Using Sample Weights in Item Response Data Analysis Under Complex Sample Designs

Xiaying Zheng and Ji Seung Yang

Abstract Large-scale assessments are often conducted using complex sampling designs that include the stratification of a target population and multi-stage cluster sampling. To address the nested structure of item response data under complex sample designs, a number of previous studies proposed the multilevel/multidimensional item response models. However, incorporating sample weights into the item response models has been relatively less explored. The purpose of this study is to assess the performance of four approaches to analyzing item response data that are collected under complex sample designs: (1) single-level modeling without weights (ignoring complex sample designs), (2) the design-based (aggregate) method, (3) the model-based (disaggregate) method, and (4) the hybrid method that addresses both the multilevel structure and the sampling weights. A Monte Carlo simulation study is carried out to see whether the hybrid method can yield the least biased item/person parameter and level-2 variance estimates. Item response data are generated using the complex sample design that is adopted by PISA 2000, and bias in estimates and adequacy of standard errors are evaluated. The results highlight the importance of using sample weights in item analysis when a complex sample design is used.

Keywords Complex sample design • Multilevel item response theory • Sample weights • Pseudo maximum likelihood

1 Introduction

Large-scale educational assessments are often conducted through complex sampling designs for the purpose of reducing costs or improving precision for subgroup analyses relative to simple random sampling (SRS) (Heeringa, West & Berglund 2010). Such design typically includes a stratification of target population and multi-stage cluster sampling within each stratum that result in unequal selection

X. Zheng • J.S. Yang (✉)
College of Education, University of Maryland, 1230 Benjamin Building, College Park, MD 20742, USA
e-mail: xyzheng@umd.edu; jsyang@umd.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 167, DOI 10.1007/978-3-319-38759-8_10

probabilities for different clusters and/or subjects within clusters. Data collected under complex sampling designs have a multilevel structure and sampling weights for units at each level. While traditional item response theory (IRT) models usually assume that examinees are independent and identically distributed across clusters (e.g. schools), these assumptions seldom hold for large-scale assessments that utilize complex sampling designs.

To address the nested structure of item response data under the complex sample designs, a number of previous studies proposed the model-based multilevel item response models (e.g., Adams, Wilson & Wu 1997; Fox & Glas 2001; Jiao, Kamata, Wang & Jin 2012; Kamata 2001), where clustering effects are treated as random effects. Multilevel IRT models have gained popularity in recent years as it addresses the person clustering that is common in education settings. But the sample weights are often not considered in estimating multilevel IRT models.

The second method to analyze complex sample data is design based, which incorporates the complex sample weights into likelihood, resulting in pseudolikelihood for point estimation (see e.g., Binder 1983; Skinner 1989). Taylor Series linearization, jackknifing or balanced repeated replication (BRR) methods are utilized for standard error estimation (see e.g., Rust 1985). However, the design-based method has been less explored in the context of item response models. One example of applying design-based method to IRT models is the work by Mislevy, Beaton, Kaplan, and Sheehan (1992), where a two-stage plausible value method is used to deal with sparse matrix of item responses. In stage 1, a unidimensional IRT calibration is conducted to obtain item parameters through marginal likelihood estimation. In stage 2, multiple imputations (Rubin 1987) of latent scores (also known as plausible values) are conducted via a latent regression model that treats item parameters from stage 1 as fixed. Sample weights are incorporated to the stage 2 model in a design-based manner to estimate parameters and standard errors. The plausible value method provides a practical framework for handling complex samples, and allows convenience for secondary data users. Another example of using design-based method in IRT modeling was explored by Cai (2013), which demonstrates that the sampling weights could be incorporated into one-level multiple-group IRT models to obtain more accurate population-level inferences.

The third approach to dealing with complex sample data combines the model-based and design-based methods by incorporating complex sampling weights in the likelihood of multilevel models. For standard errors, sandwich estimators can be used. The method has previously been evaluated in linear multilevel model (Pfeffermann, Skinner, Holmes, Goldstein & Rasbash 1998) and multilevel logistic regression (Rabe-Hesketh & Skrondal 2006), and has shown superior performance in reducing bias in point estimates. Rabe-Hesketh and Skrondal (2006) characterize this method as "a hybrid aggregated-disaggregated approach". We use "hybrid method" to refer to this combined approach throughout the manuscript. The hybrid method has also been examined using the data of the Trends in International Mathematics and Science Study (TIMSS) in the context of linear multilevel modeling (Laukaityte 2013). As far as the authors are aware of, the hybrid method has never been explored in IRT models.

The purpose of this paper is to assess the performance of four approaches to analyzing item response data that are collected under complex sample designs: (1) single-level IRT without weights, (2) the model-based method (multilevel IRT without weights), (3) the design-based method (single-level IRT with weights), and (4) the hybrid method (multilevel IRT with weights). We are particularly interested in seeing whether the hybrid method can yield the least biased item parameters and level-2 variance estimates under different conditions. To do so, we first briefly introduce complex sampling designs. A multilevel unidimensional model is then described. The marginal pseudolikelihood for the model is presented. The sandwich estimator for standard error estimation is also introduced. Finally a Monte Carlo simulation study is carried out to examine the performance of the pseudo-maximum-likelihood method in comparison with traditional design-based and model-based methods. Bias in estimates and adequacy of standard errors are evaluated across these methods.

Large-scale assessment data are routinely collected with complex sample designs. But the sample weights are often ignored in item analysis, which might lead to biased item parameter estimates and misleading inference on the target finite population. The results of the study highlight the importance of using sample weights in item analysis when a complex sample design is used.

2 Complex Sample Weights

In large-scale tests such as Programme for International Student Assessment (PISA), it is usually not practical to conduct simple random sampling (SRS) on the student level directly. Instead a complex sampling design is implemented to obtain student samples. This paper will keep using the terms "schools" and "students" for illustrative purpose.

Let's consider a complex case of cluster sampling, where stratification is carried out at both levels. The following indices are used:

- $h = 1, \ldots, H$ is the index for stratum at the school level.
- $k = 1, \ldots, K_h$ is the index for school within school-level stratum h.
- $g = 1, \ldots, G_{kh}$ is the index for within-school stratum of school k that is in school-level stratum h.
- $j = 1, \ldots, J_{gkh}$ is the index for student who is from within-school stratum g of school k, where school k is from school-level stratum h.

All schools are first separated to H school-level strata according to some grouping variables (e.g., public or private status and proportion of minority students). Let A_h and a_h be the total number of schools in stratum h and the number of schools to be sampled in stratum h, respectively. Suppose that schools in stratum 1 are over-sampled compared to schools in stratum 2. Then a_1 and a_2 are decided in such a way that $a_1/A_1 > a_2/A_2$.

Within stratum h, a two-stage sampling is carried out, where schools are sampled in the first stage, and students are then selected from each sampled school on the second stage. A common way to conduct the first-stage sampling is through Probability Proportional to Size (PPS) sampling (see e.g., Kish 1965). With PPS, the probability of a school k being sampled is proportional to the total number of students in this school, N_{kh}. Let N_h be the population of students in stratum h. Then the selection probability for school k can be written as:

$$P_{k|h} = a_h \times N_{kh}/N_h. \tag{1}$$

The level-2 weights $W_{k|h}$ is the inverse of $P_{k|h}$.

In the second stage, the stratified random sampling is implemented. Students are further stratified within each school to G groups based on some student-level grouping variables (e.g., ethnicity). Students are then randomly selected from each group. Within school k in stratum h, let N_{gkh} and n_{gkh} be the total number of students in group g, and the number of students to be sampled in group g respectively. Suppose students in group 1 are over-sampled compared to students in group 2. Then n_{1kh} and n_{2kh} are decided in such a way so that $n_{1kh}/N_{1kh} > n_{2kh}/N_{2kh}$.

The conditional selection probability of student j in group g given that his/her school has already been selected is written as:

$$P_{j|g,k,h} = n_{gkh}/N_{gkh}. \tag{2}$$

The level-1 conditional weight $W_{j|gkh}$ is the inverse of $P_{j|gkh}$.

The overall unconditional probability of a student being selected is:

$$P_{jgkh} = P_{k|h} \times P_{j|g,k,h} = a_h \times N_{kh}/N_h \times n_{gkh}/N_{gkh}. \tag{3}$$

As a result, all the students in the same group g, school k, stratum h would have the same overall unconditional selection probability, while schools and students across different strata would have different weights.

3 Multilevel IRT Model and Pseudolikelihood

For illustration purpose, this section describes a two-level 2-parameter logistic IRT model. The marginal pseudolikelihood of the model as well as the sandwich estimator for standard errors are also presented. The IRT model and its estimation could easily be extended to polytomous or mixed item types, and situations with more than two levels.

3.1 Multilevel IRT Model

Let y_{ijk} be the observed response to item i, $(i = 1, \ldots, I)$ for student j in school k. Then θ_{jk}, the latent score for student j in school k, can be expressed as the sum of school level latent mean $\xi_{.k}$ and the individual deviation score δ_{jk}. In a dichotomous two-level unidimensional IRT model, let α_i be the slope on the latent variables at both level 1 and level 2 for cross-level measurement invariance assumption. β_i is the intercept for item i. The conditional likelihood of student j from school k answering item i correctly is:

$$f_{ijk} = f(y_{ijk} = 1 \mid \xi_{.k}, \delta_{jk}) = \frac{1}{1 + \exp(-\beta_i - \alpha_i \xi_{.k} - \alpha_i \delta_{jk})}. \tag{4}$$

3.2 Conventional Likelihood

If we do not consider the complex sample weights, the conditional density for an observed response y_{ijk} is:

$$f_{\boldsymbol{\lambda}}(y_{ijk} \mid \xi_{.k}, \delta_{jk}) = f_{ijk}^{y_{ijk}}(1 - f_{ijk})^{1-y_{ijk}}, \tag{5}$$

where $\boldsymbol{\lambda}$ is a vector of parameters to be estimated. The contribution of a student's responses across all items to the marginal likelihood, conditional on level-2 random effect of school k is:

$$L_{j|k} = \int \prod_{i=1}^{I} f_{\boldsymbol{\lambda}}(y_{ijk} \mid \xi_{.k}, \delta_{jk}) g_1(\delta_{jk}) d\delta_{jk}, \tag{6}$$

where $g_1(\delta_{jk})$ is the distribution of level-1 latent variable δ_{jk}. The contribution of a level-2 school k to the marginal likelihood is:

$$L_k = \int \prod_{j=1}^{J_k} L_{j|k} g_2(\xi_{.k}) d\xi_{.k}, \tag{7}$$

where $g_2(\xi_{.k})$ is the distribution of level-2 latent variable $\xi_{.k}$. The marginal likelihood of the model to be maximized to obtain parameter estimates is the product of each school's contribution to the marginal likelihood:

$$L = \prod_{k=1}^{K} L_k. \tag{8}$$

3.3 Pseudolikelihood

Let $W_{k|h}$ be the conditional weight for school k in stratum h and $W_{j|g,k,h}$ be the conditional level-1 weight for student j in within-school stratum g, given that his/her school k has already been selected in the first stage. The contribution of student j to the marginal pseudolikelihood conditional on level-2 random effect can be obtained by rewriting Eq. (6) with weights as:

$$L^*_{j|gkh} = \int \left[\prod_{i=1}^{I} f_{\boldsymbol{\lambda}} (y_{ijk} \mid \xi_{.k}, \delta_{jk})^{W_{j|g,k,h}} \right] g_1(\delta_{jk}) d\delta_{jk}. \tag{9}$$

And the contribution of school k in stratum h to the marginal pseudolikelihood can be written as:

$$L^*_{k|h} = \int \left[\prod_{g=1}^{G_{kh}} \prod_{j=1}^{J_k} \left(L^*_{j|gkh} \right)^{W_{k|h}} \right] g_2(\xi_{.k}) d\xi_{.k}. \tag{10}$$

Finally, the likelihood of the model is:

$$L^* = \prod_{h=1}^{H} \prod_{k=1}^{K_h} L^*_{k|h} = \prod_{h=1}^{H} \prod_{k=1}^{K_h} \int \left[\prod_{g=1}^{G_{kh}} \prod_{j=1}^{J_k} \left(L^*_{j|gkh} \right)^{W_{k|h}} \right] g_2(\xi_{.k}) d\xi_{.k}. \tag{11}$$

Thus, weights are incorporated into the likelihood of the multilevel model to replicate units at both levels. As Rabe-Hesketh and Skrondal (2006) pointed out, one set of unconditional weights is not sufficient for multilevel pseudo-maximum-likelihood estimation. Level-specific weights must be used at each level.

A number of previous researchers have found that scaling of level-1 weights could affect variance estimates (e.g., Asparouhov 2006; Pfeffermann 1993; Rabe-Hesketh & Skrondal 2006; Stapleton 2002). Several scaling methods have been explored to reduce the bias in the variance components for small cluster sizes. A common scaling method is to scale the level-1 weights to sum up to actual cluster sample size, which is the scaling method used in the simulation study of this paper. Due to space limitation, a discussion of scaling issues is not presented here. A comprehensive investigation of the weight-scaling methods could be found in the work of Asparouhov (2006).

3.4 Sandwich Estimators for Standard Errors

This section summarizes the sandwich estimator for standard errors of multilevel pseudo-maximum likelihood estimates. Detailed derivations about standard error estimations could be found in the works of Asparouhov and Muthén (2006) and Rabe-Hesketh and Skrondal (2006).

When units are independent and identical, the standard errors can be computed using a sandwich estimator:

$$cov(\hat{\boldsymbol{\lambda}}) = \mathbf{A^{-1}BA^{-1}}, \tag{12}$$

A is the observed Fisher information at maximum-likelihood estimates $\hat{\boldsymbol{\lambda}}$. Let $\hat{\boldsymbol{\lambda}}'$ be the transpose of $\hat{\boldsymbol{\lambda}}$. Matrix **A** can be written as:

$$\mathbf{A} \equiv -E\left(\frac{\partial^2}{\partial\hat{\boldsymbol{\lambda}}\partial\hat{\boldsymbol{\lambda}}'} logL\right)\Bigg|_{\lambda=\hat{\lambda}}, \tag{13}$$

while **B**, the outer product of the gradient vector, can be written as:

$$\mathbf{B} \equiv E\left(\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}} logL\right)\left(\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}'} logL\right)\Bigg|_{\lambda=\hat{\lambda}}. \tag{14}$$

A and **B** are the same when the model is the true model. In complex samples, **A** becomes the observed Fisher information at pseudo-maximum-likelihood estimates:

$$\mathbf{A}^* \equiv -E\left(\frac{\partial^2}{\partial\hat{\boldsymbol{\lambda}}\partial\hat{\boldsymbol{\lambda}}'} logL^*\right)\Bigg|_{\lambda=\hat{\lambda}}. \tag{15}$$

B can be obtained by summing the contributions of each independent school (Rabe-Hesketh & Skrondal 2006). Specifically, the first derivatives of the log-pseudolikelihood is:

$$\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}} logL^* = \sum_{h=1}^{H}\sum_{k=1}^{K_h}\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}} logL^*_{k|h}. \tag{16}$$

and **B** is calculated by:

$$\mathbf{B}^* = \sum_{h=1}^{H}\frac{K_h}{K_h-1}\sum_{k=1}^{K_h}\left(\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}} logL^*_{k|h}\right)\left(\frac{\partial}{\partial\hat{\boldsymbol{\lambda}}'} logL^*_{k|h}\right). \tag{17}$$

Finally, the variances of pseudo-maximum-likelihood estimates could be estimated with:

$$cov(\hat{\boldsymbol{\lambda}}) = (\mathbf{A}^*)^{-1}\mathbf{B}^*(\mathbf{A}^*)^{-1}. \tag{18}$$

4 Simulation Design

Motivated by the versatility of the hybrid method in dealing with complex sampling weights, and by the lack of application of such technique to IRT models, this paper attempts to evaluate the performance of hybrid method in IRT models in comparison to other methods. Monte Carlo simulations are carried out to examine the performance of the above mentioned three methods in dealing with complex sample item response data.

The sample design in this paper is partly inspired by PISA 2000 sample design of the United States as described by Rabe-Hesketh and Skrondal (2006). The design includes stratification on both school and student levels, which made both the level-1 and level-2 weights informative. The simulation study chooses this design as an inspiration due to the added complexity of stratification on student level. By adopting this design, the method would be generalizable to more complex situations. Assessments with a less complex design would be a simplification of the scenario presented here.

4.1 Generating Latent Variables and Student Samples

Latent scores of the population are generated with respect to both levels (school level and student level). One level 2, the latent variable is set to follow a normal distribution with mean 0 and variance 3/7. One level 1, the latent variable is set to follow a normal distribution of mean 0 and variance 1. The setup would yield an intraclass correlation (ICC) of 0.3 for the latent variable, which is meant to mimic a fairly large clustering effect of the schools that is typically found in PISA. For example, the results from PISA2003 showed that, the ICC for math outcome was 0.345 across all countries. The ICC for USA was 0.264 (Anderson, Milford & Ross 2009). We have not found any reference on ICCs for PISA2000, but we assume them to be comparable to PISA2003. A population of 400,000 students are generated using above mentioned latent variables. The total number of schools is set to 1000 and the average school size (total number of students in a school) is set to 400. Schools are categorized into public and private schools in such a way that private schools have higher average latent scores than public schools. At level 1, students are categorized into two groups based on ethnicities. The majority group (about 70 % of all the students) are set to have a larger mean latent score than the minority group (about 30 % of the students). The proportions of minority students in each school are then identified. School type and minority status will serve as the basis for stratification in the sampling design.

The sampling method follows the design described in Sect. 2. In level-2 sampling, public schools with at least 15 % minority students are set to be twice as likely to be sampled as other schools. In level-1 sampling, minority students within public schools with at least 15 % minority students are twice as likely to be sampled as other students. As a result, higher latent scores are associated with lower selection probabilities at both levels. Since the selection probabilities are related to the outcome measures (latent scores) on both levels, the resulting sample weights are informative on both levels. Ignoring such sampling design might lead to bias in estimations of item and person parameters in the finite population.

With this method, 75 schools are first selected in the first stage. Then 30 students are selected from each school in the second stage. The final sample has 2250 students.

4.2 Generating Item Response Data

Item response data are generated using a graded response model (Samejima 1969). The generating model is chosen for illustrative purpose only, and not meant to mimic actual PISA items. The sampled latent scores are used to generate 5-category polytomous responses for 20 items using an unidimensional graded response model, with the latent variable split into level 1 and 2. Cross-level measurement invariance is assumed. Let f^*_{ijkx} be the probability of examinee j from school k scoring x or above on item i. The model is defined as:

$$f^*_{ijkx} = f(y_{ijk} \geq x \mid \xi_{.k}, \delta_{jk}) = \frac{1}{1 + \exp(-\beta_{ix} - \alpha_i \xi_{.k} - \alpha_i \delta_{jk})}. \tag{19}$$

The examinee's probability of scoring x can be expressed as:

$$f_{ijkx} = f^*_{ijkx} - f^*_{ijk(x+1)}. \tag{20}$$

4.3 Data Analysis

The generated response data are then analyzed with four methods, namely (1) one-level modeling without weights, (2) one-level modeling with weights (design-based method), (3) two-level modeling without weights (model-based method), and (4) two-level modelling with weights at both levels (hybrid method).

In total, the simulation study has four conditions (four analytical methods). Both Mplus Version 7.2 and flexMIRT® Version 3 are used for method (1), (2) and (3). Results produced by the two packages are identical across the three methods. Only Mplus is used to conduct the analysis with method (4), as no other standard IRT packages implement the hybrid method at this moment as far as the authors are aware of. For the two multilevel models, the variance of the level-1 latent variable is set to 1, leaving the level-2 factor variance to be freely estimated. 100 replications are carried out for each condition.

5 Simulation Results

5.1 Results for Item Parameter Estimates and Standard Errors

The average item parameter estimates (slopes and intercepts) over 100 replications are plotted against the generating true values in order to gauge the biases in the point estimates. As shown in Fig. 1, the biases in point estimates are all fairly small across the four models. The point estimates in the two weighted models (right two panels) are almost unbiased. Both slope and intercept estimates are slightly upward biased in the two unweighted methods (left two panels). The weighted methods are able to yield unbiased item parameters, while the unweighted methods overestimate these point estimates.

The average estimated standard errors for slopes are plotted against the Monte Carlo standard deviations of point estimates in order to evaluate the biases in standard errors in Fig. 2. Using the Monte Carlo standard deviations as the standard, the root-mean-square errors (RMSE) of the estimated standard errors are also calculated. As we can see, the two two-level models (bottom two panels in Fig. 2) yield almost unbiased slope standard errors as the points are closely distributed around the diagonal line. The two one-level models (top two panels in Fig. 2 slightly underestimated the slope standard errors as the points are mostly under the diagonal line. The RMSE also confirm the observation.

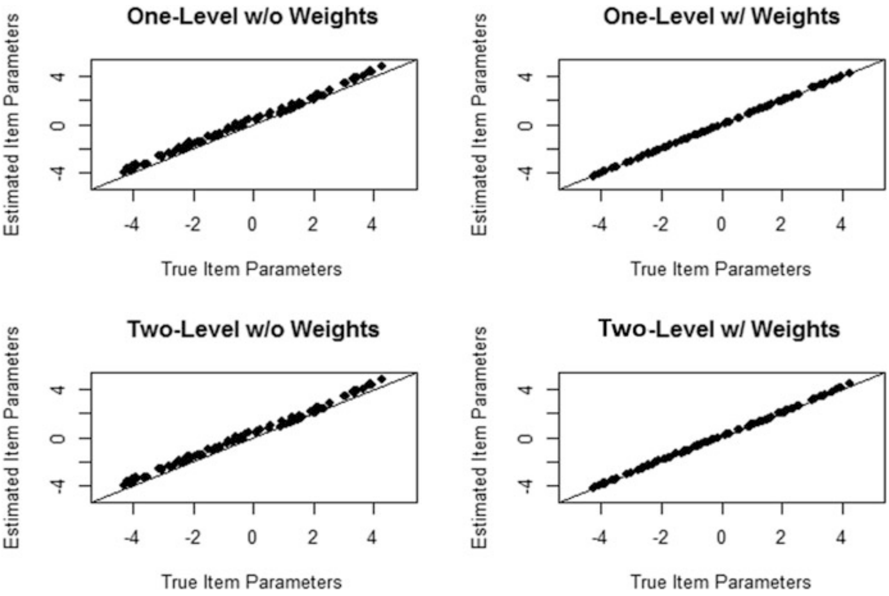

Fig. 1 True item parameters vs. estimates. The unweighted models (*left two panels*) have slightly overestimated the item parameters, while the weighted models (*right two panels*) appears to return unbiased estimates

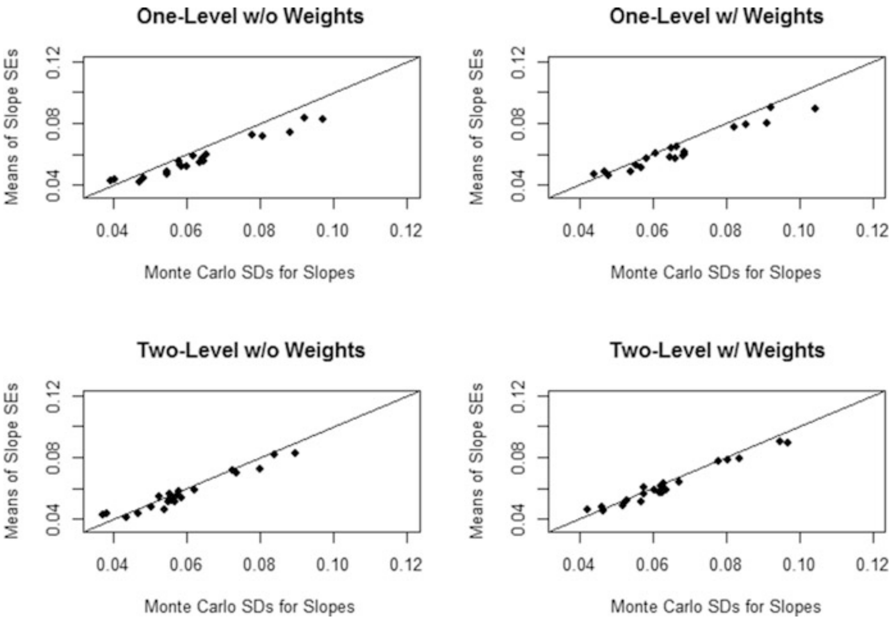

Fig. 2 Monte Carlo standard deviations for slopes vs. means of slope standard errors. The RMSEs of slope standard errors are 0.0076, 0.0065, 0.0042 and 0.0037 respectively for the one-level w/o weights, one-level w/ weights, two-level w/o weights and two-level w/ weights models

The average estimated standard errors for intercepts are plotted against the Monte Carlo standard deviations of point estimates in Fig. 3. The two two-level models (bottom two panels in Fig. 3) yield slightly biased intercept standard errors. The model-based method tends to inflate intercept standard errors, while the hybrid method slightly underestimate the intercept standard errors. The two one-level models (top two panels in Fig. 3) have severely underestimated the intercept standard errors. The RMSEs of the standard errors in the one-level models are expectedly much higher than two-level models.

The 95 % confidence intervals are constructed using the intercept estimates and their standard errors. With both the point estimate and the standard errors taken into account, the coverage rates of the true intercepts in the 95 % confidence interval are very poor in the two unweighted methods, both under 20 % across items, while the same measures for the one-level and two-level weighted methods are 86 and 90 % respectively.

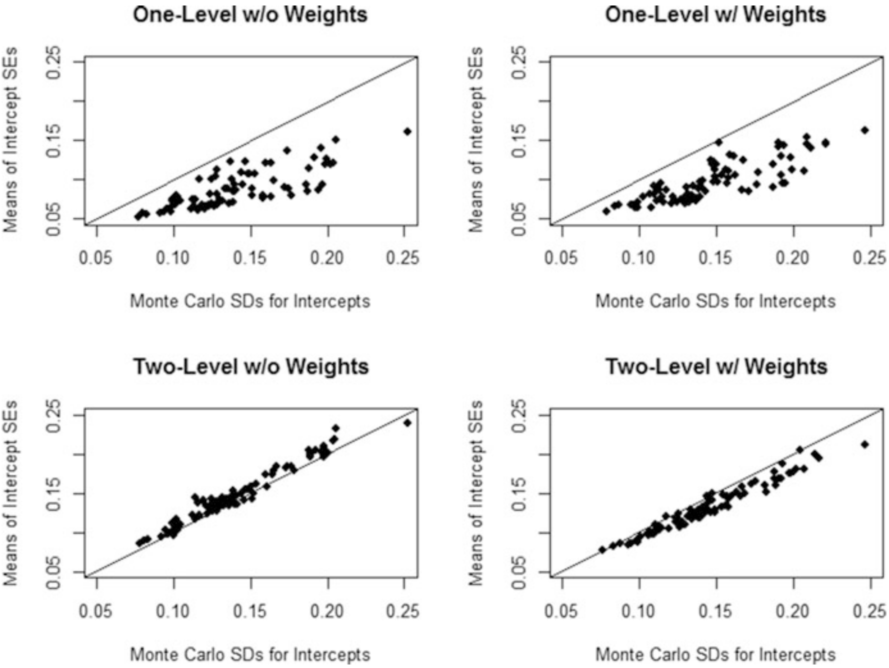

Fig. 3 Monte Carlo standard deviations for intercepts vs. means of intercept standard errors. The RMSEs of intercept standard errors are 0.0330, 0.0313, 0.0078 and 0.0129 respectively for the one-level w/o weights, one-level w/ weights, two-level w/o weights and two-level w/ weights models

5.2 Results for Level-2 Variance

Coverage of true second-level (between-school) variance in the 95 % confidence intervals of estimates is plotted in Fig. 4 for the weighted and unweighted multilevel models. It appears that the hybrid method has some advantages over traditional two-level models, as the hybrid method achieves a less biased level-2 variances and better coverage of true level-2 variance. In fact, the average percentage bias for the level-2 variance is 14 % in the unweighted model, while the same measure for the hybrid model is only −2 %. The coverage rates of true level-2 variance in the 95 % confidence intervals are 82 and 91 % respectively for the unweighted and weighted two-level models.

6 Discussion

We compared the performance of three methods to analyze item response data collected under a complex sample design, with a special interest in the performance of the pseudo-maximum-likelihood estimation method for multilevel IRT models (the hybrid method). The results show that, methods accounting for complex sample

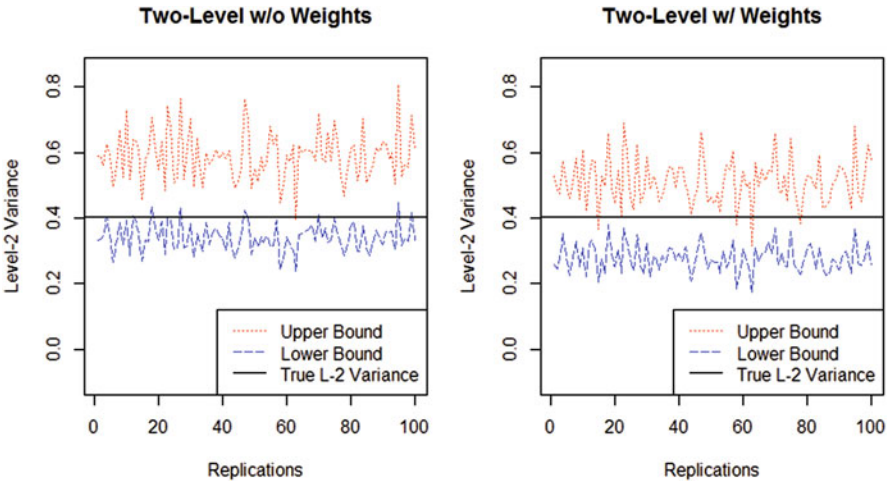

Fig. 4 Coverage of true level-2 variance in 95 % confidence intervals of estimates. The coverage rates of true level-2 variance in the 95 % confidence intervals are 82 and 91 % respectively for the unweighted and weighted two-level models

weights produce less biased point estimates for item parameters in either single-level or multilevel models, while multilevel modeling yields more accurate standard errors for item parameters than single-level models. It is worth nothing that, in the unweighted multilevel model, the coverages of the true parameters are very poor. Better standard error estimates do not seem to make up for deficiency in point estimates. The hybrid method, which accounts for both the complex sampling weights and the multilevel data structure, indeed combines the advantages of both the design-based and model-based methods. Under the unidimensional model, the performance of the hybrid method is superior to the others in terms of estimating item parameters.

The hybrid method does show great potential in analyzing testing data collected with complex sampling designs. One practical obstacle for implementing the hybrid method is the fact that it requires conditional weights for lower-level units which survey agencies generally do not release. If conditional weights are not available, and level-2 variance is not of primary interest, the authors would recommend using the total unconditional weights with single-level modeling to obtain more accurate item estimates.

There are a few limitations to the current research. First, the simulation study only uses one type of sample design. More sampling schemes should be examined to fully gauge the performance of the hybrid method, including informativeness of weights, selection mechanism, cluster size and so on. Second, the generating ICC of 0.3 in the simulation study is meant to mimic a large clustering effect. ICCs of other magnitudes should be explored to evaluate the performance of different methods. Third, an empirical illustration is missing in current research due to unavailability of level-1 conditional weights in PISA data. Last but not least, the role of weight scaling methods has not been examined.

Our future research includes comparisons of standard errors estimated with alternative methods, evaluating the weight-scaling methods under different sample designs, and expanding the hybrid method to multi-dimensional multilevel IRT models, such as simple cluster models or testlet models.

Acknowledgements This research is supported in part by the Institute for Education Sciences, U.S. Department of Education, through grants R305D150052. The opinions expressed are those of the authors and do not represent the views of the Institute or the Department of Education. We would like to thank Dr. Li Cai for providing his unpublished paper on unidimensional model with weights and flexMIRT® program.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47–76.

Anderson, J. O., Milford, T., & Ross, S. P. (2009). Multilevel modeling with HLM: Taking a second look at PISA. In *Quality research in literacy and science education* (pp. 263–286). Dordrecht: Springer Netherlands.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods, 35*(3), 439–460.

Asparouhov, T., & Muthén, B. (2006). Multilevel modelling of complex survey data. In *Proceedings of the Survey Research Methods Section, American Statistical Association 2006* (pp. 2718–2726).

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review, 51*, 279–292.

Cai, L. (2013). Multiple-group item response theory analysis in the case of complex survey data. *Contributed Paper for World Statistics Congress Session Latent Variable Modeling of Complex Survey Data*, August 2013, Hong Kong.

Fox, J., & Glas, C. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 269–286.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: CRC Press.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*(1), 82–100.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79–93.

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

Laukaityte, I. (2013). *The importance of sampling weights in multilevel modeling of international large-scale assessment data.* Paper presented at the 9th Multilevel conference, Utrecht, March 27–29.

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

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review, 61*, 317–337.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B, Statistical methodology, 60*, 23–40.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A, 169*(4), 805–827.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. NJ: John Wiley & Sons.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics, 1*(4), 381–397.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores *(Psychometric Monograph No. 17)*. Richmond, VA: Psychometric Society.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In C. J. Skinner, D. Holt & T. M. F. Smith (Eds.), *Analysis of complex surveys* (pp. 59–87). New York, NY: Wiley.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9*(4), 475–502.