Sampling Weights and Item Calibration in Large-Scale Assessment

Jiaying Xiao

Large-scale assessments (LSAs) such as Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and National Assessment of Educational Progress (NAEP) provide reliable measures of educational achievement for group of students. The results of LSAs could be used to evaluate the equity and quality of educational systems, and to help policymakers and stakeholders to make informed decisions (Arikan et al., 2020; Meinck, 2020). The reliability and validity of the LSAs rely on obtaining a representative sample in which sampling design plays a key role. In general, LSAs employ a complex multistage sampling design where students are sampled within schools, and schools are sampled within higher-level units (i.e., states or countries; Rust, 2013; Laukaityte & Wiberg, 2018; OECD, 2018). This complex sampling design results in some units in the population chosen with unequal probabilities. When the analytic model such as the multilevel modeling does not account for different selection probabilities, biased parameter estimates and incorrect conclusions would be obtained (Pfeffermann et al., 1998; Laukaityte & Wiberg, 2018). One solution is to incorporate sampling weights into the model properly when estimating the population characteristics.

While previous studies emphasize the necessity of using sampling weights for analyzing LSAs data in multilevel modeling (Rutkowski et al., 2010; Laukaityte & Wiberg, 2018; Arikan et al., 2020), applying sampling weights to the item response theory (IRT) model, especially in the calibration process, is rarely discussed (Zheng & Yang, 2016). The purpose of this study is to evaluate the use of sampling weights in IRT calibration. The paper is structured as follows. The literature review section starts by briefly describing sampling weights and IRT calibration, followed by an overview of studies at the intersection of these two topics. Then, the simulation study section provides comparison between the models with and without sampling weights. The discussion section is given in the end.

Literature Review

Sampling Weights

To accommodate the fact that some units are selected with unequal probabilities, sampling weights, taken to be the inverse of the probability of a unit being selected, are applied to LSA studies. Suppose that a sample consisting of 5 boys and 5 girls is selected from a population with 20 boys and 10 girls. The probability of a girl being selected is $5/10$ and the probability of a boy being selected is $5/20$. Accordingly, the sampling weights for girls and boys are 2 and 4, respectively. These sampling weights are referred to as raw sampling weights (also named unscaled sampling weights), meaning that the sum of the weights within a sample is equal to the

population size (Thomas & Heck, 2001). In this example, sampling weights add up to the population size 30 ($2 \times 5 + 4 \times 5 = 30$).

Ignoring sampling weights results in biased estimates of population characteristics. This happens in the previous example since boys have a larger proportion in the population, but the sample includes the same number of boys and girls. More specifically, suppose the 5 boys get $2, 3, 3, 4, 2$ points from an exam and the 5 girls get $4, 5, 5, 3, 4$ points. The unweighted mean score of this sample is $\frac{2+3+3+4+2+4+5+5+3+4}{10} = 3.5$, whereas the weighted mean score is $\frac{(2+3+3+4+2)\times 4+(4+5+5+3+4)\times 2}{30} \approx 3.27$. The weighted mean score is nearly 7% lower than the unweighted mean score. In LSA, this discrepancy could lead to the enactment of a different education policy (Rutkowski et al., 2010).

The previous example adopts a single-level sampling scheme (i.e., boy/girl). In practice, the sampling schemes of LSAs are much more complicated and typically consist of multiple levels. For example, PISA usually adopts the following two-level scheme. At the first level, we sample a school with probability proportional to its size, whereby small schools are selected with lower probability, and at the second level, we sample students randomly from the selected school (Brewer & Hanif, 1983; OECD, 2018; Arikan et al., 2020). In such cases with multilevel sampling schemes, analyzing LSAs data without sampling weights or with only single-level sampling weights could yield misleading conclusions (Laukaityte & Wiberg, 2018). Consequently, many software programs such as Mplus (Muthén, & Muthén, 2017) and HLM (Raudenbush et al., 2011) have been developed to allow users to input sampling weights at multiple levels.

IRT Calibration

IRT model (Lord & Novick, 1968) is commonly used in educational tests and LSAs for modeling the relation between the probability of answering an item correctly and the latent ability of the respondent. Among the many IRT models proposed in the literature, we focus throughout this study on the unidimensional two-parameter logistic (2PL) model. For a dichotomous item j, the probability of a respondent i getting a correct answer ($Y_{ij} = 1$) is given by

$$P(Y_{ij} = 1|\theta_i) = \frac{\exp\left(a_j(\theta_i - b_j)\right)}{1 + \exp\left(a_j(\theta_i - b_j)\right)}, \quad (2)$$

where a_j and b_j denote the item discrimination and difficulty parameters, respectively, and θ_i denotes a continuous latent ability measured by the assessment.

In LSAs like PISA, item parameters (i.e., a_j and b_j) are first calibrated based on the response data from all participating countries and then treated as known when estimating the proficiency levels of respondents and inferring the proficiency distributions of countries (OECD, 2016; Chen et al., 2022). Existing item calibration methods include separate calibration with linking (Loyd & Hoover, 1980; Stocking & Lord, 1983), fixed parameter calibration (FPC; Ban et al., 2001; Kim, 2006), and concurrent calibration (Bock & Zimowski, 1997; Cai, Albano, &

Roussos, 2021), where the first two are sometimes combined into one under the name of separate calibration (i.e., Hanson & Béguin, 2002; von Davier et al., 2019). Let us now describe these three methods in more detail. Separate calibration with linking involves two steps. First, item parameters for all items on two test forms are calibrated separately. Second, two sets of item parameters are placed onto the same scale using a linear transformation method. The transformation coefficients are estimated by using two sets of item parameters of common items (Hu, Rogers, & Vukmirovic, 2008). FPC fixes item parameters of common items and calibrates new items only (Kim, 2006). The transformation step is not needed because fixed item parameters of common items set the same scale for all the items. Alternatively, concurrent calibration estimates all items on both forms simultaneously to ensure all item parameters are on the same scale (Cai, Albano, & Roussos, 2021).

For each of the calibration methods described above, there exist many procedures to estimate the item parameters. For both FPC and concurrent calibration, a widely used approach is the marginal maximum likelihood estimation with expectation maximization algorithm (MMLE-EM) proposed in (Bock & Aitkin, 1981; Mislevy & Bock, 1985); see (Kim, 2006; Wang, Chen, & Jiang, 2020) for some recent applications. The simulation design will focus on the concurrent calibration with MMLE-EM.

It is necessary to take the complex sampling design into account when calibrating item parameters of LSAs. Roughly speaking, there are three main approaches in the literature. The first approach employs a multiple-group IRT approach (Bock & Zimowski, 1997). The multiple-group IRT assumes different subgroup distributions (von Davier & Yamamoto, 2004; von Davier et al., 2019; Zheng & Yang, 2021). The second one is the multilevel IRT model to account for the nested data structure due to the sampling design (i.e., Kamata, 2001; Jiao et al., 2012; Zheng & Yang, 2016). The third one, incorporating sampling weights into the IRT calibration procedure (i.e., MMLE-EM), however, has been rarely explored (Zheng & Yang, 2016). The current study mainly focuses on the discussion of the third approach.

Sampling Weights in IRT Calibration

Sampling weights and IRT calibration are two important techniques in LSAs. In a technical report of PISA, it mentions that one type of sampling weights, senate weight (sum of weights equals to 500 for each country), is used to minimize the effect of different country sizes on scaling during the item calibration process (Oliveri & von Davier, 2014; OECD, 2018). However, there has been few studies in the literature that discuss the effects of sampling weights on IRT calibration. In Zheng and Yang's study (2016), they incorporated sampling weights into the likelihood function, and evaluated the performance of four models when analyzing nested response data: (1) a single-level IRT without weights, (2) a multilevel IRT without weights; (3) a single-level IRT with sampling weights; and (4) a multilevel IRT with weights. The simulation results showed that models accounting for sampling weights produced less biased estimates for item parameters in both single and multilevel IRT models, while multilevel IRT models yielded more accurate standard error estimates for item parameters comparing with single-level IRT models. Another relevant study (Smits, 2016) adopted sampling weights, and multiple-group IRT model, respectively, and compared these two approaches with a standard IRT model which

ignored the sampling design. The simulation results indicated that ignoring the sampling design produced bias in both item and person parameters. There is another line of research that applied sampling weights to the IRT calibration, but their discussions mainly focused on the recovery of transformation coefficients instead of item parameters (i.e., Qian, Jiang, & von Davier, 2013; Qian, von Davier, & Jiang, 2013). In summary, all these studies highlight the importance of incorporating sampling weights into the likelihood function for IRT calibration.

Simulation Study

Simulation Design

A simulation study was conducted to investigate the effects of sampling weights and sample size on IRT calibration. The unidimensional 2PL model was used to generate response data. The test length was fixed at 30. Item parameters were generated from the following distributions: $a_j \sim logN(0, 0.5)$, $b_j \sim N(0, 1)$. These distributions were chosen from the literature (see Suh, Cho, & Wollack, 2012). The population size was fixed at 5000, and the latent ability parameter for each person (θ_i) was generated from the standard normal distribution. The population was divided into high performing group (the highest 20% of ability parameters) and low performing group (the lowest 80% of ability parameters). Two sample sizes were considered: 500 and 1000, where the sample size of 500 represented the minimal sample size required for 2PL model (König, Spoden, & Frey, 2020). The sample of 500 was obtained by randomly selecting 25% of the students from the high performing group and 6.25% of the students from the low performing group. Similarly, the sample of 1000 consisted of 50% of the students from the high performing group and 12.5% of the students from the low performing group.

Scaled Sampling Weights

To incorporate sampling weights into the IRT calibration, the raw sampling weights discussed in the Introduction section could not be included in the likelihood function directly. Instead, it is necessary to scale the sampling weights (Carle, 2009), for which several methods have been proposed in the literature (Pfeffermann et al., 1998; Asparouhov 2006; Carle, 2009). The current study adopts a popular one proposed in Asparouhov (2006), where the scaled sampling weights sum up to the sample size:

$$w_i^* = \frac{w_i n}{N}, \tag{1}$$

where w_i^*, w_i refer to the scaled and raw sampling weights for respondent i, respectively, and n and N refer to the sample and population size.

For both sample size conditions, the scaled sampling weights for high and low performing groups were fixed at 0.4 and 1.6, respectively.

MMLE-EM

The MMLE-EM algorithm was implemented to estimate item parameters. To investigate the effects of sampling weights, two likelihood functions were considered for item calibration:

weighted likelihood function (denoted as W-MML hereafter) and regular likelihood function (denoted as MML hereafter).

We first describe the MML procedure. Let $\boldsymbol{\Delta} = (\mathbf{a}, \mathbf{b})$ refer to the set of unknown item parameters, which are to be estimated in item calibration. When sampling weights are not considered, the marginal likelihood function of $\boldsymbol{\Delta}$ under a 2PL model is

$$\mathrm{L} = \mathrm{L}(\boldsymbol{\Delta}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = \prod_{i=1}^{N} \int P(\boldsymbol{y}_i|\theta_i, \boldsymbol{\Delta})\, g(\theta|\tau) d\theta, \tag{3}$$

where $P(\boldsymbol{y}_i|\theta_i, \boldsymbol{\Delta}) = \prod_{j=1}^{J} [P_j(\theta_i)^{y_{ij}} (1 - P_j(\theta_i))^{1-y_{ij}}]$ with $P_j(\theta_i)^{y_{ij}}$ computed using Equation (1), and $g(\theta|\tau)$ refers to the density function of θ with τ denoting the parameters of the normal distribution. After some algebraic transformations and quadrature approximation, the derivative of the logarithmic transformation of the marginal likelihood equation (log likelihood function) with respect to a_j can be expressed as

$$\begin{aligned} \frac{\partial}{\partial a_j} \log \mathrm{L} &= \sum_{i=1}^{N} \int [y_{ij} - P_j(\theta)](\theta - b_j) P(\theta|\boldsymbol{y}_i, \boldsymbol{\Delta}, \tau) d\theta \\ &\approx \sum_{i=1}^{N} \sum_{k=1}^{Q} [y_{ij} - P_j(\theta_k)](\theta_k - b_j)\, P(\theta_k|y_i, \boldsymbol{\Delta}, \tau) = 0, \end{aligned} \tag{4}$$

where $P(\theta_k|y_i, \boldsymbol{\Delta}, \tau) = \frac{P(\boldsymbol{y}_i|\theta_k, \boldsymbol{\Delta}) g(\theta_k|\tau)}{P(\boldsymbol{y}_i|\boldsymbol{\Delta})} = \frac{P(\boldsymbol{y}_i|\theta_k, \boldsymbol{\Delta}) g(\theta_k|\tau)}{\sum_k^Q P(\boldsymbol{y}_i|\theta_k, \boldsymbol{\Delta}) g(\theta_k|\tau)}$ refers to the posterior distribution of θ_k. To solve the equation, two auxiliary parameters are defined here:

$$\bar{n}_{jk} = \sum_{i=1}^{N} P(\theta_k \mid \boldsymbol{y}_i, \boldsymbol{\Delta}, \tau) = \sum_{i=1}^{N} \left[\frac{P(\boldsymbol{y}_i \mid \theta_k, \boldsymbol{\Delta}) g(\theta_k \mid \tau)}{\sum_k^Q P(\boldsymbol{y}_i \mid \theta_k, \boldsymbol{\Delta}) g(\theta_k \mid \tau)} \right], \tag{5}$$

$$\bar{r}_{jk} = \sum_{i=1}^{N} y_{ij} P(\theta_k \mid \boldsymbol{y}_i, \boldsymbol{\Delta}, \tau) = \sum_{i=1}^{N} \left[\frac{y_{ij} P(\boldsymbol{y}_i \mid \theta_k, \boldsymbol{\Delta}) g(\theta_k \mid \tau)}{\sum_k^Q P(\boldsymbol{y}_i \mid \theta_k, \boldsymbol{\Delta}) g(\theta_k \mid \tau)} \right]. \tag{6}$$

Using $\bar{n}_{jk}$ and $\bar{r}_{jk}$, Equation (4) can be rewritten as:

$$\frac{\partial}{\partial a_j} \log \mathrm{L} = \sum_{k=1}^{Q} (\theta_k - b_j)[\bar{r}_{jk} - \bar{n}_{jk} P_j(\theta_k)] = 0. \tag{7}$$

Similarly, the derivative with respect to b_j can be expressed as:

$$\frac{\partial}{\partial b_j} \log \mathrm{L} = \sum_{k=1}^{Q} -a_j[\bar{r}_{jk} - \bar{n}_{jk} P_j(\theta_k)] = 0. \tag{8}$$

To estimate item parameters, we use the EM algorithm detailed as follows:

1. Initialize all item parameters ($t = 1$): a_j^1, b_j^1;

2. At the t-th iteration, compute $\bar{n}_{jk}^t$, $\bar{r}_{jk}^t$ using Equations (5) and (6) [E-step];

3. Solve Equations (7) and (8) to update a_j^{t+1}, b_j^{t+1} [M-step];

4. Repeat steps 2 and 3 until convergence criterion is met. In this study, the convergence criterion is met when the maximum difference between item parameter estimates in two consecutive steps is smaller than 0.001.

Next, we describe the W-MML, where sampling weights are incorporated into the likelihood function. In this case, $P(\boldsymbol{y}_i|\theta_i, \boldsymbol{\Delta})$ becomes $P(\boldsymbol{y}_i|w_i, \theta_i, \boldsymbol{\Delta})$:

$$P(\boldsymbol{y}_i|w_i, \theta_i, \boldsymbol{\Delta}) = \prod_{j=1}^{J}[P_j(\theta_i)^{w_i y_{ij}}(1 - P_j(\theta_i))^{w_i(1-y_{ij})}]\,. \tag{9}$$

Accordingly, Equations (5) and (6) are modified as:

$$\bar{n}_{jk} = \sum_{i=1}^{N} P(\theta_k \mid \boldsymbol{y}_i, \boldsymbol{\Delta}, \tau, w_i) = \sum_{i=1}^{N}\left[\frac{w_i P(\boldsymbol{y}_i \mid w_i, \theta_k, \boldsymbol{\Delta}) g(\theta_k \mid \tau)}{\sum_k^Q P(\boldsymbol{y}_i \mid w_i, \theta_k, \boldsymbol{\Delta}) g(\theta_k \mid \tau)}\right], \tag{10}$$

$$\bar{r}_{jk} = \sum_{i=1}^{N} y_{ij} w_i P(\theta_k \mid \boldsymbol{y}_i, \boldsymbol{\Delta}, \tau) = \sum_{i=1}^{N}\left[\frac{y_{ij} w_i P(\boldsymbol{y}_i \mid \theta_k, \boldsymbol{\Delta}) g(\theta_k \mid \tau)}{\sum_k^Q P(\boldsymbol{y}_i \mid \theta_k, \boldsymbol{\Delta}) g(\theta_k \mid \tau)}\right]. \tag{11}$$

These modified auxiliary parameters are used to find solutions for Equations (7) and (8). The EM algorithm follows the same steps as described above.

Evaluation Criteria

One hundred replications were conducted for each condition. The bias and root mean squared error (RMSE) of item discrimination and difficulty parameters were adopted to evaluate the parameter recovery, which were computed across all replications per item. For instance, the bias and RMSE for a_j are

$$\text{Bias}_{a_j} = \frac{\sum_{r=1}^{R} \tilde{a}_{jr} - a_j}{R}, \tag{12}$$

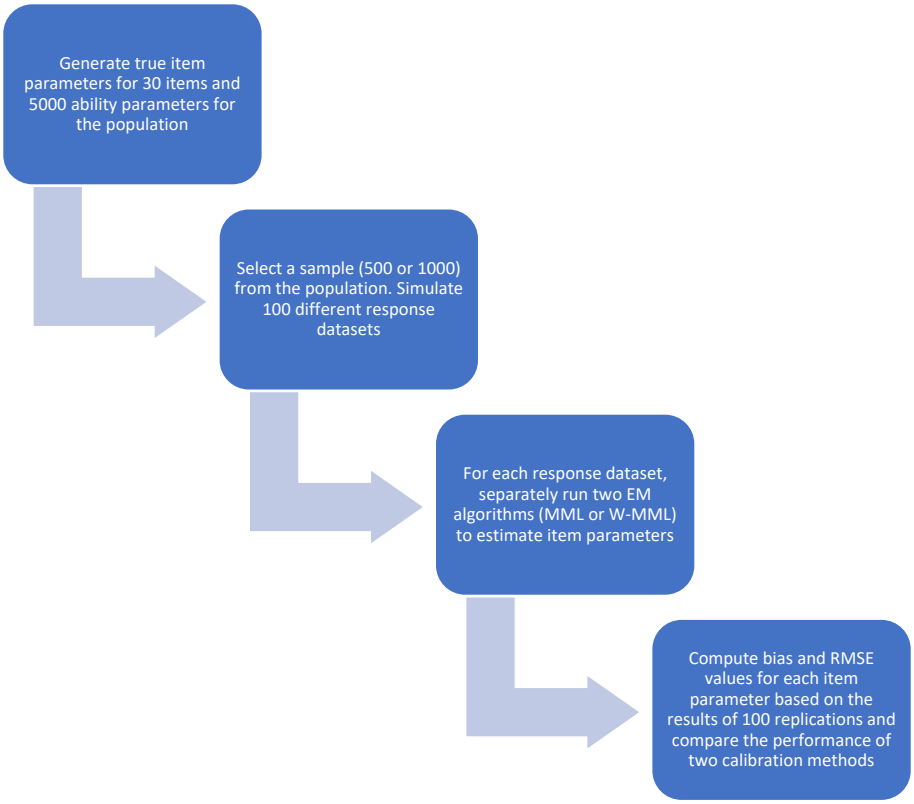
$$\text{RMSE}_{a_j} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\tilde{a}_{jr} - a_j\right)^2}, \tag{13}$$

where a_j is the true parameter, and $\tilde{a}_{jr}$ refers to its estimate in the r-th replication.

The process of the simulation study is summarized in Figure 1.

Figure 1

The process of the simulation study



Results

Figure 2 presents the bias results of item parameters using two calibration methods (MML and W-MML) under different sample sizes (500 and 1000). The boxplots represent the bias results for each item parameter computed from Equation (12). We see that, in general, item discrimination parameters (a_j) were overestimated, whereas item difficulty parameters (b_j) were underestimated, regardless of the sample size. Compared with MML, W-MML yielded less biased estimates for both item parameters, and this discrepancy was more substantial when estimating b_j. W-MML also produced unbiased results for some item discrimination parameters, with the caveat that this calibration method was more likely to produce outliers. MML yielded more variability of bias results across items. We also note that in the current setting, the effect of sample size was almost negligible.

Figure 2

Boxplot of Bias Results of 30 Items for each parameter

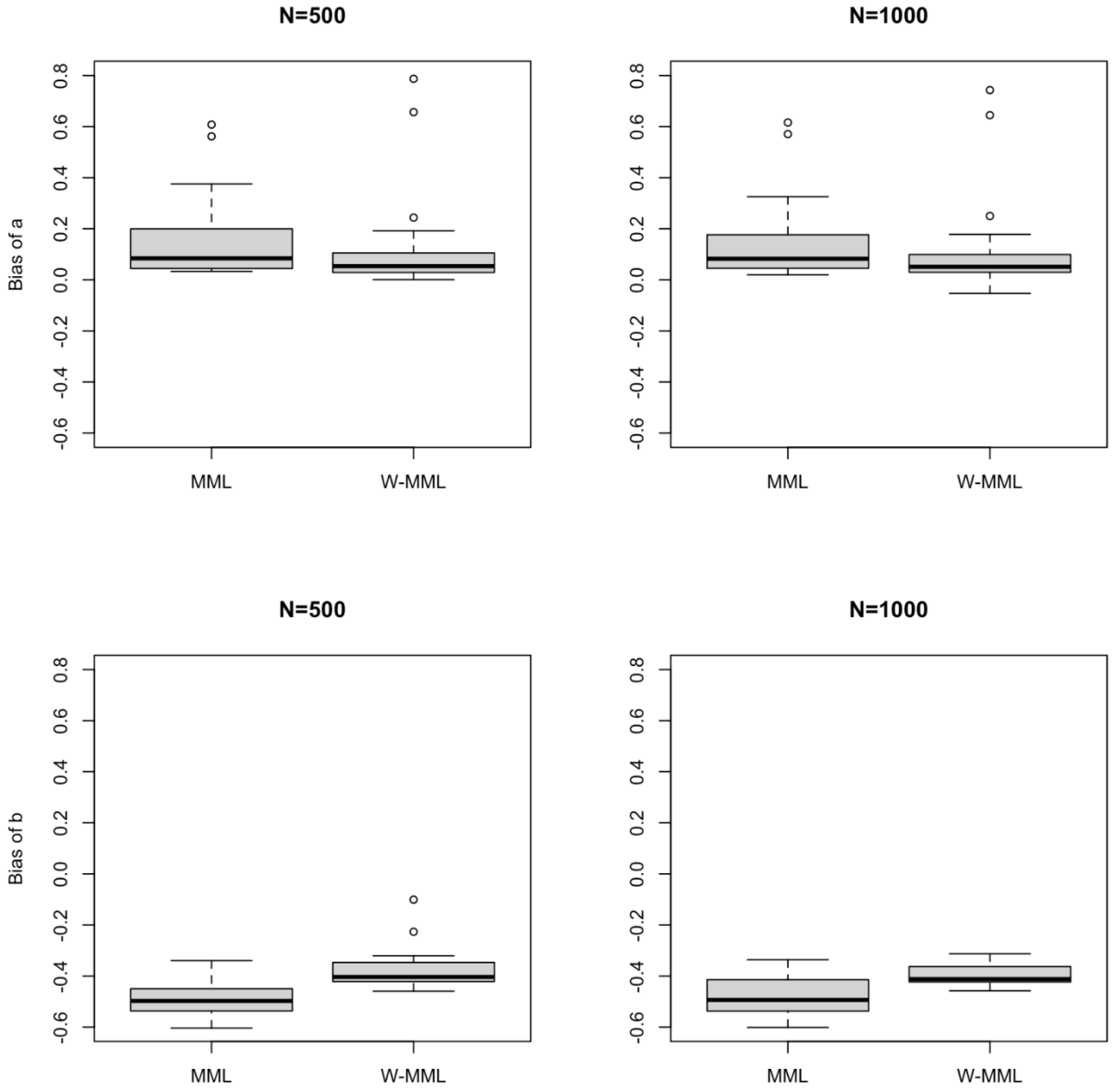
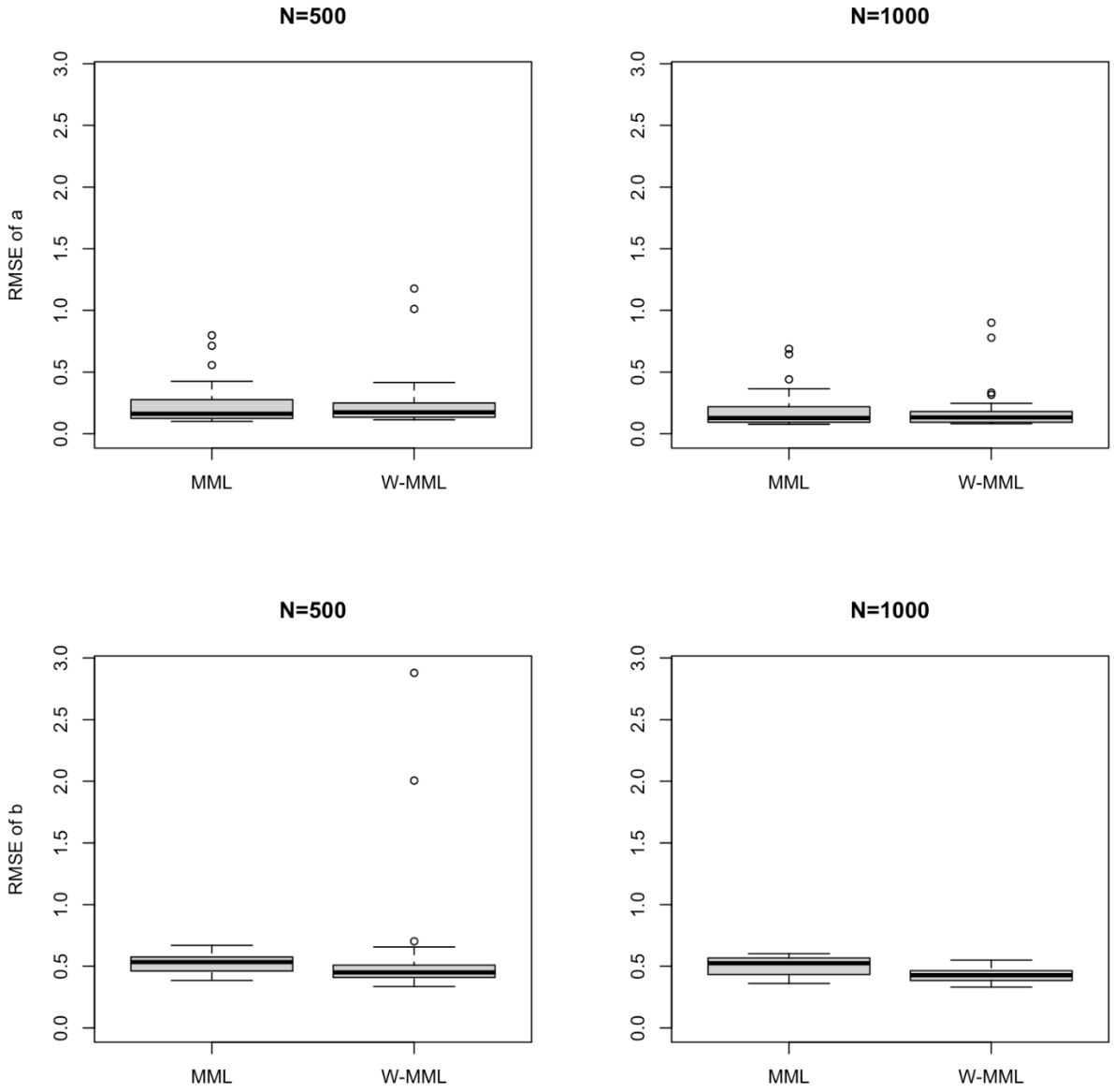


Figure 3 presents the distribution of RMSE values for each item parameter. Excluding the outliers, the RMSE results from MML and W-MML were almost indistinguishable in terms of the estimates of a_j, whereas W-MML produced smaller RMSE values for b_j. Increasing the sample size slightly improved the accuracy of item parameter estimates for both calibration methods. For W-MML, the outlier became an issue under the small sample size condition (i.e., n=500), especially for the estimate of b_j. The RMSE value for one item difficulty parameter almost reached 3.0.

Figure 3

Boxplot of RMSE Results of 30 Items for each parameter



Discussion

In the current study, the effect of sampling weights on IRT calibration was investigated. The simulation design considered two sample size conditions. Bias and RMSE values were computed to evaluate the performance of likelihood function (MML) and weighted likelihood function (W-MML). The results showed that the item calibration accounting for sampling weights produced more accurate estimates of item parameters under both conditions, which were consistent with previous findings (Smits, 2016; Zheng & Yang, 2016). However, W-MML has the tendency to induce more outliers under the small sample size condition. The current results showed that the difference between the two sample size conditions was almost unnoticeable, which support the

argument that a sample size of 500 is sufficient for accurate estimates of item parameters for the 2PL model (Baker, 1998).

There are some limitations to the current study. First, the study only focused on the 2PL model. The weighted calibration method could be generalized to other IRT models, such as the graded response model for polytomous items (Samejima, 1970), or the three-parameter logistic model (3PL) to account for guessing behaviors. Second, smaller sample sizes of other magnitudes could be explored to compare the performance of the two calibration methods. This is a meaningful direction because the calibration is typically conducted with smaller sample size in practice (de la Torre & Hong, 2010). Third, the current study only provides point estimates for item parameters. Future research could address the estimates of standard errors for the weighted calibration method.

References

Arikan, S., Özer, F., Şeker, V., & Ertas, G. (2020). The Importance of Sample Weights and Plausible Values in Large-Scale Assessments. Eğitimde Ve Psikolojide Ölçme Ve Değerlendirme Dergisi, 11(1), 522–539. https://doi.org/10.21031/epod.602765

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics. Theory and Methods*, *35*(3), 439–460. https://doi.org/10.1080/03610920500476598

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling approach. *Applied Psychological Measurement*, *22*, 153-169. https://doi.org/10.1177/01466216980222005

Ban, J.-C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest-item calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, *38*, 191–212. https://doi.org/10.1111/j.1745-3984.2001.tb01123.x

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In *Handbook of modern item response theory* (pp. 433-448). Springer, New York, NY.

Brewer, K., & Hanif, M. (1983). *Sampling with unequal probabilities, lecture notes in statistics (Vol. 15)*. New York: Springer-Verlag.

Cai, L., Albano, A. D., & Roussos, L. A. (2021). An investigation of item calibration methods in multistage testing. *Measurement: Interdisciplinary Research and Perspectives*, *19*(3), 163–178. https://doi.org/10.1080/15366367.2021.1878778

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology, 9*(1), 49–49. https://doi.org/10.1186/1471-2288-9-49

Chen, Y., von Davier, M., Weng, H., & Xie, Z. (2022). *Variable selection in latent regression IRT models via Knockoffs: An application to international large-scale assessment in education*. ArXiv. https://arxiv.org/pdf/2208.07959.pdf

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement, 34*(4), 267-285. https://doi.org/10.1177/0146621608329501

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory Item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24. https://doi.org/10.1177/0146621602026001001

Hu, H., Rogers, W. T., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement, 32*(4), 311–333. https://doi.org/10.1177/0146621606292215

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.https://doi.org/10.1111/j.1745-3984.2011.00161.x

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38,* 79–93. https://doi.org/10.1111/j.1745-3984.2001.tb01117.x

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43*, 355–381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

König, C., Spoden, C., & Frey, A. (2020). An optimized Bayesian hierarchical two-parameter logistic model for small-sample item calibration. *Applied Psychological Measurement, 44*(4), 311-326. https://doi.org/10.1177/0146621619893786

Laukaityte, I., & Wiberg, M. (2018). Importance of sampling weights in multilevel modeling of international large-scale assessment data. *Communications in Statistics. Theory and Methods, 47*(20), 4991–5012. https://doi.org/10.1080/03610926.2017.1383429

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: AddisonWesley

Loyd, B. H., & Hoover, H. D. (1980). Vertical Equating Using the Rasch Model. *Journal of Educational Measurement, 17*(3), 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Meinck, S. (2020). Sampling, weighting, and variance estimation. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment* (pp. 113-129). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-53081-5

Mislevy, R. J., & Bock, R. D. (1985). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 189–202). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory

Muthén, B., & Muthén, L. (2017). Mplus. In *Handbook of item response theory* (pp. 507-518). Chapman and Hall/CRC.

Oliveri, M. E., & Von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*(1), 1-21. https://doi.org/10.1080/15305058.2013.825265

Organisation for Economic Co-operation and Development. (2018). *PISA 2018 technical report.* Paris, France: Author. Retrieved from https://www.oecd.org/pisa/data/pisa2018technicalreport

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*(1), 23–40. https://doi.org/10.1111/1467-9868.00106

Qian, J., Jiang, Y., & von Davier, A. A. (2013). Weighting test samples in IRT linking and equating: Toward an improved sampling design for complex equating. *ETS Research Report Series, 2013*(2), i–31. https://doi.org/10.1002/j.2333-8504.2013.tb02346.x

Qian, J., von Davier, A. A., & Jiang, Y. (2013). Achieving a stable scale for an assessment with multiple forms: Weighting test samples in IRT linking. In *New Developments in Quantitative Psychology* (pp. 171-185). Springer, New York, NY.

Raudenbush S. W., Bryk A. S., Cheong Y. F., Congdon R. T., du Toit M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. Chicago, IL: Scientific Software International.

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In Rutkowski, L., von Davier, M. & Rutkowski. D. (eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 117–153). Boca Raton, FL: Chapman and Hall/CRC.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Samejima, F. (1970). Erratum estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *35*(1), 139–139. https://doi.org/10.1007/BF02290599

Smits, N. (2016). On the effect of adding clinical samples to validation studies of patient-reported outcome item banks: a simulation study. *Quality of Life Research*, *25*(7), 1635–1644. https://doi.org/10.1007/s11136-015-1199-9

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied psychological measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Suh, Y., Cho, S. J., & Wollack, J. A. (2012). A comparison of item calibration procedures in the presence of test speededness. *Journal of Educational Measurement*, *49*(3), 285-311. https://doi.org/10.1111/j.1745-3984.2012.00176.x

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education*, *42*(5), 517–540. https://doi.org/10.1023/A:1011098109834

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, *28*(6), 389-406. https://doi.org/10.1177/0146621604268734

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000-2012. Assessment in Education : Principles, Policy & Practice, 26(4), 466–488. https://doi.org/10.1080/0969594X.2019.1586642

Wang, C., Chen, P., & Jiang, S. (2020). Item calibration methods with multiple subscale multistage testing. *Journal of Educational Measurement*, *57*(1), 3-28. https://doi.org/10.1111/jedm.12241

Zheng, X., & Yang, J. S. (2016). Using sample weights in item response data analysis under complex sample designs. *Quantitative Psychology Research*, 123–137. https://doi.org/10.1007/978-3-319-38759-8_10

Zheng, X., & Yang, J. S. (2021). Multiple group item response theory applications Using Stata irt package. *Measurement: Interdisciplinary Research and Perspectives*, *19*(3), 190–198. https://doi.org/10.1080/15366367.2021.1911507