Statistical Modeling in International Large-scale Assessments

Inga Laukaitytė

Department of Statistics
Umeå School of Business and Economics
Umeå 2016

Doctoral Thesis
Department of Statistics
Umeå School of Business and Economics
Umeå University
SE-901 87 Umeå

Statistical Studies No. 51
ISBN: 978-91-7601-612-1
ISSN: 1100-8989
Cover: Salma Ben Ayed
Electronic version available at http://umu.diva-portal.org/

Printed by: UmU-tryckservice
Umeå, Sweden 2016

Table of Contents

List of Papers

The thesis is based on the following papers:

I. Laukaityte, I., and Wiberg, M. (2016). The Importance of Sampling Weights in Multilevel Modeling of International Large-Scale Assessment Data. *Manuscript.*

II. Laukaityte, I., and Wiberg, M. (in press). Using plausible values in secondary analysis in large-scale assessments. Accepted to *Communications in Statistics – Theory and Methods*.

III. Laukaityte, I. (2016). Single Imputation from a Conditional Distribution vs Multiple Imputation for Data with a Non-monotone Missing Pattern. *Manuscript.*

IV. Laukaityte, I, and Rolfsman, E. (2016). Student performance in PISA 2003-2012 in Mathematics in the Nordic countries. *Manuscript.*

Abstract

This thesis contributes to the area of research based on large-scale educational assessments, focusing on the application of multilevel models. The role of sampling weights, plausible values (response variable imputed multiple times) and imputation methods are demonstrated by simulations and applications to TIMSS (Trends in International Mathematics and Science Study) and PISA (Programme for International Student Assessment) data.

The large-scale assessments use multistage sampling design, which means that the units such as schools, classrooms, or students at some or all stages are selected with unequal probabilities. In order to make valid estimates and inferences sampling weights should be used. Thus, in the first paper, we examine different approaches and give recommendations concerning handling sampling weights in multilevel models when analyzing large-scale assessments.

Due to limitations in time and the number of students, the complex surveys use matrix sampling of items. This means that a response variable, i.e. students' performance, contains a large amount of information that is missing by design. Therefore, in order to estimate students' proficiency, TIMSS and PISA use the plausible values approach, which results in a set of five plausible values – proficiencies, computed for each student. In the second paper, different user strategies concerning plausible values for multilevel models as well as means and variances are examined with both real and simulated data. Missing information that is present because of the matrix sampling design for instance like the one used in PISA, can be arranged into a non-monotone missing data pattern, where all variables are incomplete and highly positively correlated. In the third paper, we compare a few imputation methods: a single imputation from a conditional distribution (with and without weights) and multiple imputation, for data with a non-monotone missing pattern (with no complete variables) and high positive correlation between variables.

In several of the recent international large-scale assessments, students in Sweden demonstrate a decreasing performance. Some previous research has shown that changes in performance depend on students' performance levels. In the fourth paper, we studied the relationship between student performance and the between-school variance and tried to identify factors associated with student performance in mathematics in PISA in low-, medium-, and high-performing schools in the Nordic countries.

KEYWORDS: multilevel model; plausible values; sampling weights; missing information; multiple imputation; non-monotone missing pattern; TIMSS; PISA

Sammanfattning

Att testa och jämföra elever, lärare, medborgare, eller liknade i nationellt och internationellt perspektiv har blivit allt vanligare. Den här typen av undersökningar omfattar allt från småbarnsuppfostran till vuxnas kunskaper med varierande fokus från studenter till lärare. I denna avhandling används data från två internationella storskaliga komparativa mätningar: TIMSS (Trends in International Mathematics and Science Study) och PISA (Programme for International Student Assessment). Syftet med TIMSS är att jämföra och beskriva elevers kunskaper inom matematik och naturvetenskap samt deras inställning till dessa ämnen för att förbättra lärandet hos eleverna. Syftet med PISA är att undersöka i vilken utsträckning elever är förberedda på att klara sig i samhället, genom att undersöka effekten av utbildning inom läsning, matematik och naturvetenskap. Både TIMSS och PISA syftar till att beskriva, jämföra och förstå elevers prestationer inom och mellan länder samt över tid. De internationella storskaliga komparativa mätningarna TIMSS och PISA är mycket komplexa i sina designer och analys av sådana data kräver därmed avancerade statistiska analysverktyg. För att ta hänsyn till datas hierarkiska struktur kan exempelvis flernivåanalys användas. Syftet med avhandlingen är att undersöka hur man ska hantera komplexiteten av storskaliga komparativa studier när man vill använda flernivåanalys.

De storskaliga komparativa mätningarna använder en stickprovsdesign i flera steg, vilket innebär att enheter såsom skolor, klassrum eller studenter vid några eller alla steg väljs med olika sannolikheter. För att kunna göra tillförlitliga uppskattningar och dra giltiga slutsatser ska stickprovsvikter användas. Således, i det första pappret, undersöks olika metoder för hantering av stickprovsvikterna i flernivåmodeller vid analys av storskaliga komparativa mätningar och rekommendationer ges.

På grund av begränsningar i tid och antalet studenter, så använder de komplexa mätningarna så kallad matrissampling av uppgifter. Detta innebär att en responsvariabel, dvs. elevernas provresultat, innehåller en stor mängd både ofullständig och avsiktligt saknad information. För att uppskatta elevernas kunskaper så använder TIMSS och PISA en metod som resulterar i fem s.k. plausibla värden, dvs. prestationsvärden beräknade för varje elev. I det andra pappret, så undersöks, med hjälp av både verklig och simulerad data, olika användarstrategier vi användning av plausibla värdena för medelvärden, varianser och när flernivåmodeller används.

Saknad information på grund av användandet av sampling i flera steg, exempelvis som den som används i PISA, kan ordnas i ett icke-monotont datamönster över saknad information, där alla variabler är ofullständiga och högt positivt korrelerade. I det tredje pappret, jämför vi några imputeringsmetoder: en enkel imputation från en betingad fördelning (med

och utan stickprovsvikter) och multipel imputation, för data med ett icke-monotont avsaknat mönster (utan fullständiga variabler) och hög positiv korrelation mellan variablerna.

I flera av de senaste internationella komparativa mätningarna, uppvisar elever i Sverige en minskande prestation. Tidigare forskning har visat att förändringar i prestationen beror på elevens prestationsnivå. I det fjärde pappret studeras förhållandet mellan elevernas resultat och mellanskolvariansen och ett försök görs att identifiera faktorer som är förknippade med elevernas matematikresultat på PISA i låg-, medel- och högpresterande skolor i de nordiska länderna.

Preface

Eleven years ago, when I just had finished my master studies in statistics, if somebody told me that I would become a PhD, I would not really believe. Moreover, if somebody told me that I would do this twice, I would consider this nothing more than a mere daydream. After eleven years, today is the day: writing this preface is the finishing touch on my second thesis. It has been an amazing journey through exploration of a new scientific knowledge, and not less myself. This journey would not have been so fascinating if not the people who have supported and helped me very much.

I would first like to thank my supervisor Marie Wiberg. Your inexhaustible energy, always positive attitude and smile have all along the way been for me a source of light and inspiration. Thank you for your great support and valuable advices, for leading me through the research and at the same time letting me find my own ways to reach the goals, for encouraging me when I was close to surrendering.

I would also like to thank my co-supervisors, Kenny Bränberg for reading my papers many times and for insightful comments, and Ewa Rolfsman for knowledgeable advice and inspiring cooperation on our paper.

I also want to thank Sture Holm, for sharing ideas and for valuable and constructive suggestions that greatly improved one of my papers.

In addition, I would like to thank all my co-workers at the department of Statistics. All you make this working place into cozy and inspiring. A special thanks to all my office-mates; for all the joyful moments and serious conversations (not about statistics), for your patience to my attempt to speak Swedish. Anita, it was a pleasure to share with you our girlish side of the office and I am very grateful for all your help, especially with brushing up my Swedish speaking/writing skills. Philip, thanks for broadening my knowledge in the fields of comics, jokes, songs, books, etc. By the way, have you seen my hat?

The life outside the university is also crucial for a good mental state. My life in Umeå, far away from homeland, would be quite dark and boring if not some special friends. One hundred pages would not be enough to thank them for everything, but I will try to make it short. Johan and Lisbet, thank you for amazing dinners, games and fun discussions! Jūratė and Andrius, thank you for all the support, food, trips and much much more.

And finally, whom would I be without my family. I am very grateful to my husband Anouar for encouraging, supporting, cheering me up and just being with me during this journey, and my wonderful son Jonus for filling

my days with sunshine, joy and laughter. I would also like to thank my mother for being always there whenever I needed her the most, even though we are one thousand kilometers away. Mama, ačiū už tai, kad esate!

Umeå, November, 2016
Inga Laukaitytė

Introduction

Testing and comparing students, teachers, citizens, etc. from a national and international perspective is quite popular nowadays. Educational surveys cover everything from early childhood education to adult skills with a focus varying from students to teachers. Large-scale educational assessments are complex surveys that employ multistage sampling designs, matrix sampling of items and plausible values representing the performance of students. Thus, analysis of such data requires advanced statistical techniques. For example, in order to take a multistage sampling design into account, multilevel modeling can be used. The aim of the thesis is to examine how to handle the complexity of the large-scale assessments in multilevel analysis.

Multistage sampling design in the large-scale assessments means that the units such as schools, classrooms, or students at some or all stages are selected with unequal probabilities. Hence, sampling weights have to be included in the analysis of the data. The sampling weights supplied by large-scale assessments are constructed for single-level analysis, and, thus, it is not trivial how to use them in multilevel models. In Paper I, we examine different approaches such as various cases of informative weights and various complexities of sampling designs, and give recommendations concerning handling sampling weights in two-level models when analyzing data of large-scale assessments.

Large–scale assessments contain large numbers of items as well as limited time and numbers of students. Considering these limitations, matrix sampling of items is used; therefore, students receive a small subset of all assessment items. In such a case, the measurement of individual proficiency is achieved with a measurement error (von Davier, Gonzalez & Mislevy, 2009). In order to reflect the uncertainty of the measurement, a set of scores referred to as plausible values (PVs) is computed for each student. Paper II demonstrates the role of PVs in large-scale assessments when multilevel modeling is used. Missing information that is present due to the matrix sampling design used in such assessments can be arranged into a non-monotone missing data pattern, where all variables are incomplete and highly positively correlated. There are a number of comparative studies of imputation methods, among them, those for non-monotone missing patterns (e.g. Horton & Kleinman, 2007; Wilson & Lueck, 2014; Durrant, 2009); however, all of them use data containing some or at least one complete variable. Therefore, in Paper III, we compare a single imputation from a conditional distribution (with and without weights) and multiple imputation for data with a non-monotone missing pattern (with no complete variables) and high positive correlation between variables.

Decreasing performance among students in Sweden on international large-scale assessments and increasing segregation of schools has led to many discussions concerning strategies for improving student performance. Previous research has shown that changes in student performance differ depending on the performance level of students. For example, the largest decreases in student performance are seen for low- and medium-performing students in Sweden. In the last paper of this thesis, we studied the relationship between student performance and between-school variance, and tried to identify factors associated with student performance in mathematics in PISA (Programme for International Student Assessment) in low-, medium-, and high-performing schools in the Nordic countries.

The thesis is structured as follows. In the section that follows, the large-scale assessments that are used in the thesis are briefly presented and compared. Next, the sampling design and sampling weights employed by the large-scale assessments are described, followed by an introduction to scaling methodology. In the fifth section, a brief overview of multilevel models is presented. Finally, the papers in the thesis are summarized, and some concluding remarks including further research are given.

Large-scale assessments in education: TIMSS and PISA

Data from two international large-scale assessments are used in the thesis: TIMSS (Trends in International Mathematics and Science Study) and PISA.

TIMSS is a quadrennial international comparative assessment of the mathematics and science knowledge of fourth- and eighth-grade students all over the world. TIMSS was performed for the first time in 1995 and was the largest international student assessment at that time. There were 29 participating countries for the fourth grade and 46 for the eighth grade. The most recent TIMSS assessment took place in 2015 and had 49 participating countries for the fourth grade and 38 for the eighth grade. TIMSS is devoted to helping countries to improve teaching and learning in mathematics and science. It consists of assessments in different mathematics and science domains (e.g. number, algebra, geometry, physics, biology, etc.), as well as student, teacher and school questionnaires.

PISA is a triennial international assessment that aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. In the first PISA assessment, carried out in 2000, 32 countries participated. The most recent PISA assessment performed in 2015 involved more than 70 countries/economies. PISA consists of assessments in three

areas: reading, mathematics and science; as well as questionnaires filled in by students, parents and school principals.

Scores in both assessments are scaled so that the mean of the overall achievement distribution is 500 and the standard deviation is 100. Achievement data from subsequent cycles are linked to the previous cycles through IRT (item response theory) scale linking methods.

The two assessments mentioned are similar in the sense of areas assessed, but are quite different in their goals and design. TIMSS is curriculum-based with a focus on formal mathematical knowledge, whereas PISA is not directly linked to a school curriculum and emphasizes the application of mathematics in real-life situations. Moreover, TIMSS treats both areas equally, whereas PISA has a different emphasis on the areas. In each assessment, one of the assessed areas is chosen as the major domain and is given greater prominence, while the other two areas are assessed less thoroughly. The assessments also differ in their sampling design. The TIMSS assessment samples schools and classes, mostly including all students in the class, where PISA samples schools and students (not classes).

Sampling design and sampling weights

Most large-scale assessments, including TIMSS and PISA, employ the two-stage random sampling design. In such sampling design, the population is partitioned into groups (or clusters) and a simple random sample of the groups is selected. Afterwards, a simple random subsample of elements is sampled within each of the selected groups. Two-stage sampling is mostly used when the sizes of the groups are very large, making it difficult and/or expensive to observe all the units inside them.

Prior to the sampling procedure, in TIMSS and PISA, the schools are stratified, i.e. the schools in the target population are arranged into strata that share some common characteristics. Two types of stratification are used: explicit and implicit. In explicit stratification, schools are grouped into strata that will be treated independently from one another or as if they were separate school sampling frames. Regions of a country, type of school, and size of school are some examples of explicit stratification variables in TIMSS and PISA. Implicit stratification consists of sorting the schools within each explicit stratum or within the entire sampling (if explicit stratification is not used) by one or more implicit stratification variables. Geographic location, type of school, school gender, degree of urbanization, or minority composition is examples of possible implicit stratification variables. This type of stratification is a very effective and simple way of ensuring a proportional sample allocation of students across all implicit strata. Implicit stratification can also lead to improved reliability of

achievement estimates, if the implicit stratification variables are correlated with student achievement at the school level (Jaeger, 1984; OECD, 2012).

After stratification, both assessments, TIMSS and PISA, sample schools at the first sampling stage. Schools are sampled from a national list of all schools in the country that have students enrolled in the target grade (TIMSS) or schools having 15-year-old students (PISA), with probabilities that are proportional to size. The second stage is different in the two assessments. TIMSS samples classes, i.e. within each sampled school, all classes with students at the target grade are listed, and one or more intact classes are selected with an equal probability of selection using systematic random sampling. PISA samples students, i.e. within each sampled school, all 15-year-old students are listed, and typically, 35 students are selected with equal probability. In the schools having fewer than 35 15-year-old students, all students on the list are selected (Joncas, 2008; OECD, 2012).

Both assessments require a minimum of 150 schools to be selected in each country. In participating countries that have fewer than 150 schools, all schools are selected.

Weights

The two-stage sampling design used in TIMSS and PISA means that each student in the target population is chosen with unequal probability. In order to make valid estimates and inferences of the target population and to calculate appropriate estimates of sampling error, sampling weights should be used.

TIMSS

Sampling weights in the TIMSS assessment are calculated according to a three-step procedure involving selection probabilities for schools, classrooms, and students (Joncas, 2008). TIMSS offers six sets of weights, which are downloadable with the data:

- three versions of overall student sampling weights: the total student weight (*TOTWGT*), the student house weight (*HOUWGT*), and the student senate weight (*SENWGT*);
- the overall and by area (mathematics and science) teacher weights (*TCHWGT*, *MATWGT*, and *SCIWGT*);
- the school weight (*SCHWGT*); and
- the sum of student weights (*STOTWGT)*.

The difference between the three overall student sampling weights is in scaling. *TOTWGT* sums to each national population, *HOUWGT* sums to the student sample size in each country, and *SENWGT* sums to 500 in each country.

The overall student sampling weight is the product of the three basic weights reflecting selection probabilities and nonparticipation adjustments. Thus, the total student weight for class *k* in the school *j* can be expressed as

$$TOTWGT_{j,k} = w_{sc}^{j} \cdot A_{sc} \cdot w_{cl}^{j} \cdot A_{cl} \cdot w_{st}^{j,k} \cdot A_{st}^{j,k},$$

where w_{sc}^{j}, w_{cl}^{j}, and $w_{st}^{j,k}$ are the basic school, class and student weights respectively. The basic weight is the inverse of the probability of selection at that level. A_{sc}, A_{cl}, and $A_{st}^{j,k}$ are the weighting adjustments for non-participation of schools, classes, and students, respectively. A more detailed description of the sampling weights can be found in Joncas (2008).

PISA

PISA data set contains four weight variables. The final student weight (*W_FSTUWT*), the senate student weight (*SENWGT_STU*), the school weight (*W_FSCHWT*), and the senate school weight (*SENWGT_SCQ*). *SENWGT_STU* sums to 1000.

The final student weight, W_FSTUWT_{ji}, for student *i* in school *j*, is composed of the school base weights, the within-school base weights, and five weighting adjustments. It is defined as

$$W\_FSTUWT_{ji} = w_{1j} \cdot w_{2ji} \cdot f_{1j} \cdot f_{1ji}^{A} \cdot f_{2ji} \cdot t_{1j} \cdot t_{2ji},$$

where w_{1j} and w_{2ji} are the school and within-school (student) base weights respectively. As in the TIMSS case, the base weight is the inverse probability of selection at that level. f_{1j} is a weighting adjustment to compensate for non-participation by other schools that are somewhat similar in nature to school *j*, f_{1ji}^{A} is a weighting adjustment to compensate for schools in some participating countries where only 15-year-old students who were enrolled in the modal grade for 15-year-old students were included in the assessment, f_{2ji} is a weighting adjustment to compensate for non-participation by students within the same school non-response cell and explicit stratum, and where permitted by the sample size, within the same high/low grade and gender categories, t_{1j} is a school base weight trimming factor, used to reduce unexpectedly large values of w_{1j}, t_{2ji} is a final student weight trimming factor, used to reduce weights of students with exceptionally large values for the product of all the preceding weight components (OECD, 2012a).

Scaling methodology

The aim of large-scale assessments is to get a maximum of information with a minimum time spent on a test. Thus, participants are given only a small sample of all the items and their responses are placed on a common scale to provide an overall picture. This is called matrix-sampling of items or a

rotated block design. In such sampling design, assessment items are assigned to a number of blocks that are then combined in systematic ways into a set of booklets, with each student completing just one booklet. In the TIMSS assessment, items are grouped into 28 blocks, with half of blocks containing mathematics items and the other half containing science items. Each student booklet is a combination of two mathematics and two science item blocks. Thus, each participating student is given a sample of items from both assessed areas. In the PISA assessment, items are grouped into thirteen clusters. For example, in PISA 2012, there were seven mathematics clusters and three each of reading and science clusters. PISA also offers the option of administering an easier set of booklets. This option is offered for countries that had previously achieved or are expected to achieve a mean scale score in the area on focus of 450 or less. In the easier set of booklets, two standard clusters are substituted with two easier clusters. For example, standard mathematics clusters M6A and M7A in PISA 2012 are substituted with easier clusters M6B and M7B. Each student booklet is composed of various combinations of four clusters. Thus, in PISA, only some of the students are given items from all the three areas, the others take part solely in one (the area on focus) or two areas (the area on focus and one from the other two areas). Nevertheless, all students are assessed in all the three areas, even though they had no items from some of them. Examples of student achievement booklet designs in TIMSS and PISA are given in Figure 1.

TIMSS 2011

Assessment blocks				
Booklet ID	**Part 1**		**Part 2**	
Booklet 1	M01	M02	S02	S02
Booklet 2	S02	S03	M02	M03
Booklet 3	M03	M04	S03	S04
Booklet 4	S04	S05	M04	M05
Booklet 5	M05	M06	S05	S06
Booklet 6	S06	S07	M06	M07
Booklet 7	M07	M08	S07	S08
Booklet 8	S08	S09	M08	M09
Booklet 9	M09	M10	S09	S10
Booklet 10	S10	S11	M10	M11
Booklet 11	M11	M12	S11	S12
Booklet 12	S12	S13	M12	M13
Booklet 13	M13	M14	S13	S14
Booklet 14	S14	S01	M14	M01

PISA 2012

Standard test booklets				
Booklet ID	**Cluster**			
Booklet 1	M5	S3	M6A	S2
Booklet 2	S3	R3	M7A	R2
Booklet 3	R3	M6A	S1	M3
Booklet 4	M6A	M7A	R1	M4
Booklet 5	M7A	S1	M1	M5
Booklet 6	M1	M2	R2	M6A
Booklet 7	M2	S2	M3	M7A
Booklet 8	S2	R2	M4	S1
Booklet 9	R2	M3	M5	R1
Booklet 10	M3	M4	S3	M1
Booklet 11	M4	M5	R3	M2
Booklet 12	S1	R1	M2	S3
Booklet 13	R1	M1	S2	R3

Figure 1. Examples of student achievement booklet design in TIMSS 2011 and PISA 2012. M – mathematics, S – science, R – reading.

Item response theory models

In the complex international large-scale assessments, the student achievement must be estimated on the entire assessments considering the incomplete or missing information. In order to define student achievement on an assessment and to provide accurate measures of trends, international studies rely on item response theory (IRT) scaling. IRT models the relationship between an unobserved variable, commonly an examinee's ability, and the probability of the examinee responding correctly to any item in the test (e.g., Harris, 1989). Different assessments use different IRT methods. The TIMSS assessment uses a two-parameter logistic IRT model for the constructed response items that were scored dichotomously, a three-parameter logistic IRT model for multiple choice items, and a generalized partial credit model with the constructed response items that were scored polytomously (Yamamoto & Kulick 2000). In PISA, data are scaled using the mixed-coefficients multinomial logit model (a generalized form of the Rasch model) which is a categorical response model (Adams & Wu, 2007).

One-, two- and three-parameter IRT models

The three-parameter model employs three item parameters: discrimination (a), also referred to as the slope of the item characteristic curve, that allows items to differentially discriminate among examinees; item difficulty (*b*), also referred to as the location parameter; and pseudo guessing (*c*), also referred to as the lower asymptote parameter, that reflects the chances of students with low ability selecting the correct answer. Mathematically, the three-parameter model gives the probability that student with ability θ_k will correctly respond to item *i*:

$$P\left(x_i = 1 \middle| \theta_k, a_i, b_i, c_i\right) = c_i + \left(1 - c_i\right)\frac{\exp\left(-1.7a_i\left(\theta_k - b_i\right)\right)}{1 + \exp\left(-1.7a_i\left(\theta_k - b_i\right)\right)},$$

where x_i is the response to item *i*, with possible alternatives: 1 if correct and 0 if incorrect; θ_k is the ability (in TIMSS referred to as proficiency) of student on a scale *k*. The two-parameter model is equivalent to the three-parameter model with the $c_i = 0$. Setting the c_i parameter to zero implies that guessing the correct answer is highly unlikely. The one-parameter model can be obtained by additionally setting a_i to a constant.

Generalized partial credit model

TIMSS also includes some extended constructed-response items. Each of such items is scored on a multipoint scale with scores ranging from 0 to 2, and are referred to as polytomous items. The polytomous items are scaled using a generalized partial credit model (Muraki 1992) which models the

probability that a person with ability θ_k will have, for the i^{th} item, a response x_i that is scored in the l^{th} of m_i ordered score categories:

$$P\left(x_i = l \middle| \theta_k, a_i, b_i, d_{i,1}, \ldots, d_{i,m_i-1}\right) = \frac{\exp\left(\sum_{v=0}^{l} 1.7a_i\left(\theta_k - b_i + d_{i,v}\right)\right)}{\sum_{g=0}^{m_i-1}\left(\exp\left(\sum_{v=0}^{g} 1.7a_i\left(\theta_k - b_i + d_{i,v}\right)\right)\right)},$$

where m_i is the number of response categories for item *i*, usually 3; $d_{i,l}$ is the category *l* threshold parameter.

Indeterminacies in the parameters of the described model are resolved by setting $d_{i,0} = 0$ and $\sum_{j=1}^{m_i-1} d_{i,j} = 0$.

Mixed-coefficients multinomial logit model

This scaling model is used in PISA, thus we will adopt the notations of Monseur & Adams (2009) and OECD (2012).

Let us assume that we have $i = 1, \ldots, I$ items and $k = 0, \ldots, K_i$ possible response categories for each item. Let $\boldsymbol{X}^T = \left(\boldsymbol{X}_1^T, \ldots \boldsymbol{X}_I^T\right)$ be a response vector, where $\boldsymbol{X}_i = \left(\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{iK_i}\right)^T$ and

$$X_{ij} = \begin{cases} 1, & \text{if response to item } i \text{ is in category } j \\ 0, & \text{otherwise} \end{cases}.$$

The zero category or vector of zeroes is a reference category, needed for model identification. The probability of response in category *j* of item *i* is assumed to have the form

$$\Pr\left(\boldsymbol{X}_{ij} = 1; \mathbf{A}, \mathbf{B}, \xi \middle| \theta\right) = \frac{\exp\left(\boldsymbol{b}_{ij}\theta + \boldsymbol{a}_{ij}'\xi\right)}{\sum_{k=1}^{K_i} \exp\left(\boldsymbol{b}_{ik}\theta + \boldsymbol{a}_{ik}'\xi\right)},$$

where vector $\xi^T = \left(\xi_1, \ldots, \xi_p\right)$ describes items (ξ are used to describe the empirical characteristics of the response categories of each item), $\mathbf{A}^T = \left(\boldsymbol{a}_{11}, \boldsymbol{a}_{12}, \ldots, \boldsymbol{a}_{1K_1}, \boldsymbol{a}_{21}, \ldots, \boldsymbol{a}_{2K_2}, \ldots, \boldsymbol{a}_{IK_I}\right)$ is a design matrix constructed from a set of design vectors $\boldsymbol{a}_{ij}$ $(i = 1, \ldots, I;\ j = 1, \ldots, K_i)$, $\mathbf{B} = \left(\mathbf{B}_1^T, \mathbf{B}_2^T, \ldots, \mathbf{B}_I^T\right)^T$ is a scoring matrix for the entire test, $\mathbf{B}_i = \left(\boldsymbol{b}_{i1}, \boldsymbol{b}_{i2}, \ldots, \boldsymbol{b}_{iD}\right)^T$ is the scoring sub-matrix for item *i*, and $\boldsymbol{b}_{ik} = \left(b_{ik1}, b_{ik2}, \ldots, b_{ikD}\right)^T$ is a vector containing the scores across *D*

dimensions, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_D)'$ represents and individual's position in the D-dimensional latent space. The response vector is

$$f(\boldsymbol{x}; \xi \mid \theta) = \psi(\theta, \xi) \exp(\boldsymbol{x}'(\mathbf{B}\theta + \mathbf{A}\xi)),$$

$$\psi(\theta, \xi) = \left(\sum_{z \in \Omega} \exp\left(\boldsymbol{z}^T \left(\mathbf{B}\theta + \mathbf{A}\xi \right) \right) \right)^{-1},$$

where Ω is the set of all possible response vectors, $\boldsymbol{x}$ is a particular case of the $\boldsymbol{X}$.

Plausible values

To reflect the estimates of population characteristics, the multiple imputation method, also referred to as the plausible values approach (Rubin, 1987; Mislevy, 1991), is mostly used. This means that several scores, called plausible values (PVs), are generated for each student. The PVs approach uses students' responses to the items together with all background data in order to estimate directly the characteristics of student populations and subpopulations (Yamamoto & Kulick, 2000). In such a way, estimates of student performance may be obtained on the assessment as a whole, even though each student responded to just a subset of the assessment items. PVs should not be treated as test scores for individuals, but rather as a measure of performance of the population.

Generating PVs

Let $f(\boldsymbol{x} \mid \theta)$ be the item response probability, mostly referred to as the item response model, given an item response pattern $\boldsymbol{x}$, and ability θ (a measure of student's proficiency). In large-scale assessments, $f(\boldsymbol{x} \mid \theta)$ corresponds mostly to one-, two- or three parameter IRT models. Next, assume that $g(\theta)$, referred to as the population model, has a multivariate normal distribution with a mean given by a linear model with regression parameters, and a common variance. Then, PVs for a student with item response pattern $\boldsymbol{x}$, are random draws from a posterior distribution given by (Wu, 2005)

$$h(\theta, \boldsymbol{x}) = \frac{f(x \mid \theta) g(\theta)}{\int f(x \mid \theta) g(\theta) d\theta}.$$

Commonly, five PVs are generated for each student. The National Assessment of Educational Progress (NAEP) is the only assessment that at the moment generates twenty sets of PVs.

Analyzing data with PVs

Denote generated PVs by $\hat{Q}_m$, $m = 1, \ldots, \mathrm{M}$, where $M > 1$ is the number of PVs drawn. Any analysis containing PVs should be done separately for each

imputed PV data set, and then the obtained values should be combined to a single estimate, taking the form of

$$Q^* = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m \, .$$

The estimated total variance of Q^* is the sum of two components: the within-imputation variance, V^*, and the between-imputation variance, B_M:

$$\mathrm{Var}\left(Q^*\right) = V^* + \left(1+\frac{1}{M}\right)B_M \, ,$$

where $V^* = \frac{1}{M}\sum_m V_m$, V_m is the sampling variance for $\hat{Q}_m$, and $B_M = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Q}_m - Q^*\right)^2$ (Mislevy, 1991; Schafer, 1997).

A more detailed description of TIMSS and PISA scaling methodology can be found in Yamamoto & Kulick (2000) and Monseur & Adams (2009) respectively.

Multilevel models

The international large-scale assessments, as mentioned previously, have a two-stage sampling design. Students are grouped in schools, so the population of students consists of subpopulations of schools that contain students. Ignoring the sampling design would mean that students (secondary units) were selected independently, which is not true. Having selected a school (primary unit) increases the chances of selection of students from that school. Therefore, in such two-stage sampling design, observations are dependent, and ignoring this fact in the statistical analysis may lead to inaccurate inferences (Snijders & Bosker, 2012).

In order to take into account the multistage sampling design, the multilevel statistical models should be used. Multilevel modeling is an extension of the well-known multiple regression method. Multilevel model or also called hierarchical linear model can be seen as a model with simultaneous multiple regressions at different levels that includes nested random coefficients. Multilevel analysis accounts for correlated responses at levels where dependencies of observations occur (Ma, Ma, & Bradley, 2008).

One common way of starting the multilevel modeling is by analyzing a *null model*, also referred to as an *empty model*. Assume that we have *J* schools (groups) with a varying number of students (individuals) n_j in each school. The outcome Y_{ij}, which in our case is the student's achievement, is

measured at the student level, here denoted as Level 1. Then the null model can be expressed as

Level 1 (within schools): $Y_{ij} = \beta_{0j} + r_{ij}$,

Level 2 (between schools): $\beta_{0j} = \gamma_{00} + u_{0j}$,

where β_{0j} is a random intercept, γ_{00} is a general mean, u_{0j} is a random effect at the school level, here denoted as Level 2, and r_{ij} is a random effect at the student level.

The null model can be useful for several research purposes. It is used as the baseline model for comparison of more complicated models. It also allows researchers to estimate variance components at the student and school levels, as well as to estimate the amount of dependence of observations, that can be expressed as the intraclass correlation (ICC). More specifically, the ICC is the proportion of variance at the school level in relation to the total variance, and can be estimated by the equation

$$\rho = \frac{\sigma_{00}}{\sigma_{00} + \sigma_r^2},$$

where σ_{00} is the variance of the residual errors u_{0j}, i.e., the between-school variance, and σ_r^2 is the variance of the student level residuals.

The next step in multilevel modeling, similarly to the multiple regression analysis, is to include explanatory variables. Multilevel models with explanatory variables can have fixed or random slopes. In the random intercept models with fixed slopes, the relationship between the explanatory variables and the dependent variable is the same in every school. However, sometimes the effects of the explanatory variables can vary from one school to another. This can be modeled by multilevel models with random slopes.

Explanatory variables can be introduced at all levels of a model. They can be included at student level only, with the intention of identifying the possible characteristics of students that may be confounded with school effects. Such models are called level-one or student models. If we augment a level-one model with aggregated student level variables, included at school level, then we get a so-called contextual model. Such models aim to examine school contextual effects (Ma et. al, 2008). Finally, multilevel models can include explanatory variables at all levels, and they are often called full models. The full models may also contain interaction variables or nonlinear transformations of basic variables. For example, a two-level full model with one student-level explanatory variable X_{ij}, one school-level explanatory variable Z_j, and an interaction term $Z_j X_{ij}$ can be expressed as follows:

Level 1 (within schools): $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$,

Level 2 (between schools): $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$,

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}.$$

Assumptions of the multilevel models

It is always very important to test the validity of assumptions related to the studied models. Multilevel models have similar assumptions to most of the other general linear models; however, some of the assumptions are adjusted for the hierarchical nature of the design. The major assumptions for multilevel models are (Maas & Hox, 2004):

- Linearity (linear relationship between variables);
- Normality. The student level residuals r_{ij} have a normal distribution with mean zero and constant variance σ_r^2, the school level residuals u_{0j} and u_{1j} have multivariate normal distribution with mean zero and constant variance σ_{00};
- Independence. The pairs of residuals (u_{0j}, u_{1j}) are independent and identically distributed, and they are independent of the student level residuals r_{ij}. The residuals r_{ij} are also independent and identically distributed. Observations at highest level are independent of each other.

It is very important to check the assumptions, as model misspecification can result in the misrepresentation of the relations in the data and invalid hypothesis tests. The various aspects of the model specification are somewhat entangled together, i.e. the model misspecification in one respect may lead to consequences in other respects. For example, an unidentified Level 1 heteroscedasticity may lead to fitting a model with a significant random slope variance, which then disappears if the heteroscedasticity is taken into account; and unidentified non-linear effects of some explanatory variables X_{ij} may appear as heteroscedasticity at Level 1 or as a random slope (Snijders & Berkhof , 2007).

Summary of papers

In this section, short summaries of the four papers are presented.

Paper I: The importance of sampling weights in multilevel modeling of international large-scale assessment data.

The aim of this paper is to examine different approaches and to give recommendations concerning handling design weights in multilevel models when analyzing large-scale assessments such as TIMSS. For this purpose, we examine real data from two countries, Sweden and the USA, and perform a simulation study. Three types of two-level model are used in the empirical and simulation studies. The student model and the full model were of primary interest, while the null model is used for reference purposes. In order to examine the impact of weights in the models mentioned, four different cases are examined: without weights, with unscaled weights, with scaled weights, and with different combinations of weights.

The analyses in the empirical study showed that using no weights or only student-level weights sometimes could lead to misleading conclusions. The simulation study only showed small differences in estimation of the weighted and unweighted models when informative design weights were used. The use of unscaled or not rescaled weights however cause significant differences in some parameter estimates.

Paper II: Using plausible values in secondary analysis in large-scale assessments.

The objective of this paper is to demonstrate the role of plausible values in large-scale assessment surveys when multilevel modeling is used. In order to reach the objective, different user strategies (different numbers) of PVs for multilevel models as well as means and variances are examined using both simulations and real data from TIMSS 2011. Three countries are chosen for the real data analysis based on their average mathematics achievement. The countries chosen represent different parts of the achievement scale, i.e. below the international mathematics average score (Sweden), close to the average score (Slovenia), and above the average score (USA). Simulated data are made up to mimic the TIMSS database as closely as possible. A full multilevel model is used in both, the empirical and the simulation, studies.

Analysis of the real data shows that biased results are obtained if PVs are used inappropriately in the analysis. When using only one or a few PVs, parameter estimates vary greatly and the quality of estimation varies greatly depending on which PV is chosen. Our study also shows that the estimation

results in multilevel modeling using the average of PVs are very close to those, obtained using all five PVs, when analyzing TIMSS 2011 data, but as expected, the standard errors and the within-school variance differ. The results of the simulation study indicate that PV-based estimates have a better recovery of the population parameters than any of the point estimators, although in general the differences between all estimates are quite small. From the simulation study, we can also conclude that it is possible for us to increase the precision of the estimates in some cases if more than five PVs are used.

Paper III: Single imputation from a conditional distribution vs multiple imputation for data with a non-monotone missing pattern.

Missing information is common in real data studies. When missingness is large, it should not be ignored and, instead a missing data imputation method should be considered. The choice of the imputation method depends on the type or pattern of missing information, as well as on the nature of data. For instance, observations in large-scale educational assessments are incomplete by missing some components and based on usually positively correlated results within the students. In all types of analysis of such data, the correlation has to be considered in a reliable calculation of properties of estimates. The aim of this paper is to compare a single imputation from a conditional distribution (with or without weights) and multiple imputation for data with a non-monotone missing pattern and high positive correlation between variables. For this purpose, such estimates as mean and variance are compared.

The results of the simulation study show that for the complete-data set, imputation from a conditional distribution with unweighted estimates (method I) and with the weighted estimates (method II) estimate the average mean and variance better than multiple imputation (method III). In most of the cases, method I with a slight difference tends to outperform method II. It is worth noting that method III often results in smaller variations of estimates compared to other methods. The analogous conclusions are obtained even for higher correlations between the variables. If the three variables are weekly correlated, or two out of the three variables are weekly correlated, then method II estimates the average mean and variance better than other methods studied.

In conclusion, imputation from a conditional distribution with weighted and unweighted estimates, as well as complete-data set, estimate variances in the studied conditions more reliably than does multiple imputation.

Paper IV: Low-, medium-, and high-performing schools in the Nordic countries. Student performance at PISA mathematics 2003-2012.

Decreasing performance among students in Sweden on international educational large-scale assessments and increasing segregation of schools, has led to a number of discussions concerning strategies for improving student performance. The previous research has shown that changes in student performance differ depending on the performance level of students. For example, the largest decreases in student performance are seen for low- and medium-performing students in Sweden. This raises the question of whether different performance levels may be related to different kinds of school factors. Hence, the school unit and its characteristics regarding the composition of students with reference to performance level are of great importance. The purpose of this study is to analyze the between school variance and to identify factors associated with student performance in mathematics in PISA at different school performance levels in the Nordic countries. In order to separate the effect of school-level variables, from the effect of student's background factors and to take the multistage sampling design used in PISA into account, multilevel analysis is used. Contrary to previous studies conducted on science performance in PISA, the results of our study show that no evidence regarding the relationship between the average student performance in mathematics and the between-school variance is found. Regarding school-level factors, our results overall have shown that few school-level factors (having a positive or a negative effect) seem to be associated with performance. School-level factors associated with performance have mainly been identified only for low- and medium-performing schools, and to a less extent for high-performing schools (only in Sweden and Denmark). This is a result which is in line with other studies showing the educational system's incapacity to provide support for high-performing students and to enhance their learning.

Final remarks and further research

This thesis contributes to the area of research based on data from large-scale educational assessments, with a focus on the application of multilevel models. The role of sampling weights, plausible values (multiply imputed response variable), and imputation methods is shown by simulations and applications to TIMSS and PISA data.

International large-scale assessments have a large amount of missing data in both items and background information, with planned missing data at item level and non-planned missing data for background information. Hence, the

researcher using multilevel models for analysis of such data must deal with multiply imputed response variable, and missing data at all levels of the model. This leads to quite complicated models that are hard to implement in existing software devoted to multilevel modeling. In the future, it would be of interest to compare various methods of handling all this missing data and to analyze their effects on the results.

In Paper II, we analyzed plausible values that were generated from single-level imputation models, as it is done in the large-scale assessments. However, the data and student proficiencies are of a hierarchical nature, and so it would be of great interest to perform a similar simulation study with plausible values generated using a multilevel latent variable plausible values approach.

In Paper III, only two different imputation methods with a few variables were compared. Further research should focus on comparing more methods with more complicated data setups. It would also be of great interest to study a case of discrete data with a large amount of missing information for items and background information, with all the complexity possessed by large-scale assessment data.

In 2017, data from TIMSS 2015 and PISA 2015 will be released for public use. This is only the second time in history when both assessments have been conducted in the same year. Hence, it would be of great interest to perform comparative studies with these data in the future.

References

Adams, R.J. & Wu, M.L. (2007). The Mixed-Coefficients Multinomial Logit Model: A Generalized Form of the Rasch Model. In M. von Davier & C.H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications*. (pp.57–75). New York: Springer-Verlag.

Durrant, G.B. (2009). Imputation Methods for Handling Item-Nonresponse in Practice: Methodological Issues and Recent Debates. *International Journal of Social Research Methods*, 12(4), 293–304.

Harris, D. (1989). Comparison of 1-, 2-, and 3-Parameter IRT Models. *Educational Measurement: Issues and Practice*, 8(1), 35–41.

Horton, N.J. & Kleinman, K.P. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1), 79–90.

Jaeger, R.M. (1984). *Sampling in Education and the Social Sciences*. New York: Longman.

Joncas, M. & Foy, P. (2012). Sample Design in TIMSS and PIRLS. In M.O. Martin, I.V.S. Mullis (Eds.). *Methods and procedures in TIMSS and*

PIRLS 2011 (pp.1–21). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Ma, X., Ma, L., & Bradley, K.D. (2008). Using multilevel modeling to investigate school effects. In A.A. O'Connell, D.B. McCoach (Eds.), *Multilevel modeling of educational data* (pp.59–110). Charlotte, NC: Information age publishing inc.

Maas, C.J.M., & Hox, J.J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational statistics and data analysis*, 46, 427–440.

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

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

Monseur, C. & Adamas, R. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement*, 10(3), 320–334.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176.

Snijders, T.A.B. & Bosker, R.J. (2012). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*, 2nd edition. London: Sage.

OECD (2012). Sample design. *PISA 2009 Technical Report*, PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en

OECD (2012a). Survey Weighting and the calculation of sampling variance. *PISA 2009 Technical Report*, PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chopman & Hall.

Snijders, T.A.B, & Berkhof, J. (2007). Diagnostic checks for multilevel models. Handbook of Multilevel Analysis. In J. de Leeuw & M. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 139–173). New York, NY: Springer.

von Davier, M., Gonzalez, E. & Mislevy, R.J. (2009). What are plausible values and why are they useful? *IERI Monograph Series. Issues and Methodologies in Large–Scale Assessments*, 2, 9–36. Retrieved from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf

Wilson, M.D. & Lueck, K. (2014). Working with missing data: imputation of nonresponse items in categorical survey data with a non-monotone missing pattern. *Journal of Applied Mathematics*, 368791, 1–9.

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

Yamamoto, K. & Kulick, E. (2000). Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales. In M.O. Martin, K.D. Gregory & S.E. Stemler (Eds.), *TIMSS 1999 Technical Report.* (pp. 237–263). Retreived from http://timssandpirls.bc.edu/timss1999i/tech_report.html. Accessed 11 November 2016.