such as choice and competition drawn from the private sector. In conclusion, this study presented a wide variety of analysis and discussed them in light of the previous literature to examine the relationship between SES and achievement and how it is moderated by different facets of educational organization using data from a number of countries. The results of this study expand on both the conceptual and empirical base available as research on this topic moves forward.

Summary

This thesis focuses on the application of item response theory (IRT) in the context of large scale international educational surveys like PISA (OECD), TIMSS and PIRLS (both IEA). Although IRT methodology has been widely used in educational applications such as test construction, norming of examinations, detection of item bias and computerized adaptive testing, large scale surveys present a number of specific problems. A number of these problems are addressed in this thesis. The procedures are illustrated using student questionnaire data of the 2006 and 2009 cycles of the PISA study.

The first problem in international comparative educational tests relates to the detection of cultural bias over countries. In this thesis, we targeted a problem know as country-specific Differential Item Functioning (CDIF) or country-by-item-interaction. Statistical tests to detect differential item functioning are available, but the huge number of students and countries presents feasibility problems related to the power of the tests and presentation and interpretation of the results. The power problem is related to the fact that with a sample size of students exceeding half a million even the tiniest model violation becomes significant. However, many well-founded test statistics (Orlando & Thissen, 2000; Glas & Suárez-Falcón, 2003) are based on residuals (differences between predictions from the model used and actual observations) that can shed light on the severity of the model violation. Further, this information can be used to model CDIF using country specific item parameters. In this approach, it is assumed that a scale consist of both items which are free of CDIF and items that may be subject to CDIF. The first set of items ensures the validity of the measure across countries. The second set of items is calibrated concurrently with the first set of items and both sets of items contribute to measurement precision. Residual analysis is used to establish that the two sets of items relate to the same latent variable, that is, the same construct, yet with different item parameters.

In Chapter 2, this methodology is outlined and applied to the field trial of the background questionnaires of the PISA 2009 cycle (this chapter was published in the *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, edited by Rutkowski, von Davier, and Rutkowski, as Glas & Jehangir (2014)). The impact of using country-specific item parameters was evaluated by comparing the ordering of

the countries on the latent variables measured without and with a model for CDIF. The analyses showed that certain scales of the student background questionnaire and the ICT questionnaire are indeed affected by the presence of CDIF. The scale most affected by CDIF was 'Like Reading'. Other scales where DIF was evident were 'Online Reading activities', 'Memorization strategies', 'Use of libraries' and 'Home use of ICT'. Correlations between ordering of countries showed that the detected CDIF did indeed have an impact. Finally, using either a model with or without discrimination parameters made little impact. The conclusions regarding CDIF items were not substantially affected by the model used.

However, besides in the background questionnaires, CDIF can also play a role in the assessment of cognitive outcomes. In fact, in an article in *Psychometrika*, titled '*Analysis of model fit and robustness, a new look at the PISA scaling model underlying ranking of countries according to reading literacy*', Svend Kreiner and Karl Bang Christensen (K&C) heavily criticize the methodology of the PISA project, both with respect to the use of the Rasch model (Rasch, 1960) and the presence of CDIF. According to K&C, their analyses provides strong evidence of misfit of the PISA scaling model and especially very strong evidence of CDIF in the PISA 2006 reading dataset. Based on these findings they assert that the country rankings reported by PISA 2006 are not reliable. In Chapter 3, K&C's main criticism concerning the impact of CDIF on the ranking of countries in PISA 2006 was investigated with the conclusion that the K&C critique is inappropriate. K&C had ignored or misrepresented some important methodological choices made in the PISA approach and based their analysis on a limited subset of the data such that their findings did not generalize to the PISA data at large. The results presented in this thesis showed that the impact of CDIF on the ranking of countries is far less prominent than suggested and becomes almost negligible when the statistical uncertainty regarding the country means is properly taken into account.

In Chapter 4, the practical significance of modeling item bias on the background questionnaire scales was studied not only in terms of the ordering of countries on the respective scales but also its impact on the results of regression analyses with latent variables in survey research. Chapter 4 was published in *Measurement: Journal of the International Measurement Confederation* (Jehangir, Van Den Berg, & Glas, 2015). We selected an important latent background scale 'Home Possessions' from PISA 2009 to study the impact of cultural bias or DIF in the framework of IRT. For our analyses we investigated the impact of DIF using different measurement models and scoring methods. We did this for the 1PLM and the 2PLM as the measurement models and the WLM and EAP scoring methods. We modeled DIF in the

framework of these different approaches using country specific item parameters and analyzed the results from the subsequent regression analyses. We observed that DIF in the latent scale under investigation impacts the results of the regression analyses and the use of country specific item parameters helps in mitigating the impact of DIF. The regression coefficients are affected more by DIF in the measurement model when using a model without discrimination parameters for estimating the latent covariates and less when using a model with different item discrimination parameters, so the use of country specific item parameters has greater need if the model without discrimination parameters is the measurement model. However as the ordering of the countries and the significances were quite similar for both models, which is often of interest in cross national surveys, either model may be used for future analyses provided DIF is adequately accounted for with country specific item parameters.

Another problem related to using IRT in large-scale international educational surveys pertains to issues of test administration. IRT gives flexibility in managing practical issues that a large scale survey entails. IRT separates person and item parameters and thus allows for the use of incomplete item administration designs (in educational measurement usually referred to as booklet designs) that support domain coverage through the administration of a large number of items while limiting the response burden of students. However, early in the history of the PISA project, it became clear that the positon of an item in a booklet influenced the item difficulty parameters. The problem was addressed by the introduction of so-called booklet parameters. These booklet parameters are assumed to be valid for all countries. In Chapter 5, the validity of this approach is evaluated by comparing it to alternatives, one that allows for country-by-booklet interaction, one using position parameters, and one using country-by-position interaction parameters. The conclusion is that booklet and position parameters improve model fit and that adding country-specific interactions is unnecessary.

The final topic relates to the combination of the results of IRT measurement models with multilevel structural models to relate cognitive outcomes to background variables. Several procedures to estimate these models are studied. A much used procedure is to generate so-called plausible values from the measurement model, that is, the IRT model, conditionally on principle components of background variables, and to estimate latent regression models conditional on these plausible values. Alternatives that are potentially more precise are concurrent estimation of the measurement and latent regression model and a two-step procedure where the measurement model is estimated first and the latent regression model is estimated using the item parameters obtained in the first step as covariates. The motivation for the plausible values

approach is that concurrent and two-step estimation methods are complicated and require dedicated software that is not generally available to practitioners. Therefore, in datasets like the PISA dataset, plausible values for outcome variables and maximum likelihood estimates for latent background variables are already provided in the dataset for use by secondary researchers. In Chapter 6, a study is reported that investigates if the results of the different estimation procedures lead to comparable inferences in latent regression analyses. According to our findings, results obtained using separately estimated plausible values and latent covariates (as provided in large scale data sets like PISA) were comparable with the results obtained using a concurrent estimation with raw data for the various kinds of IRT measurement models embedded in the structural (regression) model.

Finally, Chapter 7 gave an example of an advanced latent regression model based on plausible value methodology that explores the relation between socio-economic status and reading achievement in PISA 2009 through an intercepts-and-slopes-as-outcomes paradigm. Chapter 7 was published in the *Journal of Educational Research* (Jehangir, Glas, & Van den Berg, 2015). The aim of this study was to investigate an advanced methodological approach for analyzing educational achievement and not to draw ultimate conclusions on the state of educational equality in the population sample we studied. It was shown that some of the findings suggested new insights into equity related issues beyond those mentioned in the PISA 2009 international report. Several background aspects were selected that were deemed to have a bearing on educational attainment. The first was an attribute of the student, his or her socio-economic status. School level aspects included were school type (public, governmental private and non-governmental-private) and the orientation of the study program enrolled in. A main motivating factor in the intercepts-and-slopes-as-outcomes approach is that it enables a study of correlations between intercepts and slopes. That is, we can investigate whether the within-school relationship between SES and student performance differs according to the mean level of the students' scores. A problem was that for some countries there was not enough variance in the SES slopes. However, where this was not the case, the model gave new results and insights in the context of PISA; Taipei was a good example. The results showed a clear impact of study orientation and school type on the intercepts and slopes, and their relation. Some of the noteworthy results were that the within school SES effects were much weaker when compared to the between school effect of SES (except for Finland) and there was little variation in the within school SES slopes across many countries. The effect of school type was mixed across countries, in some countries public schools fared better while in other countries private

schools fared better. However after accounting for SES the public schools did equally good or better than private schools in all the countries, thus dispelling the notion of the inherent private school advantage. The correlation between intercepts and slopes gave a mixed pattern: some were positive and some negative. The study orientation had a marked effect on student performance with students in general programs doing much better than their counterparts in vocational programs in a number of countries. Even after accounting for SES, students in vocational programs performed worse than their peers in the general programs in a number of countries. The correlation between intercepts and slopes was positive in all countries and all study orientations (where significant), except for Taipei, where the correlation was negative for the general orientation.

Samenvatting

Dit proefschrift behandelt toepassing van IRT (item responsie theorie) in de context van grootschalige internationale onderwijs-surveys zoals PISA (OECD), TIMSS en PIRLS (beiden IEA). Hoewel IRT methoden op grote schaal worden gebruikt in onderwijskundige toepassingen zoals toets constructie, normering van examens, het opsporen van vraagonzuiverheid (Eng. differential item functioning, DIF) en adaptief toetsen via de computer, leidt de toepassing in grootschalige surveys tot een aantal specifieke problemen. Een aantal van deze problemen worden in dit proefschrift behandeld. De voorgestelde oplossingen worden geïllustreerd aan de hand van de data van de student-vragenlijsten van de 2006 en 2009 edities van PISA.

Het eerste probleem is het opsporen van DIF gerelateerd aan culturele verschillen tussen landen, ook bekend als CDIF (country-specific differential item functioning) of landen-bij-items-interactie. Hoewel er theoretisch goed gefundeerde statistische toetsen voor het identificeren van CDIF beschikbaar zijn, leidt het grote aantal studenten (meer dan een half miljoen) tot praktische problemen gerelateerd aan de power van de tests en de presentatie van de resultaten. Het power probleem is dat door de grote steekproef de kleinste afwijkingen van het model al significant zijn. Echter, de meeste statistische toetsen (Orlando & Thissen, 2000; Glas & Suárez-Falcón, 2003) zijn gebaseerd op residuen en grootte van deze residuen kan gebruikt worden om de ernst van de modelovertredingen vast te stellen. Verder kan deze informatie gebruikt worden om CDIF te modelleren met behulp van land-specifieke itemparameters. In deze benadering wordt verondersteld dat een schaal bestaat uit een combinatie van items met en zonder CDIF. De items zonder CDIF garanderen de validiteit van de schaal over de landen heen. De items met CDIF wordt simultaan met de eerste set items mee-geschaald, zodat beide sets items bijdragen aan de meetnauwkeurigheid voor het schaalconstruct. Residuen-analyse wordt gebruikt om aan te tonen dat de vragen van de schaal ook werkelijk gerelateerd zijn aan hetzelfde schaalconstruct, d.w.z. aan dezelfde latente variabele.

Deze methodologie wordt uiteengezet in Hoofdstuk 2 en toegepast op de vragenlijsten van de PISA 2009 editie (Dit hoofdstuk werd gepubliceerd in het Handbook of International Large-

Scale Assessment: Background, Technical Issues, and Methods of Data Analysis, geredigeerd door Rutkowski, von Davier, en Rutkowski (Glas & Jehangir (2014)). Het belang van land-specifieke itemparameters werd geëvalueerd door de ordening van landen op de latente variabelen met en zonder land-specifieke itemparameters te vergelijken. De analyses lieten zien dat een aantal schalen van de student-vragenlijsten en de ICT-vragenlijsten inderdaad CDIF vertoonden. Het grootste effect werd gevonden bij de schaal 'Like Reading' (Plezier in Lezen). Andere schalen waar CDIF werd gevonden waren 'Online Reading activities' (Online Lees Activiteiten), 'Memorization strategies' (geheugen strategieën), 'Use of libraries' (Bibliotheek Gebruik) en 'Home use of ICT' (ICT Gebruik Thuis). Uit de correlaties tussen de ordeningen van landen bleek dat CDIF inderdaad een aantoonbaar effect had. Verder werd aangetoond dat het gebruik van een IRT model met of zonder een discriminatie parameter geen effect had.

Naast in student- en ICT vragenlijsten kan CDIF ook een rol spelen bij de cognitieve toetsen die het vaardigheidsniveau in lezen, wiskunde en natuurkunde van de leerlingen meten. In een artikel in Psychometrika met de titel 'Analysis of model fit and robustness, a new look at the PISA scaling model underlying ranking of countries according to reading literacy', leveren Svend Kreiner en Karl Bang Christensen (K&C) zware kritiek op de methodologie van het PISA project, zowel op het gebruik van het Rasch model (Rasch, 1960) als op de aanwezigheid van CDIF. Volgens K&C laten hun analyses op de 2006 data voor het onderwerp lezen zien dat het in het PISA project gebruikte model erg slecht past en dat er een belangrijk effect van CDIF is. Gebaseerd op deze bevindingen stellen zij dat de ranglijst van landen zoals gerapporteerd voor PISA 2006 niet betrouwbaar is. In Hoofdstuk 3 is de kritiek van K&C onderzocht en verworpen. De belangrijkste bevinding was dat K&C hun conclusies baseren op een erg beperkte data set, terwijl de gevonden effecten zo goed als geheel verdwijnen als de hele data set geanalyseerd wordt, zeker als daarin ook nog de betrouwbaarheid van de schatting van de rangordening wordt betrokken.

In Hoofdstuk 4 wordt het effect van het modelleren van CDIF op multilevel regressie analyses met latente variabelen onderzocht. (Het hoofdstuk werd gepubliceerd in Measurement: Journal of the International Measurement Confederation (Jehangir, Van Den Berg, & Glas, 2015)). De invloed van de variabele 'Home Possesions' (Dingen in Huis) op de cognitieve uitkomsten in de 2009 PISA studie werd geëvalueerd. Hierbij varieerden we of de latente variabele voor de voorspeller geschat met een IRT model met en zonder discriminatie parameters, of CDIF al-dan-niet gemodelleerd was met land-specifieke itemparameters, en de methode voor het

schatten van de peroonsparameters, d.w.z. of WML (weighted maximum likelihood) of EAP (expected posterior) schatters voor de latente variabele voor de voorspeller gebruikt werden. De regressiecoëfficiënten in de verschillende latente regressie modellen werden beïnvloed door CDIF, maar meer in een model zonder dan in een model met discriminatie parameters. Het significantie-patroon onder de verschillende modellen was vergelijkbaar.

Een ander probleem in grootschalige onderwijskundige surveys hangt samen met de samenstelling en distributie van items over de studenten. Het gebruik van IRT voor de analyses levert op dit gebied een hoge mate van flexibiliteit. Doordat de invloed van de items en de studenten gemodelleerd zijn in afzonderlijke parameters, kan gebruik gemaakt worden van onvolledige designs, waarin niet alle leerlingen alle items beantwoorden. In onderwijskundig onderzoek worden dit design vaak een boekjes-designs genoemd. Ieder boekje bevat een (overlappend) deel van de items, en groepen leerlingen krijgen verschillende boekjes. Hierdoor kan het cognitieve domein dat door de items bestreken wordt breed zijn, zonder dat iedere leerling alle items hoeft te beantwoorden. Echter, al in de eerste ronde van PISA werd het duidelijk dat de positie van de items in een boekje invloed had op de moeilijkheidsgraad. Dit probleem werd opgelost door aan het IRT model z.g. boekjes-parameters toe te voegen. Deze parameters waren gelijk voor alle landen. De bedoeling van deze parameters is om te corrigeren voor het verschil in moeilijkheid van de verschillende boekjes. In Hoofdstuk 5 wordt een aantal alternatieven voor deze aanpak onderzocht. Het eerste alternatief is om boekjes-bij-landen interactie in het IRT model op te nemen. Het tweede alternatief is om de positie van een item in een boekje expliciet te modelleren. En het derde alternatief is om aan dit model ook weer interactie parameters voor landen-bij-positie interactie toe te voegen. De conclusie van dit onderzoek was dat zowel boekjes als positie parameters goede resultaten geven en dat het niet nodig is om ook nog interactie termen op te nemen.

Het laatste onderwerp, uitgewerkt in de hoofdstukken 6 en 7, heeft betrekking op multilevel latente regressieanalyse. Als voorbeeld worden modellen met de cognitieve variabelen als uitkomst en leerling-variabelen (plezier in lezen en leesstrategieën) en sociaaleconomische status (SES) als voorspellers gebruikt. Verschillende methoden om deze modellen te schatten zijn vergeleken. Een veelgebruikte methode is om zogenaamde ‘plausible values’ te genereren uit het meetmodel (het IRT model), conditioneel op de principale componenten van de achtergrondvariabelen. Vervolgens worden de latente regressie modellen met deze ‘plausible values’ als afhankelijke variabelen geschat. Alternatieven zijn het simultaan schatten van het

meetmodel en het latente regressie model en een twee-staps procedure waarin het meetmodel eerst geschat wordt, waarna de geschatte itemparameters als covariaten in het complete model (meetmodel plus latent regressiemodel) worden geïmputeerd. Deze twee methoden zijn theoretisch nauwkeuriger dan de benadering met 'plausible values', maar ze zijn ook complex en vereisen geavanceerde software. Daarom levert het PISA project voor secundaire analyses standaard 'plausible values' in de database. In Hoofdstuk 6 worden de resultaten van de verschillende methoden met elkaar vergeleken. De resultaten van de latente regressiemodellen bleken over het algemeen vergelijkbaar te zijn.

Tenslotte wordt in hoofdstuk 7 een 'slopes-as-outcomes' multilevel latente regressie model gepresenteerd voor het onderzoeken van de relatie tussen SES en leesvaardigheid. (Hoofdstuk 7 werd gepubliceerd in het Journal of Educational Research (Jehangir, Glas, & Van den Berg, 2015)). Het doel van dit hoofdstuk is om een geavanceerde methodologische aanpak te onderzoeken die nieuw licht kan werpen op onderwijskundige processen en het is niet de bedoeling om vergaande inhoudelijke conclusies te trekken. Drie achtergrond variabelen waarvan bekend is dat ze meeropbrengsten beïnvloeden werden geselecteerd. Op leerlingniveau, SES, en op schoolniveau, schooltype (openbaar, onafhankelijk privé, en overheidsafhankelijk privé) en studietype (algemeen versus beroepsgeoriënteerd). Een 'intercepts-and-slopes-as-outcomes' model heeft correlatietermen tussen de intercepten en de regressiecoëfficiënten van individuele regressiemodellen voor scholen. Hierdoor is te onderzoeken of de relatie tussen SES en lezen binnen een school afhangt van het gemiddelde prestatieniveau van een school. Verder kunnen deze relaties uitgesplitst worden naar schoolkenmerken, zoals schooltype en studietype. Een probleem bij de analyses was dat sommige landen niet genoeg variantie in de regressiecoëfficiënten voor SES hadden. Daar waar dit niet het geval was, gaf het model nieuwe interessante inzichten. Er bleek een duidelijke invloed van studietype and schooltype op de intercepten en regressiecoëfficiënten, en op hun relatie, te zijn. Opmerkelijk was dat het binnenschoolse effect van SES zwakker was dan het effect tussen scholen (behalve voor Finland, waar scholen zeer egalitair zijn). Overigens was er in de meeste landen zeer weinig variatie in binnenschoolse variantie van de coëfficiënten van SES. Het beeld van de invloed van school type varieerde sterk over landen. In sommige landen haalden openbare scholen betere leerresultaten dan privé scholen; in andere landen deden de privé scholen het beter. Echter, na het controleren voor SES deden in alle landen openbare scholen het even goed, of beter, dan privé scholen. Het effect van studietype was zoals verwacht: studenten met een algemene oriëntatie scoorden hoger dan studenten in een

opleiding met een beroepsoriëntatie, ook na controle voor SES. De correlatie tussen intercepten en regressiecoëfficiënten was positief in alle landen en alle studie typen, behalve in Taipei, waar de correlatie negatief was voor de algemene oriëntatie.

References

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficient multinomial logit model: A generalized form of the Rasch model. In M. v. Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57 - 76): Springer Verlag.

Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., & Syme, S. L. (1994). Socioeconomic status and health: The challenge of the gradient. American Psychologist, 49(1), 15-24.

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

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

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

Albert, J.H. and Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, 82, 747-769.

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007). Lasting consequences of the summer learning gap. American Sociological Review, 72(2), 167–180.

Ansalone, G. (2010). Tracking: educational differentiation or defective strategy. Educational Research Quarterly, 34(2) 3-17.

Archer, M.S. (1984). *Social Origins of Educational Systems*. Beverly Hills : Sage.

Berkson J. (1942) Tests of significance considered as evidence. JASA, 37:325–335.

Brennan, J & Osborne, M. (2008). Higher education's many diversities: of students, institutions experiences and outcomes. *Research Papers in Education*, 23(2), 170-190.

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

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer and G. N. Wilkinson (Eds.) *Robustness in Statistics* (pp. 201-236). New York Academic Press.

Bradlow, E.T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. Psychometrika, 64, 153-168.

Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical linear models*: *Applications and data analysis*. Newbury Park, CA: Sage.

Cabrera, A. F., & La Nasa, S. M. (2001). On the path to college: Three critical tasks facing America's disadvantaged. Research in Higher Education, 42(2), 119-149.

Camilli, G., & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items.* Thousand Oaks, CA: Sage.

Chall, J. S. (1996). American reading achievement: Should we worry? Research in the Teaching of English, 30, 303-310.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response function in polytomously scored item response models. *Psychometrika, 59,* 391-404.

Chao, R. K., & Willms, J. D. (2002). The effects of parenting practices on children's outcomes. In J. D. Willms (Ed.), Vulnerable children: Findings from Canada's National Longitudinal Survey of Children and Youth (pp. 149-166). Edmonton, AB: University of Alberta Press.

Chubb, J. E., & Moe, T. M. (1990). Politics, Markets, and America's Schools. Washington, D.C.: Brookings Institution.

Cohen, M. (1982). Effective schools: Accumulating research findings. American Education, 18(1), 13-16.

Coleman, J.S., Hoffer, T., & Kilgore, S. (1981*). Public and Private Schools. A Report to the National Center for Educational Statistics*. Chicago: National Opinion Research Center.

Coleman, J.S., Hoffer, T., & Kilgore, S. (1982). *High School Achievement: Public, Catholic, and Private Schools Compared.* New York: Basic Books.

Coleman, J.S., & Hoffer, T. (1987). *Public and Private High Schools: The Impact of Communities.* New York: Basic Books.

Coleman, J. S. (1997). The Design of Schools as Output-Driven Organizations. In R. Shapira & P. W. Cookson (Eds.), Autonomy and Choice in Context: An International Perspective (pp. 249-270). Oxford, UK: Pergamon.

Corten, R., & Dronkers, J. (2006). School achievement of pupils from the lower strata in public, private government-dependent and private government-independent schools: Across-national test of the Coleman-Hoffer thesis. *Educational Research and Evaluation,12,* 179-208.

Efron, B. (1977). Discussion on maximum likelihood from incomplete data via the EM algorithm (by A. Dempster, N. Liard, and D. Rubin). *J. R. Statist. Soc. B. 39,* 29.

Fox, J.-P. (2010). Statistics for Social and Behavioral Sciences. New York: Springer.

Fox, J.-P. & Glas, C.A.W. (2001). Bayesian Estimation of a Multilevel IRT Model using Gibbs Sampling. Psychometrika 66, 271-288.

Fox, J.-P. & Glas, C.A.W. (2003), Bayesian modeling of measurement error in predictor variables using Item Response Theory, Psychometrika, 68, 169--191.

Gamoran, A. (1992). The variable effects of high school tracking. American Sociological Review, 57(6), 812-828.

Gardner, M. J., & Altman, D. G. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed), March 15; 292(6522): 746–750.

Geerlings, H., Glas, C.A.W., & van der Linden, W.J. (2011). Modeling Rule-Based Item Generation. *Psychometrika, 76,* 337-359.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association 85,* 398-409.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). Bayesian Data Analysis, London, Chapman & Hall/CRC, Boca Raton, FL.

Glas, C.A.W., & Verhelst, N.D. (1995). Tests of fit for polytomous Rasch models. In G. H. Fischer & I.W. Molenaar (eds.). *Rasch models. Their foundation, recent developments and applications.* (pp.325-352). New York: Springer.

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

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

Glas, C.A.W., & Suarez-Falcon, J.C. (2003). A Comparison of Item-Fit Statistics for the Three-Parameter Logistic Model. *Applied Psychological Measurement, 27,* 81-100.

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

Glas C.A.W. (2010). MIRT, Software program and Manual. University of Twente, Enschede, The Netherlands. http://www.utwente.nl/gw/omd/en/employees/employees/glas/

Glas, C.A.W., & Jehangir, K. (2014). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.). *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis.* (pp. 97-115). New York, Springer.

Goldhaber, D.D. (1996). Public and Private High Schools: Is School Choice an Answer to the Productivity Problem? *Economics of Education Review,* 15(2), 93-109.

Guo, G., & Harris, K. (2000). The mechanisms mediating the effects of poverty on children's intellectual development. Demography, 37(4), 431-447.

Gunn (Eds.), Consequences of growing up poor (pp. 190-238). New York: Russell Sage Foundation.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential family response models. *The Annals of Statistics, 5*, 815-841.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*, 313-334.

Hanson, T. L., McLanahan, S., & Thomson, E. (1997). Economic resources, parental practices, and children's well-being. In G. J. Duncan & J. Brooks-

Harkreader, S., & Weathersby, J. (1998). Staff development and student achievement: Making the connection in Georgia schools. Atlanta, GA: The Council for School Performance.

Holland, P.W. and Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun (Eds.), *Test validity.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Holland, P.W., & Wainer, H. (1993). *Differential Item Functioning.* Hillsdale, N.J., Erlbaum.

Janssen, R., Tuerlinckx, F., Meulders, M. & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. Journal of Educational and Behavioral Statistics, 25, 285-306.

Jehangir, K., Van Den Berg, S.M., Glas, C.A.W. (2015). Correcting for differential item functioning in multi-level regression models in cross-national surveys. Measurement: Journal of the International Measurement Confederation, 71, 1-15.

Jehangir, K., Glas, C.A.W., van den Berg, S. (2015). Exploring the relation between socio-economic status and reading achievement in PISA 2009 through an intercepts-and-slopes-as-outcomes paradigm. Journal of Educational Research, 66, 263-271.

Johnson, V.E. & Albert, J.H. (1999). *Ordinal Data Modeling.* New York: Springer.

Kerckhoff, A.C. (1993). *Diverging pathways: Social structure and career deflections. Cambridge*, England; New York: Cambridge University Press.

Kerckhoff, A., Raudenbush, S., & Glennie, E. (2001). Education, cognitive skill, and labor force outcomes. Sociology of Education, 74(1), 1-24.

Kok, F.G., Mellenbergh, G.J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22,* 295-303.

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

Kulik, J.A.; Kulik, C.C. (1992). Meta-analytic findings on grouping programs. Gifted Children Quarterly, 36(2), 73-77.

Lareau, A. (2002). Invisible inequality: Social class and childrearing in black families and white families. American Sociological Review, 67(5), 747-776.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading M.A: Addison-Wesley.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, N.J., Erlbaum.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44*, 226-233.

Loveless, T.(1999). The tracking wars: State reform meets school policy. Washington, DC: Brookings Institution Press.

Lubienski, C. (2001). Redefining "Public" Education: Charter Schools, Common Schools, and the Rhetoric of Reform. Teachers College Record, 103(4), 634-666.

Lubienski, C. (2003). Instrumentalist Perspectives on the "Public" in Public Education: Incentives and Purposes. Educational Policy, 17(4), 478-502.

Lubienski, S.T., Lubienski, C. (2005). *Charter, Private, Public Schools and Academic Achievement: New Evidence from NAEP Mathematics Data*. A Report to the National Center for Educational Statistics.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) *WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility.* Statistics and Computing, 10:325--337.

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

Ma, X. (2000). Socioeconomic gaps in academic achievement within schools: Are they consistent across subject areas? Educational Research and Evaluation, 6(4), 337-355.

Macaskill, G (2008). *PISA TAG(0809)6a_1.doc: Alternative Scaling Models and Dependencies.* Available from mypisa.acer.edu.au.

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

Mirel, J., & Angus, D. (1994). "Rhetoric and Reality: American high Course Taking, 1928-1990" in *Learning from the Past: What History Teaches Us About School Reform*, Baltimore: Johns Hopkins University Press.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J., & Bock, R. D. (1989). A hierarchical item-response model for educational testing. In R. D. Bock (Ed.), *Multilevel analysis for educational data* (pp. 57-74). San Diego, CA: Academic Press.

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56 ,* 177-196.

Mislevy, R., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29 ,* 133-161.

Mislevy, R., Johnson, E., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17,* 131-154.

Molenaar, I. W. (1997). *Lenient or strict application of IRT with an eye on practical consequences.* In J. Rost & R. Langeheine (Eds.), Applications of latent trait and latent class models in the social sciences (pp. 38-49). Münster: Waxmann.

Muthén, L.K. and Muthén, B.O. (1998–2012) *Mplus User's Guide*, 7th ed., Muthén & Muthén, Los Angeles, CA.

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

Oaks, J. (1985).Keeping track. New Haven: Yale University Press.

OECD (2007). *PISA 2006 Science Competencies for Tomorrows World, Vol. 1: Analysis*. OECD, Paris.

Oliveri, M. E. & von Davier, M. (2011). *Investigation of Model Fit and Score Scale Comparability in International Assessments.* Psychological Test and Assessment Modeling, 53 (3) 315-333. Retrieved 9/29/2011 from: http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Stiuka, M. F. (1994). Ability-group effects: Instructional, social, or institutional? Sociology of Education, 67(1), 27-46.

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

Plummer, M. (2003). *JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling*. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna, Austria.

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

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Raudenbush, S.W., & Kasim, R.M. (1998). Cognitive skill and economic inequality: Findings from the National adult Literacy Study. Harvard Educational Review, 68(1), 33-79.

Rubin, D.B (1978). Multiple Imputations in sample surveys —A phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section* (pp.20-34). Alexandria, VA American Statistical Association.

Rubin, D.B. (1987). *Multiple imputation for non response in surveys*. New York: Wiley.

Rutter, M., & Maughan, B. (2002). School effectiveness findings 1979-2002. Journal of School Psychology, 40(6), 451-475.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph Supplement, No. 17.*

Schofield, J.W. (2010). International evidence on ability grouping with curriculum differentiation and the achievement gap in secondary schools. Teachers College Record, 112(5), 1492-1528.

Schwarz, G. E. (1978), "Estimating the dimension of a model", *Annals of Statistics* **6** (2): 461–464.

Slavin, R.E. (1990). Achievement effects of ability grouping in secondary schools: A best evidence synthesis. Review of Educational Research, 60 (3), 471-499.

Schofield, J.W. (2010). International evidence on ability grouping with curriculum differentiation and the achievement gap in secondary schools. Teachers College Record, 112 (5), 1492-1528.

Scheerens, J. (1992). Effective schooling: Research, theory, and practice. London, UK: Cassell.

Schutz, G., Ursprung, H.W., & Woßmann, L. (2008). *Education Policy and Equality of Opportunity*, Kyklos, Wiley Blackwell, vol. 61(2), 279-308.

Silvey, S. D. (1959). The Lagrangian multiplier test. *Annals of Mathematical Statistics, 30*, 389-407.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417–453.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). "Bayesian measures of model complexity and fit (with discussion)". *Journal of the Royal Statistical Society, Series B, 64* (4): 583–639.

Thomas, N., & Gan, N (1997). Generating multiple imputations for Matrix Sampling Data Analyzed with Item Response Models. *Journal of Educational and Behavioral Statistics*, 22(4), 425-445.

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

Van de Werfhorst, H.G. & Mijs, J.J. B. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. Annual Review of Sociology, 36, 407-428.

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

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika, 54,* 427-450.

Weeks, J., von Davier, M. & Yamamoto, K. (2013). *Design Considerations for the Program for International Student Assessment.* In Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). Handbook International Large-scale Assessment: Background, Technical Issues, and Methods of Data Analysis. Taylor & Francis.

White, K,R. (1982). The relation between socioeconomic status and academic achievement, *Psychological Bulletin, 91(3)* 461-481.

Willms, J. D. (Ed.). (2002).Vulnerable children: Findings from Canada's National Longitudinal Survey of Children and Youth. Edmonton, AB: University of Alberta Press.

Wilson, D., Burgess, S. and Briggs, A. (2006): "The Dynamics of School Attainment of England's Ethnic Minorities", CASE Papers /105, Centre for Analysis of Social Exclusion, LSE.

Yeung, W. J., Linver, M. R., & Brooks-Gunn, J. (2002). How money matters for young children's development: Parental investment and family processes. Child Development, 73(6), 1861-1879.