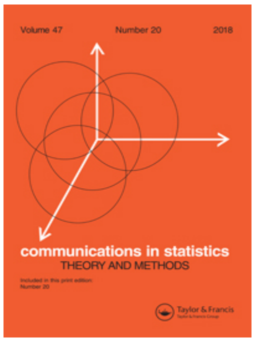

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Importance of sampling weights in multilevel modeling of international large-scale assessment data

Inga Laukaityte & Marie Wiberg

To cite this article: Inga Laukaityte & Marie Wiberg (2018) Importance of sampling weights in multilevel modeling of international large-scale assessment data, Communications in Statistics - Theory and Methods, 47:20, 4991-5012, DOI: 10.1080/03610926.2017.1383429

To link to this article: https://doi.org/10.1080/03610926.2017.1383429

Published online: 13 Nov 2017.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2018, VOL. 47, NO. 20, 4991–5012
https://doi.org/10.1080/03610926.2017.1383429

Taylor & Francis
Taylor & Francis Group

Importance of sampling weights in multilevel modeling of international large-scale assessment data

Inga Laukaityte and Marie Wiberg

Department of Statistics, Umeå School of Business and Economics, Umeå University, Umeå, Sweden

ABSTRACT

Multilevel modeling is an important tool for analyzing large-scale assessment data. However, the standard multilevel modeling will typically give biased results for such complex survey data. This bias can be eliminated by introducing design weights which must be used carefully as they can affect the results. 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 (The Trends in International Mathematics and Science Study). To achieve the goal of the paper, we examined real data from two countries and included a simulation study. The analyses in the empirical study showed that using no weights or only level 1 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 caused significant differences in some parameter estimates.

ARTICLE HISTORY
Received 25 March 2017
Accepted 19 September 2017

KEYWORDS
informative weights; two-stage sampling; rescaling weights; simulation study.

MATHEMATICS SUBJECT CLASSIFICATION

Introduction

The use of the results from international large-scale assessment databases such as TIMSS (Trends in International Mathematics and Science Study) and PISA (Programme for International Student Assessment) for research of educational achievement has increased markedly in recent years. In these large-scale assessments, subjects such as schools, classrooms, or students are selected randomly from a target population. All such complex surveys employ multistage sampling designs with the units at some or all stages selected with unequal probabilities. For example, each school, classroom, and student has a certain probability to be selected. Complex survey data are frequently analyzed with multilevel models, which are used to study the effect of group-level variables on the individual outcomes. However, sampling features of the survey data are commonly ignored in the analysis. When units at any level of the hierarchy are selected with unequal probabilities in ways that are not accounted for by the model, biased parameter estimates are typically obtained if standard multilevel modeling is used (Pfeffermann et al. 1998; Zaccarin and Donati 2008). One way to eliminate this bias is to use sampling weights at one or all levels in the multilevel model, but these weights must be included properly in the model because they can affect the results of the analysis.

CONTACT Inga Laukaityte inga.laukaityte@gmail.com Department of Statistics, Umeå School of Business and Economics, Umeå University, Umeå, Sweden.

Chantala and Suchindran (2006) have analyzed how to construct the sampling weights for multilevel models for different software. They used a two-level model with continuous outcome, and analyzed a case of using a subpopulation. Jenkins (2008) studied three-level longitudinal multilevel model with informative weights on student and school levels. He also compared results using different software programs. Cai (2013) has investigated ways of handling sampling weights in multilevel model analyses by simulating a home/context model. He compared different approaches of inclusion of sampling weights (multilevel pseudo-maximum likelihood, probability-weighted iterative generalized least squares and naive (ignoring sampling weights)) under different conditions, like informative and non informative design, different levels of variation of sampling weights. Rutkowski et al. (2010) have also advised researches how weights given by TIMSS and PIRLS (Progress in International Reading Literacy Study) databases should be used in single- and multilevel models. However, they have not performed any real data analysis or any simulation study to show how the use or misuse of weights affects the results and the standard errors. Stapleton (2013) overviewed different definitions of weights included in international large-scale assessments and presented simple examples of single- and multilevel analysis. She provided information on the usage of sampling weights in different statistical software as well. Kim, Anderson, and Keller (2013) comprehensively demonstrated how multilevel analysis could be applied for large-scale assessment data using PIRLS 2006 data. Our paper differs from earlier mentioned papers as it contains three types of two-level models (null, student, and full) for both real and simulated data. Furthermore, we studied different cases of informative weights and presented a simulation study that contains models with different sampling designs including various complexities in order to get a better understanding.

Many researchers have used multilevel models for studying students' achievement. Some researchers have not used (or have not mentioned) weights at all (e.g., Jianjun 2000; Lamb & Fullarton 2002; Kyriakides and Charalambous 2004; Sabah and Hammouri 2010), or if they have used sampling weights they do not specify which weights have been used or how (e.g., Howie and Plomp 2004; Webster and Fisher 2010). This raises the questions of whether omitting the weights affects the results of analyses, and whether the right or wrong choice of weights has a statistically significant impact on the modeling results.

The main purpose of this paper is to examine the use of sampling weights in multilevel modeling when analyzing data from TIMSS, and to give recommendations concerning the handling of sampling weights when analyzing large-scale assessment studies such as TIMSS, PIRLS, and PISA. This paper is structured as follows. Multilevel models are introduced in the next section, followed by description of sampling weights. The Intraclass Correlation and Design Effects section contains brief definitions of intraclass correlation and design effects, followed by a description of informativeness of weights. The design of the empirical and simulation studies is given in the Design of the Empirical and Simulation Studies section. The Results section contains the results, and conclusions are given in the last section.

Multilevel models

Three types of two-level models were used in the later empirical and simulation studies; the student model and the full model were of primary interest, while the null model was used for reference purposes. The response variable in all the models was mathematics achievement – the five plausible values for mathematics given in TIMSS 2011 database. Each analyzed model must be estimated for every plausible value separately, and the results must be combined in a special way (Schafer 1997).

Null model. The within- and between-school variances were determined in the null model. Mathematics achievement for each student was estimated as a function of the school average with a random error.

Level 1 (within schools):

$$Y_{ij} = \beta_{0j} + r_{ij}, \;\; i = 1, \ldots, N$$

Level 2 (between schools):

$$\beta_{0j} = \gamma_{00} + u_{0j}, \;\; j = 1, \ldots, J$$

where Y_{ij} denotes mathematics achievement for student i within school j, β_{0j} is the average mathematics achievement for school j, and r_{ij} is the error term representing a unique effect associated with student i in school j. Level 2 term γ_{00} denotes the grand mean of mathematics achievement and u_{0j} is the error term representing a unique effect associated with school j.

Student model. In this model, the mathematics achievement for each student was estimated as a function of the school mean achievement and the student effect. The f student-level factors are denoted as $H_{1ij}, \ldots, H_{fij}$.

Level 1 (within schools):

$$Y_{ij} = \beta_{0j} + \beta_{1j} \cdot (H_{1ij}) + \beta_{2j} \cdot (H_{2ij}) + \cdots + \beta_{fj} \cdot (H_{fij}) + r_{ij}, \; i = 1, \ldots, N$$

Level 2 (between schools):

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j}, \quad j = 1, \ldots, J \\ \beta_{1j} &= \gamma_{10}, \;\; \beta_{2j} = \gamma_{20}, \ldots, \;\; \beta_{fj} = \gamma_{f0} \end{aligned}$$

Full model. In the full model, the mathematics achievement for each student was estimated as a function of the school factors controlling for student-level variables. On the student-level, we included student-related factors, and on the school-level we included all school-level factors, here denoted as $S_{1j}, \ldots, S_{lj}$.

Level 1 (within schools):

$$Y_{ij} = \beta_{0j} + \beta_{1j} \cdot (H_{1ij}) + \beta_{2j} \cdot (H_{2ij}) + \cdots + \beta_{5j} \cdot (H_{fij}) + r_{ij}, \quad i = 1, \ldots, N$$

Level 2 (between schools):

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \cdot (S_{1j}) + \cdots + \gamma_{0l} \cdot (S_{lj}) + u_{0j}, \quad j = 1, \ldots, J$$

$$\beta_{1j} = \gamma_{10}, \;\; \beta_{2j} = \gamma_{20}, \ldots, \;\; \beta_{5j} = \gamma_{50}$$

Sampling weights

Let p_j be the probability of selecting group (school) j, and $p_{i|j}$ is the conditional probability that individual (student) i is selected, given that group (school) j is selected. Most commonly, the sampling (or design) weights w at level 1 (the individual level) are obtained by $w_{i|j} = 1/p_{i|j}$ and at level 2 (the group level) by $w_j = 1/p_j$. The sampling weights used for most surveys are composed of weighting factors, which are the inverse of the probability of selection for each stage of the sampling design (school, class, and student), and adjustment factors for non participation at each stage (Joncas 2008).

An important issue is the scaling of the weights that is required for proper inclusion of the sampling weights into the estimation. It must be noted that scaling is applied only for level 1 weights because scaling of the level 2 weights generally does not influence the parameter estimates and their standard errors (Asparouhov 2006). There are many ways of scaling weights, but there are no agreements on which is the best (Pfefferman et al. 1998; Zaccarin and Donati 2008; Asparouhov 2006; Carle 2009). The sum of the scaled weights usually results in some meaningful characteristics of the sample or some certain value common for each subgroup (the sum of the scaled weights is 500 for each country in TIMSS). In multilevel models, however, the weights must be scaled differently across clusters so that the sum of the weights is equal to some cluster characteristics. Asparouhov (2006) states that in multilevel models the ratio between the true cluster sample size and the total weight within the cluster is an important quantity because it affects the distribution of the level 2 random effects conditional on all observed data. The different scaling methods, however, may have different effects on different estimation techniques. Simple raw weights cannot be used in calculations (Carle 2009). The two most common scaling methods for level 1 weights (Rabe-Hesketh and Skrondal 2006; Pfefferman et al. 1998) are used in this paper and described here:

- Method 1. The factor $\tilde{w}_j = \sum_i w_{i|j}^2 / \sum_i w_{i|j}$ can be interpreted as the "design effect" required to reduce the naïve sample size $\hat{N}_j$ in the unscaled estimator to the *effective sample size*

$$w_{i|j}^{*} = \frac{w_{i|j}}{\tilde{w}_j} = \frac{w_{i|j} \sum_i w_{i|j}}{\sum_i w_{i|j}^2}.$$

- Method 2. The factor $\bar{w}_j = \sum_i w_{i|j}/n_j$ reduces the naïve sample size $\hat{N}_j$ to the *actual sample size* n_j

$$w_{i|j}^{**} = \frac{w_{i|j}}{\bar{w}_j} = \frac{w_{i|j} n_j}{\sum_i w_{i|j}}.$$

All of the international large-scale assessment databases contain ready-to-use scaled weights and their components. However, these weights are prepared for a single-level analysis and are not appropriate for multilevel analysis.

Sampling weights in TIMSS

One international large-scale assessment is TIMSS, which is an international comparative assessment dedicated to improve teaching and learning in mathematics and science for students around the world (Olson, Martin, and Mullis 2008). A two-stage stratified cluster sample design is used in TIMSS, with schools at the first stage and intact classes (with all students in the class) at the second stage (Olson et al. 2008). The two-stage probability proportional to size sampling technique is used as the student sampling selection method. Thus, each student in the target population is chosen with unequal probability. To compensate for the unequal probabilities of selection and to avoid possible bias in the parameter estimates, sampling weights should be used.

Sampling weights in the TIMSS study are calculated according to a three-step procedure involving selection probabilities for schools (the first stage), classrooms (the second stage), and students (the last, third stage) (Joncas 2007). The overall student sampling weight is the product of the three weights and includes non participation adjustments. TIMSS offers six sets of weights to be downloadable with the data:

(a) Total student weight (*totwgt*) – sums to each national population,
(b) Student house weight (*houwgt*) – sums to the student sample size in each country,

(c) Student senate weight (*senwgt*) – sums to 500 in each country,
(d) Overall and by subject teacher weights (*tchwgt, matwgt*, and *sciwgt*) – *totwgt* divided by the number of teachers (overall, math, or science) a student has.
(e) School weight (*schwgt*) – sums to the number of schools in a given country,
(f) Sum of student weights (*stotwgt*) – the sum of the student weights within a school. This is equivalent to the number of students that are represented by the students in the school.

For each student, by grade within each country, where i is the individual student and g is the grade of the student, the weights *senwgt* and *houwgt* are obtained by

$$senwgt_{g,i} = totwgt_{g,i}\frac{500}{\sum_{i=1}^{I} totwgt_{g,i}}$$

and

$$houwgt_{g,i} = totwgt_{g,i}\frac{n}{\sum_{i=1}^{I} totwgt_{g,i}}$$

where I is the total number of students in the country for the grade g. The total student weight is composed from weighting factors $wgtfac1$, $wgtfac2$, and $wgtfac3$, which are the inverse probabilities of selection of the school, a classroom within the school, and the individual student within the classroom, respectively. The weighting adjustments $wgtadj1$, $wgtadj2$, and $wgtadj3$ are applied to the weighting factors to account for non participating schools, classes, and students, respectively. The total student weight is thus defined as

$$totwgt_{j,i} = wgtfac1_{j,i} \cdot wgtfac2_{j,i} \cdot wgtfac3_{j,i} \cdot wgtadj1_{j,i} \cdot wgtadj2_{j,i} \cdot wgtadj3_{j,i}.$$

A more detailed description of the weighting factors and adjustments can be found in Joncas (2008). The school weight *schwgt* is obtained by multiplying the variables $wgtadj1$ and $wgtfac1$.

All of the weights (a)–(f) mentioned above are appropriate weights for single-level analysis, but must be very carefully used in the multilevel case. For example, Rutkowski et al. (2010) advise against the use of the provided *schwgt* weights at level 2 and *totwgt* at level 1. The total student weight *totwgt* is the joint probability that the school, the class, and the student are selected. However, in two-level models an assumption is made that the level 1 student weight (when level 2 weights are included) is inversely proportional only to the probability of a student being selected given that the school was selected. Thus, it is recommended to calculate weights manually for each level using separate components of weights that are included in the TIMSS database. It is worth noting that the weights (a)–(d) should also be rescaled when used in the multilevel models, as was mentioned in the introduction. All level 1 weights are calculated to compute population means which are different from two-level means when the groups (schools) are not balanced. It is recommended to rescale them in the software of one's choice.

Sampling weights in multilevel models in TIMSS

In order to use the proper sampling weights in multilevel modeling, the level 1 weights supplied with data of large-scale assessments should be recalculated. The right weights here denoted as *studwgt*, and can be obtained by multiplying the variables *wgtadj2, wgtfac2, wgtadj3*, and *wgtfac3*. The level 2 weights *schwgt* do not require recalculation. Again, it is important to emphasize that scaled level 1 weights should be used in the analysis, but scaling of level 2 weights is not obligatory and should not influence parameter estimates. To examine the impact of sampling weights, four different cases were examined in the later simulations and real data studies using the null, student, and full two-level models:

(i) without weights,
(ii) with unscaled weights,
(iii) with scaled weights (method 1 and 2),
(iv) with different combinations of weights.

Intraclass correlation and design effects

Intraclass correlation (ICC) shows how homogeneous the data are within group-level clusters. More specifically, the ICC indicates the proportion of total variance in the data that is accounted for by the group-level variance. ICC is typically useful to run with the unconditional model before evaluating and comparing different multilevel models. The ICC can be computed at any stage of modeling by ICC $= \omega^2/(\omega^2 + \sigma^2)$, where ω^2 is the between-school variance, and σ^2 is the within-school variance. Moreover, the ICC is used for the calculation of the design effect, which shows how much the standard errors are underestimated. The design effect can be obtained from

$$\text{Designeffect} = 1 + (\text{averageclustersize} - 1) \cdot \text{ICC}$$

A design effect greater than 2 indicates that the clustering of the data needs to be taken into account during estimation (Kish 1965).

Informativeness

The informativeness of weights (see, e.g., Pfeffermann 1993) is a useful measure. If weights are informative, then their influence on results is not negligible and they should be used in the multilevel analysis of the data. Non informative weights should not affect the results noticeably. The weights are informative when effective sample size is smaller than the real sample size. Effective sample sizes for two-level models can be obtained as follows. For effective sample size at level 2 (between schools)

$$N^{\text{eff}} = \frac{\left(\sum_j w_j\right)^2}{\sum_j \left(w_j^2\right)}$$

and for effective sample size at level 1 (within schools) for cluster j

$$n_j^{\text{eff}} = \frac{\left(\sum_i w_{i|j}\right)^2}{\sum_i \left(w_{i|j}^2\right)}$$

In general, the effective sample size is defined in a way so that a weighted sample gives the same amount of information as a simple random sample with sample size equal to the effective sample size (Pfeffermann 1993; Snijders and Bosker 2012).

Design of the empirical and simulation studies

Large-scale assessments are extremely complex studies, so to keep control over as many conditions as possible a simulation study was included. In the empirical study, all problems relating to the use of real data, for example, missing data, had to be taken into account. In the simulated study, we simplified the problem by simulating constructs, not separate item responses, as described more explicitly in the Simulation Study section below.

Student-level factors and school-level factors

The response variable used was mathematics achievement described by five plausible values, and the weights were taken from the TIMSS 2011 database. To model mathematics achievement, we used a number of level 1 factors and level 2 factors presented in Table 1. These definitions were generally in line with Mohammadpour and Ghafar's (2012) operationalization of some important concepts for mathematics achievement. Note, not all of their defined factors were used, because our interest here was in the sample weights and not explicitly about which factors explain mathematics achievement. In addition, some definitions are somewhat different from Mohammadpour and Ghafar (2012), because they used data from IEA (2007), and some items are changed or were not present in TIMSS 2011. *Socioeconomic status* is the background factor that has been altered the most. The characteristics of the factors are given in detail in the Appendix Table A1.

Software

Researchers use different software for multilevel modeling and one must be aware of how the software behaves with weights. For example, weights scaled outside of the program cannot be used in the programs HLM or SPSS because these software programs perform automatic scaling of the weights (Carle 2009). Thus, if one uses already scaled weights (appropriate for multilevel modeling), they would not be properly included in the likelihood estimation. In Mplus, a researcher can choose not to scale weights, or to scale them by using method 1 (referred to as Ecluster) or method 2 (referred to as Cluster) described in the previous Sampling Weights section. In this paper, the IEA IDB analyzer (IDB Analyzer 2009) was used for merging data files, and the multilevel analysis was implemented with Mplus 7 (Muthén and Muthén 1998–2011), as it is designed to handle both plausible values and sample weights. The simulations were performed in R software (R Development Core Team 2014).

Empirical study

Real data from TIMSS 2011 grade 8 in Mathematics (IEA 2011) in Sweden and the USA were used. Sweden can be viewed as country of an average size with respect to the number of schools. The USA was chosen as the country having one of the largest sampling size with the intention that sampling weights should play a greater role in this case, especially when analyzed together with other countries. Level 1 and level 2 factors were constructed as described previously. Descriptive statistics of these factors can be found in Table A1 in the Appendix. The factors were then modeled with the three different multilevel models and examined in the four cases of interest for the two countries.

The missing data range at the level 1 was low, ranging from a minimum of 1.28% for *ses* in the USA to a maximum of 5.08% for *atm* in Sweden. Because missing data at the level 1 was low, slightly more than 5%, listwise deletion was used for handling missing data (Tabachnik and Fidell 2007). It is, of course, better to use multiple imputations, but because the exact models were of less importance here and there were only a few cases, listwise deletion was

Table 1. Level 1 and level 2 characteristics of the investigated factors.

	Description
Factor	Student-level factors
Math self-concept [*msc*]	This factor measures students' mathematics self-concept through the items: (1) I usually do well in mathematics, (2) mathematics is not one of my strengths, (3) mathematics is more difficult for me, and (4) I learn things quickly in mathematics. Possible responses were 1 – Agree a lot, 2 – Agree a little, 3 – Disagree a little, 4 – Disagree a lot. Note that (2) and (3) were reverse coded. The responses were averaged and classified into two categories: 0 = Low, average is greater than 2.5; 1 = High, average is less than or equal to 2.5.
Attitude toward mathematics [*atm*]	This factor was based on students' responses to: (1) I enjoy learning mathematics, (2) mathematics is boring, and (3) I like mathematics. Note (2) was reverse coded. The possible responses and the constructed scale were the same as those of Math self-concept.
Valuing mathematics [*vm*]	This factor was based on students' responses to: (1) Learning mathematics will help me in my daily life, (2) I need mathematics to learn other school subjects, (3) I need to do well in mathematics to get into university, and (4) I need to do well in mathematics to get the job I want. The possible responses and the constructed scale were the same as those of Math self-concept.
Socioeconomic status [*ses*]	This was based on students' responses on the following two indicators. (1) Books at home: 1 = 0–10 books, 2 = 11–25 books, 3 = 26–100 books, 4 = 101–200 books, and 5 = >200 books. This was recoded to 1 = Low (0–25 books), 2 = Medium (26–200 books), or 3 = High (>200 books). (2) Possession of educational home resources (computer and study desk). This was categorized as 0 = Low, if the student had none or one item; 1 = High, if the student had two items. The two indicators were averaged and classified into 0 = Low, average was less than 2; 1 = High, average was equal to 2 or greater.
Sex [*sex*]	0 = Female, 1 = Male.
	School-level factors
School climate [*sclim*]	This factor was the school principals' answers to how they would characterize the following statements within their school: (1) teachers' job satisfaction, (2) teachers' understanding of the school's curricular goals, (3) teachers' degree of success in implementing the school's curriculum, (4) teachers' expectations for student achievement, (5) parental support for student achievement, (6) parental involvement in school activities, (7) students' regard for school property, and (8) students' desire to do well in school. Possible responses were 1 = Very high, 2 = High, 3 = Medium, 4 = Low, 5 = Very low. The responses were averaged and categorized as 0 = Low, average was greater than 2.5; 1 = High, average was less than or equal to 2.5.
Good attendance [*ga*]	The school principals' answers to what degree is arriving late at school/absenteeism a problem among 8th grade students in their school. Possible responses were: 1 – Not a problem, 2 – Minor problem, 3 – Moderate problem, and 4 – Serious problem. The responses were summed and classified into two categories: 0 = Low, sum is greater than or equal to 5; 1 = High, sum is less than 5.
School location [*sloc*]	The school principals' answers to the number of people living in the area where the school is located. The responses were classified into two categories: 0 = Rural areas, less than or equal to 50,000 people; 1 = Urban areas, greater than 50,000 people.
School resources [*sre*]	The school principals' answers to whether their school's capacity to provide instruction is affected by a shortage or inadequacy of any of the following items: (1) computers for mathematics instruction, (2) computer software for mathematics instruction, (3) calculators for mathematics instruction, (4) library materials for mathematics instruction, or (5) audio-visual resources for mathematics instruction. The original responses were 1 = Not at all; 2 = A little; 3 = Some; or 4 = A lot. These responses were averaged and classified into two categories: 0 = Low, average is greater than 2; 1 = High, average is less or equal to 2.

chosen to simplify this problem. Missing values at higher levels, that is, at the level 2, cannot simply be removed because this will have an impact on the lower level as well. Missing data were assumed to be missing at random (MAR), which means that a participant's probabilities of response are related only to his or her own set of observed items, a set that can change from one participant to another (Schafer and Graham 2002). The two most popular ways of dealing with MAR data are the multiple imputation method (Rubin 1987; Howell 2008) and the full-information maximum likelihood (FIML) procedure (Marsh et al. 2005; Pauli, Reusser, and Grob 2007; Danielsen et al. 2010). Missing data at higher levels have a complicated structure that is difficult to mimic. For this reason, Schafer and Graham (2002) recommend using FIML for handling missing data instead of imputation, which was used in this study.

Empirical study procedure

In the empirical study, we started by analyzing the informativeness of the weights. Next, multilevel analysis of the data was performed by starting with the weighted and unweighted null models. Then, ICC was calculated to determine the homogeneity of the data within groups, that is, the existence of school effects was examined. Next, level 1 factors were introduced, and weighted as well as unweighted student models were constructed. Finally, the full multilevel models containing different factors at level 1 and level 2 were examined. For all models, parameter estimates, standard errors, Akaike information criterion (AIC), Bayesian information criterion (BIC), and deviance were examined.

Simulation study

Initially, we performed a simple simulation study based on Pfefferman et al.'s (1998) paper. The simulation study consists of two cases (A and B) for simulating the finite population and a more complex case C, which aimed to be as similar as possible to the TIMSS 2011 study. Case A is equivalent to their study and is displayed to illustrate the null model case. Case B was constructed to illustrate the full model when one level 1 factor and one level 2 factor were included, and is not studied by Pfefferman et al. (1998). This model helps to understand how the estimation of model parameters changes when different factors are included. Finally, case C, to the best of our knowledge, has not appeared in the literature before.

(A) The finite population values Y_{ij} were simulated from the two-level null model $Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$, $u_{0j} \sim N(0, \omega^2)$, $r_{ij} \sim N(0, \sigma^2)$, $j = 1, \dots, 300$, $i = 1, \dots, N_j$; $\gamma_{00} = 1$, $\omega^2 = 0.2$, $\sigma^2 = 0.5$. Values of N_j were obtained from $N_j = 75 \exp(\tilde{u}_j)$, $\tilde{u}_j \sim N(0, \omega^2)$, and $-1.5\omega \leq \tilde{u}_j \leq 1.5\omega$.

(B) The finite population values Y_{ij} were simulated from the full two-level model $Y_{ij} = \gamma_{00} + \gamma_{01}S_j + u_{0j} + \gamma_{02}H_{ij} + r_{ij}$, $S_j \sim \text{Bernoulli}(0.4)$, $H_{ij} \sim \text{Bernoulli}(0.6)$, $u_{0j} \sim N(0, \omega^2)$, $r_{ij} \sim N(0, \sigma^2)$, $j = 1, \dots, 300$, $i = 1, \dots, N_j$; $\gamma_{00} = 1$, $\omega^2 = 0.2$, $\sigma^2 = 0.5$. Values of N_j were obtained from $N_j = 75 \exp(\tilde{u}_j)$, $\tilde{u}_j \sim N(0, \omega^2)$, and $-1.5\omega \leq \tilde{u}_j \leq 1.5\omega$.

The following two sampling schemes were used for cases A and B.

(1) A sample in the A and B cases was obtained such that weights would be *informative only at level 2*. In total, 35 level 2 units were sampled with probabilities π_j proportional to a measure of size (e.g., school size) X_j, where $\pi_j = 35X_j / \sum_1^{300} X_j$. The measure X_j is obtained in the same way as N_j, although u_{0j} is used instead of $\tilde{u}_j$. Then, a simple random sample of 38 level 1 units was selected from each sampled level 2 unit.

(2) A sample in the A and B cases was obtained such that weights would be *informative at both levels*. Here the scheme is the same as in (1), except the sampling of the level 1 units. The level 1 units in each sampled level 2 unit were divided into two strata according to whether $r_{ij} > 0$ or $r_{ij} \leq 0$. Then, simple random samples of sizes 10 and 29 were selected from the respective strata.

(3) (C) The mathematics achievement values for student i within school j, Y_{ij}, were simulated from the full two-level model with level 1 and level 2 factors, which mimic the real data as described in Table 1.

$Y_{ij} = \gamma_{00} + \gamma_{01}S_{1j} + \gamma_{02}S_{2j} + \gamma_{03}S_{3j} + \gamma_{04}S_{4j} + u_{0j} + \gamma_{05}H_{1ij} + \gamma_{06}H_{2ij} + \cdots + \gamma_{09}H_{5ij} + r_{ij}$, $S_{1j} \sim \text{Bernoulli}(0.51)$, $S_{2j} \sim \text{Bernoulli}(0.64)$, $S_{3j} \sim \text{Bernoulli}(0.40)$, $S_{4j} \sim \text{Bernoulli}(0.83)$, $H_{1ij} \sim \text{Bernoulli}(0.70)$, $H_{2ij} \sim \text{Bernoulli}(0.52)$, $H_{3ij} \sim \text{Bernoulli}(0.92)$, $H_{4ij} \sim \text{Bernoulli}(0.69)$, $H_{5ij} \sim \text{Bernoulli}(0.52)$, $u_{0j} \sim N(0, \omega^2)$, $r_{ij} \sim N(0, \sigma^2)$, $\gamma_{00} = 1$, $\omega^2 = 0.2$, $\sigma^2 = 0.5$. The finite population consisted of 1519 schools. The simulated number of classes varied from two to six classes per school. Values for number of students per class k were obtained from $N_k = 20\exp(\tilde{u}_k)$, $\tilde{u}_k \sim N(0, \omega^2)$, and $-0.5\omega \le \tilde{u}_k \le 0.5\omega$.

A sample was obtained so that weights would be informative only at level 2. In total, 153 schools were sampled with probabilities π_j proportional to a measure of size X_j, where $\pi_j = 153X_j / \sum_1^{1519} X_j$. The measure X_j is obtained in the same way as N_k, although u_{0j} is used instead of $\tilde{u}_k$. Then, a simple random sample of classes was selected, randomly picking two classes from each sampled school (in the cases A and B we randomly selected students). All students of the sampled class were included in the sample. The average effective sample size at level 2, N^{eff}, for the simulated samples was 94 (compared with the actual sample size of 153). The process of simulating the finite population and selecting a sample was repeated 500 times for all cases. For each obtained sample, multilevel models were applied.

Sampling and calculation of the weights were done using the R package *Sampling*, version 2.6 (Tillé and Matei 2013). Note, that Pfefferman et al. (1998) used probability-weighted iterative generalized least squares (PWIGLS) method for the weighted analysis, while we used multilevel pseudo-maximum likelihood (MPML) estimation method implemented in Mplus 7. The MPML method directly estimates the population likelihood function by weighting the sample likelihood function (Asparouhov 2004, 2006)

$$L(\theta_1, \theta_2) = \prod_j \left(\int \left(\prod_i f\left(y_{ij}, H_{ij}, \eta_j, \theta_1\right)^{w_{i|j}\lambda_{1j}} \right) \phi\left(\eta_j, S_j, \theta_2\right) d\eta_j \right)^{w_j \lambda_{2j}}$$

where θ_1 and θ_2 are parameters for the fixed effects for the level 1 and level 2, H_{ij} is level 1 covariates (factors) for individual i in group j, y_{ij} is the observed variable (in our case mathematics achievement), η_j is level 2 random effects in group j, and λ_{1j} ($\lambda_{1j} = (\tilde{w}_j)^{-1}$ or $\lambda_{1j} = (\bar{w}_j)^{-1}$) and λ_{2j} are level 1 and level 2 scaling constants. The variance is estimated by the robust variance estimator

$$\left(\frac{\partial^2 \log L}{\partial \theta^2}\right)^{-1} \left(\sum_j \left(\lambda_{2j} w_j\right) \frac{\partial \log L}{\partial \theta} \left(\frac{\partial \log L}{\partial \theta}\right)' \right) \left(\frac{\partial^2 \log L}{\partial \theta^2}\right)^{-1}$$

The PWIGLS estimator is used by MLwiN software (Rasbash et al. 2005). Asparouhov (2005) claims that the bias produced by the two estimators is nearly identical. However, parameter estimates and standard errors differ. The largest difference is found in the estimation of standard errors. The MPML method performs better in terms of coverage of the true values. Note, that MLwiN was not suitable for us, as it cannot handle plausible values. A comparison of PWIGLS and MPML estimators can be found in Cai (2013).

Results

Empirical study

First, the informativeness of the weights was examined. The level 1 effective sample sizes n_j^{eff} are equal to the actual sample sizes for both chosen countries, meaning that level 1 weights are not informative. However, the weights of the level 2 are informative for both, Sweden and the USA. For Sweden data, N^{eff} is equal to 87, which is less than 153 – the actual number of schools; for the USA data, N^{eff} is equal to 57, which is far less than 501. Thus, the level 2 weights should noticeably affect the results of our multilevel analysis.

The results of the weighted and unweighted null models are presented in Table 2. Important differences in obtained estimates of the variance can be seen even in this simple model. Having no weights or only the scaled level 1 weights (*studwgt* or *totwgt*) produces similar within- and between-school variance estimates. If level 1 weights are unscaled, then the variances change remarkably, especially when *totwgt* is used. The null model with *schwgt* weights only gives similar estimates as the one with scaled weights at both levels. If weights are used at both levels, but level 1 weights are unscaled, then the estimates of variances change drastically, especially when unscaled *totwgt* is used (in the case of Sweden). Note, both scaling methods of level 1 weights produced nearly identical results. Scaled and unscaled level 2 weights, as expected, did not show significant differences in estimation of parameters.

The ICC (see Table 2) showed that 12%, 57%, and 47% (numbers are taken from the unweighted analysis) of the total variance in mathematics achievement was attributable to schools in Sweden, the USA, and the model with both countries, respectively. The design effect was 4.69 in Sweden, 11.62 in the USA, and 11.18 in the combined model of Sweden and the USA. All these three values were greater than 2, which indicate that the grouping of the data needs to be taken into account during estimation.

Table 3 shows results of student models for Sweden and the USA data. As in the null model, unweighted estimates are very close to the ones obtained from the model with the scaled level 1 weights. This finding is not surprising, as the level 1 weights are not informative and should not affect the results. If level 1 weights are not rescaled, then estimates differ more noticeably compared with other cases. We can also note that the significance of the factors *vm* and *sex* has changed in this case. The results of the estimation models with unscaled *studwgt* compared with the ones with unscaled *totwgt* differ remarkably. In some cases, for example, for factor *sex*, a significance level can vary too. Even the indices of goodness of fit AIC, BIC, and deviance are extremely large when level 1 weights are not rescaled, especially when *totwgt* is used. The parameter estimates, standard errors, and fit indices of the student model with *totwgt, houwgt*, and *senwgt* rescaled weights were almost indistinguishable. The use of the level 2 weights (*schwgt*) affects all the estimates, having the largest impact on the estimation of the standard errors and the variances. Almost all standard errors increase, especially for the intercept. The within-school variance slightly decreases by 0.3% for Sweden, and increases by 8% for the USA, compared with the model with scaled *studwgt* only. The between-school variance increases in both Sweden (3%) and the USA (7%). Very similar results are obtained when both, the level 1 and the level 2 scaled weights, are used in the model.

The results of the student models with both studied countries included together were equivalent to the ones from the models using a single country, thus they were excluded in the paper but can be obtained upon request.

Table 2. Null model for Sweden and the USA based on real data (standard errors).

Parameter Sweden	No weights	*studwgt* (unscaled)	*totwgt* (unscaled)	*totwgt, studwgt* (scaled in Mplus*'**)	*schwgt* (scaled in Mplus)	*studwgt* and *schwgt* (both unscaled)	*totwgt* and *schwgt* (both unscaled)	*studwgt* (unscaled) and *schwgt* (scaled in Mplus)	*totwgt* (unscaled) and *schwgt* (scaled in Mplus)	*totwgt, studwgt,* and *schwgt* (both scaled in Mplus**)
Intercept	485.01[3]	484.51[3]	484.49[3]	485.01[3]	488.63[3]	488.11[3]	489.28[3]	488.14[3]	489.49[3]	488.64[3]
	(2.27)	(2.31)	(2.35)	(2.26)	(2.84)	(2.79)	(3.36)	(2.80)	(3.51)	(2.84)
Variance components										
Within-school variance, σ^2	4024.21	3936.22	3861.60	4023.98	3984.79	3926.84	3759.96	3924.58	3757.04	3984.62
Between-school variance, ω^2	574.89	484.506	778.68	568.51	590.26	646.98	831.18	663.44	2243.77	587.89
Proportion of variance explained										
ICC	0.119	0.142	0.163	0.118	0.123	0.138	0.182	0.138	0.182	0.122
USA										
Intercept	507.64[3]	507.43[3]	507.38[3]	507.66[3]	498.55[3]	498.60[3]	497.07[3]	499.15[3]	494.73[3]	498.56[3]
	(2.69)	(2.70)	(2.70)	(2.69)	(12.67)	(12.54)	(13.38)	(12.15)	(13.37)	(12.67)
Variance components										
Within-school variance, σ^2	2441.06	2215.07	2331.08	2439.49	2661.46	2354.50	2262.44	2355.93	2264.89	2661.84
Between-school variance, ω^2	3268.31	3344.35	3402.25	3242.71	4148.57	4069.04	4360.46	2890.58	2854.14	4148.95
Proportion of variance explained										
ICC	0.570	0.601	0.592	0.570	0.588	0.623	0.644	0.623	0.644	0.588

[1] $p < 0.05$.
[2] $p < 0.01$.
[3] $p < 0.001$.
*Scaling method 1.
**Scaling method 2 is used for level 1.

Table 3. Fixed slope student model for Sweden and the USA based on real data (standard errors).

Country	Sweden							USA						
Parameter	No weights	*studwgt*, *totwgt*, *houwgt* (scaled in Mplus**)	*studwgt* (unscaled)	*totwgt* (unscaled)	*schwgt* (scaled in Mplus)	*studwgt* and *schwgt* (both unscaled)	*totwgt*, *studwgt*, and *schwgt* (both scaled in Mplus **)	No weights	*studwgt*, *totwgt*, *houwgt* (scaled in Mplus**)	*studwgt* (unscaled)	*totwgt* (unscaled)	*schwgt* (scaled in Mplus)	*studwgt* and *schwgt* (both unscaled)	*totwgt*, *studwgt*, and *schwgt* (both scaled in Mplus **)
Intercept	485.37^{3} (1.90)	485.37^{3} (1.90)	410.54^{3} (4.51)	484.88^{3} (1.98)	488.51^{3} (2.32)	413.41^{3} (4.61)	412.60^{3} (4.94)	507.95^{3} (2.49)	459.26^{3} (3.38)	462.35^{3} (3.77)	507.70^{3} (2.51)	501.62^{3} (9.64)	452.09^{3} (10.19)	446.78^{3} (10.30)
Student-level factors														
Fixed effects														
msc	59.80^{3} (1.79)	59.80^{3} (1.80)	59.82^{3} (2.07)	59.85^{3} (1.90)	59.51^{3} (2.22)	59.61^{3} (1.88)	59.49^{3} (2.24)	34.17^{3} (1.48)	34.14^{3} (1.48)	32.86^{3} (1.55)	33.05^{3} (1.61)	34.84^{3} (3.21)	33.24^{3} (1.61)	34.82^{3} (3.21)
atm	15.14^{3} (1.87)	15.17^{3}(1.87)	15.33^{3} (1.99)	15.52^{3} (2.01)	16.25^{3} (2.30)	15.63^{3} (2.01)	16.27^{3} (2.30)	8.09^{3} (1.28)	8.09^{3} (1.28)	6.63^{3} (1.50)	9.26^{3} (1.43)	12.14^{3} (2.96)	9.27^{3} (1.43)	12.13^{3} (2.96)
vm	7.84^{1} (3.61)	8.01^{1} (3.66)	7.42 (3.98)	8.14 (4.48)	7.79 (4.73)	8.20 (4.40)	7.98 (4.77)	9.07^{3} (2.25)	8.97^{3} (2.28)	8.40^{2} (1.73)	10.10^{3} (2.58)	11.10^{2} (4.03)	9.97^{3} (2.57)	11.02^{2} (4.01)
ses	31.63^{3} (2.11)	31.64^{3} (2.11)	29.96^{3} (2.60)	27.80^{3} (2.08)	30.18^{3} (2.07)	30.00^{3} (2.11)	30.19^{3} (2.07)	14.94^{3} (1.28)	14.92^{3} (1.28)	12.57^{3} (1.36)	13.45^{3} (1.30)	17.33^{3} (1.82)	13.85^{3} (1.30)	17.32^{3} (1.81)
sex	−5.30^{2} (2.00)	−5.34^{2} (2.01)	−5.69^{1} (2.33)	−6.04^{2} (1.95)	−5.30^{1} (2.10)	−5.97^{2} (1.96)	−5.31^{1} (2.10)	4.90^{2} (1.66)	4.89^{2} (1.65)	5.48^{2} (1.74)	4.19^{1} (1.68)	4.08 (2.98)	4.23^{1} (1.68)	4.06 (2.98)
Variance components														
Within-school variance, σ^2	2746.86	2745.94	2673.55	2640.86	2737.67	2684.92	2738.561	2099.14	2100.48	1913.90	2004.39	2277.41	2024.44	2276.52
Between-school variance, ω^2	382.25	379.86	458.816	534.68	392.77	444.17	387.811	2760.32	2774.78	2940.39	2931.51	2965.30	3169.55	2964.53
Proportion of variance explained														
ICC	0.082	0.082	0.099	0.117	0.087	0.098	0.086	0.522	0.522	0.559	0.546	0.504	0.555	0.504
AIC	52983.01	52980.67	127333.85	913511.38	52983.29	917429.87	52982.21	104768.80	104769.52	1086942.21	32459129.11	105652.48	32623350.23	105652.49
BIC	53034.98	53032.65	127392.88	913586.20	53035.27	917504.69	53034.19	104826.35	104827.06	1087018.65	32459232.71	105278.60	32623453.83	105710.04
−2LL (deviance)	52967.01	52964.67	127317.85	913495.38	52967.29	917413.87	52966.21	104752.80	104753.48	1086926.21	—	105710.03	—	105636.49

[1] $p < 0.05$.
[2] $p < 0.01$.
[3] $p < 0.001$.
**Scaling method 2 is used for level 1.

Table 4. Full model for Sweden based on real data (standard errors).

	Sweden						
Country Parameter	No weights	*studwgt, totwgt, houwgt* (scaled in Mplus**)	*studwgt* (unscaled)	*totwgt* (unscaled)	*schwgt* (scaled in Mplus)	*studwgt* and *schwgt* (both unscaled)	*studwgt, totwgt,* and *schwgt* (both scaled in Mplus **)
Intercept	397.54[3]	397.40[3]	398.15[3]	399.05[3]	402.19[3]	403.45[3]	402.04[3]
	(7.19)	(7.17)	(7.53)	(7.91)	(8.67)	(8.54)	(8.68)
Student-level factors							
Fixed effects							
msc	60.72[3]	60.70[3]	60.65[3]	60.97[3]	60.35[3]	60.58[3]	60.33[3]
	(2.07)	(2.07)	(2.44)	(2.24)	(2.59)	(2.23)	(2.60)
atm	13.47[3]	13.47[3]	13.80[3]	14.09[3]	14.47[3]	14.14[3]	14.47[3]
	(2.02)	(2.01)	(2.26)	(2.22)	(2.59)	(2.23)	(2.59)
vm	8.39[1]	8.36[1]	6.91	8.06	9.63[1]	8.21	9.67[1]
	(3.88)	(3.88)	(4.10)	(4.64)	(4.82)	(4.62)	(4.86)
ses	30.73[3]	30.72[3]	29.53[3]	27.19[3]	29.44[3]	28.37[3]	29.43[3]
	(2.20)	(2.19)	(2.79)	(2.17)	(2.24)	(2.20)	(2.24)
sex	−4.18	−4.22	−3.94	−4.68[1]	−4.14	−4.63[1]	−4.16
	(2.21)	(2.21)	(2.61)	(2.22)	(2.47)	(2.21)	(2.48)
School-level factors							
Fixed effects							
sclim	11.25[2]	11.23[2]	11.79[2]	12.16[2]	10.13[1]	10.41[1]	10.13[1]
	(3.94)	(3.92)	(4.00)	(4.18)	(5.12)	(5.02)	(5.13)
ga	9.37[1]	9.55[1]	10.56[1]	11.88[2]	13.94[2]	14.68[2]	14.04[2]
	(4.36)	(4.34)	(4.50)	(4.58)	(4.99)	(5.03)	(4.98)
sloc	3.99	4.04	4.19	3.32	−0.16	0.02	−0.06
	(3.98)	(3.96)	(4.07)	(4.24)	(4.92)	(4.82)	(4.90)
sre	−1.84	−1.78	−1.95	−3.36	−6.62	−6.60	−6.58
	(5.09)	(5.08)	(5.30)	(5.65)	(6.54)	(6.67)	(6.53)
Variance components							
Within-school variance, σ^2	2749.07	2748.24	2660.61	2634.92	2745.82	2680.11	2745.52
Between-school variance, ω^2	313.45	311.09	382.34	461.54	325.50	373.61	323.47
Proportion of variance explained							
ICC	0.085	0.084	0.102	0.122	0.091	0.102	0.090
AIC	40157.18	40155.40	95219.30	690540.18	40164.63	693417.48	40163.90
BIC	40231.82	40230.04	95304.36	690649.05	40239.27	693526.36	40238.54
−2LL (deviance)	40133.18	40131.40	95195.30	690516.18	40140.63	693393.48	40139.90

[1] $p < 0.05$.
[2] $p < 0.01$.
[3] $p < 0.001$.
**Scaling method 2 is used for level 1.

Table 4 shows results of the full model for Sweden data. The findings from comparison of weighted and unweighted analyses are similar to the ones obtained with the student model. In the full model, we can additionally see that the significance of level 2 factors change depending on what type of weights is used in the model. All standard errors increase when *schwgt* is used, compared with the full model with *studwgt* only. The usage of unscaled level 1 weights affects now both level 1 and level 2 factors. The significance of factors can change at both levels (e.g., for factors *sex* and *ga*) depending on which level 1 weight is used.

Table 5 shows modeling results of the full model for the USA data. In this case, one can see the significance of some level 1 factors changing radically depending on the choice of the

Table 5. Full model for the USA based on real data (standard errors).

	USA						
Country Parameter	No weights	*studwgt, totwgt, houwgt* (scaled in Mplus**)	*studwgt* (unscaled)	*totwgt* (unscaled)	*schwgt* (scaled in Mplus)	*studwgt* and *schwgt* (both nscaled)	*studwgt, totwgt,* and *schwgt* (both scaled in Mplus**)
Intercept	439.23[3]	439.19[3]	442.85[3]	438.79[3]	401.85[3]	405.66[3]	401.84[3]
	(8.93)	(8.93)	(9.09)	(9.06)	(16.28)	(16.60)	(16.28)
Student-level factors							
Fixed effects							
msc	34.04[3]	34.04[3]	32.73[3]	32.78[3]	33.62[3]	32.90[3]	33.62[3]
	(1.69)	(1.69)	(1.76)	(1.86)	(3.60)	(1.85)	(3.60)
atm	7.45[3]	7.44[3]	5.62[3]	8.54[3]	10.80[2]	8.55[3]	10.79[2]
	(1.35)	(1.35)	(1.57)	(1.56)	(3.15)	(1.56)	(3.15)
vm	8.76[3]	8.78[3]	7.47[2]	9.95[2]	11.75[1]	9.86[2]	11.76[1]
	(2.47)	(2.47)	(2.82)	(2.91)	(4.67)	(2.89)	(4.67)
ses	14.64[3]	14.64[3]	12.71[3]	13.58[3]	16.46[3]	13.88[3]	16.45[3]
	(1.43)	(1.43)	(1.49)	(1.46)	(2.06)	(1.46)	(2.06)
sex	4.53[1]	4.53[1]	5.31[2]	3.83[1]	2.22	3.83[1]	2.22
	(1.77)	(1.77)	(1.97)	(1.76)	(2.25)	(1.76)	(2.52)
School-level factors							
Fixed effects							
sclim	24.38[3]	24.38[3]	24.54[3]	24.59[3]	39.87[3]	40.25[3]	39.87[3]
	(5.54)	(5.54)	(5.58)	(5.56)	(10.73)	(10.72)	(10.73)
ga	19.99[2]	19.99[2]	20.24[2]	20.28[2]	40.24[2]	40.73[2]	40.24[2]
	(7.12)	(7.12)	(7.17)	(7.15)	(14.03)	(14.02)	(14.03)
sloc	−6.90	−6.89	−6.97	−6.94	−2.53	−2.15	−2.52
	(5.01)	(5.01)	(5.05)	(5.02)	(8.48)	(8.59)	(8.48)
sre	−8.64	−8.63	−8.83	−8.75	−11.75	−12.34	−11.74
	(6.67)	(6.67)	(6.73)	(6.69)	(9.87)	(9.84)	(9.87)
Variance components							
Within-school variance, σ^2	2046.00	2046.10	2660.61	1964.76	2168.89	1984.18	2169.10
Between-school variance, ω^2	2450.60	2450.69	2597.06	2585.51	2040.12	2177.35	2040.40
Proportion of variance explained							
ICC	0.522	0.522	0.555	0.544	0.525	0.563	0.525
AIC	88187.27	88187.66	914505.15	27514423.56	88667.95	27645573.38	88668.78
BIC	88271.55	88271.94	914617.76	27514577.00	88752.24	27645726.82	88753.06
−2LL (deviance)	88163.27	88163.65	914481.15	—	88643.95	—	88644.78

[1] $p < 0.05$.
[2] $p < 0.01$.
[3] $p < 0.001$.
**Scaling method 2 is used for level 1.

weights. It can change, for example, for factor *vm*, from slightly significant ($p < 0.05$) to very significant ($p < 0.001$), or, as for factor *sex*, from non significant to moderate significant ($p < 0.01$).

Simulation study

The simulation study started with sampling scheme (1), where weights are informative only at level 2. Averaged results of 500 simulations of the simple null model in case A are presented in Table 6. The weighted and the unweighted cases with the level 1 weights give the

Table 6. Simulation means when using the null model in case A and the full model in case B; weights are informative only at level 2.

	Null model				Full model			
Parameter	No weights	Level 1 weights (scaled in Mplus**)	Level 2 weights (scaled in Mplus)	Level 1 and Level 2 weights (both scaled in Mplus**)	No weights	Level 1 weights (scaled in Mplus**)	Level 2 weights (scaled in Mplus)	Level 1 and Level 2 weights (both scaled in Mplus**)
Intercept	1.190	1.190	1.003	1.003	1.199	1.199	1.011	1.011
Level 1 factors								
Fixed effects								
H					0.999	0.999	0.999	0.999
Level 2 factors								
Fixed effects								
S					0.997	0.997	1.003	1.003
Variance components								
Within-group variance, σ^2	0.498	0.498	0.499	0.499	0.500	0.500	0.500	0.500
Between-group variance, ω^2	0.189	0.189	0.182	0.182	0.185	0.185	0.181	0.181

The true values of intercept, *H* and *S* are equal to 1, $\sigma^2 = 0.5$, and $\omega^2 = 0.2$; 500 replications.
**Scaling method 2 is used for level 1.

same results, as well as there is no difference between the model with the level 2 weights only and the model with the weights at both levels. These findings agree with the ones obtained in the empirical study. From the presented simulation results, it is evident that the parameter estimates for the intercept and the within-group variance are less biased in the weighted case, while the between-group variance is closer to the true value in the unweighted case. The difference between the unweighted and the weighted models is quite small. The conclusions from the analysis of the simulated full model in case B are the same as in the null model case (see Table 6). Estimates for the intercept are always overestimated and variances are underestimated. Fixed effects are mostly slightly underestimated or sometimes overestimated.

Findings, obtained using sampling scheme (2), are reported in Table 7. It can be seen that even in this case there is no absolutely best method. Results show that parameter estimates of the factors and the intercept are more precise in the weighted analysis. Note that when level 1 weights are informative, the right choice of scaling method for level 1 weights is important.

The simulated full model in case C presented in Table 8 was more complex in design than the models in cases A and B, but obtained simulation results are similar. In case B, we could see that the parameter estimates for level 1 and level 2 factors were estimated equally well in both the unweighted and the weighted models. However, in case C the parameter estimates for level 1 and level 2 factors were sometimes more precise in the unweighted model, while at other times in the weighted one. In general, it was difficult to find a clear pattern. Note, we have also examined what happens if different ICC values (0.5, 0.6, and 0.7) were chosen for sampling scheme (1); the results did however not change from the earlier described ones, and thus the results were omitted here but can be obtained upon request.

Table 7. Simulation means when using the null model in case A and the full model in case B; weights are informative at both levels.

	Null model					Full model					
Parameter	No weights	Level 1 weights (scaled in Mplus*)	Level 1 weights (scaled in Mplus**)	Level 2 weights (scaled in Mplus)	Level 1 and Level 2 weights (both scaled in Mplus**)	No weights	Level 1 weights (scaled in Mplus*)	Level 1 weights (scaled in Mplus**)	Level 2 weights (scaled in Mplus)	Level 1 and Level 2 weights (both scaled in Mplus**)	True values
Intercept	1.198	1.198	1.198	1.007	1.007	1.209	1.021	1.209	1.021	1.021	1
Level 1 factors											
Fixed effects											
$\underline{H}$						1.001	1.000	1.000	1.001	1.000	1
Level 2 factors											
Fixed effects											
S						0.981	0.990	0.981	0.989	0.990	1
Variance components											
Within-group variance, σ^2	0.499	0.498	0.494	0.498	0.494	0.496	0.494	0.491	0.495	0.490	0.5
Between-group variance, ω^2	0.194	0.194	0.199	0.191	0.195	0.191	0.183	0.196	0.182	0.187	0.2

$\sigma^2 = 0.5$ and $\omega^2 = 0.2$; 500 replications.
*Scaling method 1
**Scaling method 2 is used for level 1.

Table 8. Simulation means when using the full model in case C.

Parameter	No weights	Level 1 weights (scaled in Mplus**)	Level 2 weights (scaled in Mplus)	Level 1 and Level 2 weights (both scaled in Mplus**)	True values
Intercept	4.542	4.542	4.354	4.354	1
Level 1 factors					
Fixed effects					
H_1	0.999	0.999	0.999	0.999	1
H_2	1.000	1.000	1.001	1.001	1
H_3	0.999	0.999	0.998	0.998	1
H_4	1.000	1.000	1.000	1.000	1
H_5	0.999	0.999	0.999	0.999	1
Level 2 factors					
Fixed effects					
S_1	1.002	1.002	1.003	1.003	1
S_2	1.004	1.004	1.002	1.002	1
S_3	1.006	1.006	1.009	1.009	1
S_4	1.002	1.002	0.996	0.996	1
Variance components					
Within-group variance, σ^2	0.499	0.499	0.499	0.499	0.5
Between-group variance, ω^2	0.193	0.193	0.189	0.189	0.2

500 replications.
**Scaling method 2 is used for level 1.

Conclusions

This study aimed to analyze the impact of sampling weights in multilevel models when analyzing complex large-scale assessment data. Although multilevel models are becoming a common way to analyze data of large-scale assessments, still the usage of weights in such models is questioned. For example, one common practice in two-level analysis is to use only the total student weight *totwgt*, which is constructed to be used with single-level models. The analysis with real data shows that usage of scaled *totwgt* or appropriate level 1 weight (*studwgt*) is equivalent to using no weights in the estimation model at all. Additionally, the parameter estimates, the standard errors, and sometimes significance of factors differ from those obtained when level 1 weights or scaled weights on both levels are used in the estimation model. This is not very surprising, because level 1 weights, even if correct ones are used, are not informative and should not affect the results considerably. We note that use of not scaled (rescaled) level 1 weights causes significant differences in some parameter estimates.

Concerning the scaling, the two common scaling methods of level 1 weights produce similar results. The empirical study showed that results sometimes could differ dramatically depending on which weighted or unweighted method an analyst choose. In real data the truth is however unknown, thus a simulation study was used, which enables us to compare models by calculating the bias of the estimates. Our findings from the null and full models for the simulated data using sampling scheme with weights being informative only at level 2 agree with Cai's (2013) results obtained for a student model. For both the null and the full models, less biased between-group variance is obtained using no weights.

The more complex simulation study showed that the differences between unweighted and weighted analysis are not always obvious. It is worth to note that informativeness of the level 2 weights in the complex simulation study was quite high. Findings obtained using sampling scheme with informative weights at both levels indicate the importance of a correct scaling method for the level 1 weights. In our simulation study, the results from the simulated null model can be compared with the ones presented in Pfefferman et al. (1998). Similar estimates were obtained for the intercept and the within-group variance, but opposite differences for the between-group variance. Although we got contradictory results, the differences between the two simulation studies are very small. This finding lifts the significance of the choice of the software used for the data analysis. Different estimators and/or software can have different impact on the results.

In summary, we recommend that researchers examine the informativeness and impact of weights when modeling complex large-scale assessment data such as TIMSS. We also want to emphasize the importance of using scaled weights, if any are used in the model. Scaled weights constructed for a single-level analysis must be rescaled when used with multilevel models otherwise, the parameter estimates become more biased.

In this study, we have analyzed only the usage of the sampling weights in two-level models with students nested in schools. However, students could be nested in classes or teachers; schools could be nested in districts, etc. Future research should focus on how to handle design weights in such cases, also in three-level models.

Funding

This study was funded by the Swedish Research Council, grant number 2015–02160.

References

Asparouhov, T. 2006. General Multilevel modeling with sampling weights. *Communications in statistics. Theory and methods* 35 (3):439–60. doi:10.1080/03610920500476598.

Asparouhov, T., and Muthen & Muthen. 2004. Weighting for unequal probability of selection in multilevel modeling. Mplus Web Notes: No. 8. Retrieved from: https://www.statmodel.com/download/webnotes/MplusNote81.pdf. Accessed 03 July 2014.

Asparouhov, T. 2005. Sampling weights in latent variable modeling. *Structural Equation Modeling* 12 (3):411–34. doi:10.1207/s15328007sem1203_4.

Cai, T. 2013. Investigation of ways to handle sampling weights for multilevel model analyses. *Sociological methodology* 43 (1):178–219. doi:10.1177/0081175012460221.

Carle, A. C. 2009. Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology* 9 (49). Retrieved from: http://www.biomedcentral.com/1471-2288/9/49. Accessed 03 July 2014.

Chantala, K., and C. Suchindran. 2006. Adjusting for unequal selection probability in multilevel models: A comparison of software packages. In *JSM proceedings*, survey research methods section. Alexandria, VA: American Statistical Association. 2815–24. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.627.4453&rep=rep1&type=pdf. Accessed 03 July 2014.

Danielsen, A. G., N. Wiium, B. U. Wilhelmsen, and B. Wold. 2010. Perceived support provided by teachers and classmates and students' self-reported academic initiative. *Journal of School Psychology* 48 (3):247–67. doi:10.1016/j.bbr.2011.03.031.

Howell, D. C. 2008. The analysis of missing data. In *Handbook of social science methodology*, ed. W. Outhwaite and S. Turner, (208–224). London, GB: Sage.

Howie, S. J., and T. Plomp. 2004. The importance of national options in IEA International Comparative Studies: Exploring the effects of language proficiency upon secondary students' performance in Timss'99 mathematics in South Africa. *Proceedings of the IRC-2004*. Retrieved from: http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2004/Papers/IRC2004_Howie_Plomp.pdf. Accessed 03 July 2014.

IEA. 2006. PIRLS 2006 international database. Retrieved from: https://timssandpirls.bc.edu/pirls2006/user_guide.html. Accessed 03 July 2014.

IEA. (2007). TIMSS 2007 international database. Retrieved from: https://timssandpirls.bc.edu/TIMSS2007/idb_ug.html. Accessed 03 July 2014.

IEA. 2011. TIMSS 2011 international database. Retrieved from: http://timssandpirls.bc.edu/timss2011/international-database.html. Accessed 03 July 2014.

IDB Analyzer (Version 2). 2009. [computer software]. Hamburg, Germany: IEA Data Processing and Research center.

Jenkins, F. 2008. Multilevel analysis with informative weights. In JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 2226–33. Retrieved from https://www.amstat.org/sections/SRMS/Proceedings/y2008/Files/301419.pdf. Accessed 03 July 2014.

Jianjun, W. 2000. Relevance of the hierarchical linear model to TIMSS data analyses. opinion papers. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Joncas, M. 2008. TIMSS 2007 sampling weights and participation rates. In *TIMSS 2007 technical report*, ed. J. F. Olson, M. O. Martin, and I. V. S. Mullis, 153–92. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Kim, J. S., C. J. Anderson, and B. Keller. 2013. Multilevel analysis of assessment data. In *Chapman and Hall/CRC statistics in the social and behavioral: Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, ed. L. Rutkowski, M. von Davier, D. Rutkowski, 389–424. Boca Raton, FL, USA: Chapman & Hall/CRC Press. Retrieved from http://www.ebrary.com. Accessed 13 October 2015.

Kish, L. 1965. *Survey sampling*. New York, NY: Wiley.

Kyriakides, L., and C. Charalambous. 2004. Extending the scope of analyzing data of IEA studies: Applying multilevel modelling techniques to analyse TIMSS data. Proceedings of the IRC–2004. Retrieved from http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2004/Papers/IRC2004_Kyriakides_Charalambous.pdf

Lamb, S., and S. Fullarton. 2002. Classroom and school factors affecting mathematics achievement: A comparative study of Australia and the United States using TIMSS. *Australian Journal of Education* 46 (2):154–71. doi:10.1177/000494410204600205.

Marsh, H. W., U. Trautwein, O. Lüdtke, O. Köller, and J. Baumert. 2005. Academic self-concept, interest, grades and standardized test scores: Reciprocal effects models of causal ordering. *Child development* 76 (2):397–416. doi:10.1111/j.1467-8624.2005.00853.x.

Mohammadpour, E., and M. Ghafar. 2012. Mathematics achievement as a function of within- and between school differences. *Scandinavian Journal of Educational Research*, 58 (2):189–221. doi:10.1080/00313831.2012.725097

Muthén, L. K., and B. O. Muthén. 1998–2011. *[computer software]. MPLUS Version 7*. Los Angeles, CA: Muthén & Muthén.

Olson, J. F., M. O. Martin, and I. V. S. Mullis. (Ed.). 2008. *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Pauli, C., K. Reusser, and U. Grob. 2007. Teaching for understanding and/or self-regulated learning? A video-based analysis of reform-oriented mathematics instruction in Switzerland. *International Journal of Educational Research* 46:294–305. doi:10.1016/j.ijer.2007.10.004.

Pfeffermann, D. 1993. The role of sampling weights when modeling survey data. *International Statistical Review* 61 (2):317–37. doi:10.2307/1403631.

Pfeffermann, D., C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash. 1998. Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B* 60:23–40. doi:10.1111/1467-9868.00106.

R Development Core Team. 2014. [Computer software]. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: http://www.R-project.org/. Accessed 03 July 2014.

Rasbash, J., C. Charlton, W. J. Browne, M. Healy, and B. Cameron. 2005. [Computer software]. MLwiN Version 2.02. Centre for Multilevel Modelling, University of Bristol.

Rabe-Hesketh, S., and A. Skrondal. 2006. Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A* 169 (4):805–27. doi:10.1111/j.1467-985X.2006.00426.x.

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

Rutkowski, L., E. Gonzalez, M. Joncas, and M. von Davier. 2010. International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher* 39 (142):142–51. doi:10.3102/0013189X10363170.

Sabah, S., and H. Hammouri. 2010. Does subject matter? Estimating the impact of instructional practices and resources on student achievement in science and mathematics: Findings from TIMSS 2007. *Evaluation & Research in Education* 23 (4):287–99. doi:10.1080/09500790.2010.509782.

Schafer, J. L. 1997. *Analysis of incomplete multivariate data*. London, GB: Chapman & Hall.

Schafer, J. L., and J. W. Graham. 2002. Missing data: Our view of the state of the art. *Psychological Methods* 7 (2):147–77. doi:10.1037//1082-989X.7.2.147.

Snijders, T. A. B., and R. J. Bosker. 2012. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. 2nd ed. London: Sage.

Stapleton, L. 2013. Incorporating sampling weights into single- and multilevel analyses. In *Chapman and Hall/CRC statistics in the social and behavioral: Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, ed. L. Rutkowski, M. von Davier, and D. Rutkowski, 363–88. Boca Raton, FL, USA: Chapman & Hall/CRC Press. Retrieved from http://www.ebrary.com. Accessed 13 October 2015.

Tillé, Y., and A. Matei. 2013. Sampling: Survey sampling. R package version 2.6. Retrieved from: http://CRAN.R-project.org/package=sampling. Accessed 03 July 2014.

Webster, B. J., and D. L. Fisher. 2010. Accounting for variation in science and mathematics achievement: A multilevel analysis of Australian data third international mathematics and science study (TIMSS). *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice* 11 (3):339–60. doi:10.1076/0924-3453(200009)11:3;1-G;FT339.

Zaccarin, S., and C. Donati. 2008. The effects of sampling weights in multilevel analysis of PISA data. Working Paper n. 119. Retrieved from: http://www2.units.it/nirdses/sito_inglese/working%20papers/files%20for%20wp/wp119.pdf. Accessed 03 July 2014.

Appendix

Table A1. Level 1 and level 2 characteristics of the investigated factors from simulated and real data.

		Real data	
Factor	Simulated data Proportion of high	Sweden Proportion of high	USA Proportion of high
Student factors			
Math-self-concept	0.70	0.70	0.73
Attitude toward mathematics	0.52	0.52	0.59
Valuing mathematics	0.92	0.92	0.94
Socioeconomic status	0.69	0.69	0.56
Sex (male)	0.52	0.52	0.49
School factors			
School climate	0.51	0.51*	0.70*
Good attendance	0.64	0.64*	0.80*
School location	0.40	0.40*	0.48*
School resources	0.83	0.83*	0.81*

*Taking into account only the complete observations.