Research Memorandum
ETS RM–23-03

Impact of Per-Item Sample Size on Group-Score Proficiency Estimates

Nuo Xi
John Mazzeo

May 2023

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Impact of Per-Item Sample Size on Group-Score Proficiency Estimates

Nuo Xi and John Mazzeo
ETS, Princeton, New Jersey, United States

May 2023

Corresponding author: N. Xi, E-mail: nxi@ets.org

Suggested citation: Xi, N., & Mazzeo, J. (2023). *Impact of per-item sample size on group-score proficiency estimates* (Research Memorandum No. RM-23-03). ETS.

Action Editor: Usama Ali

Reviewers: John Donoghue and Tim Davey

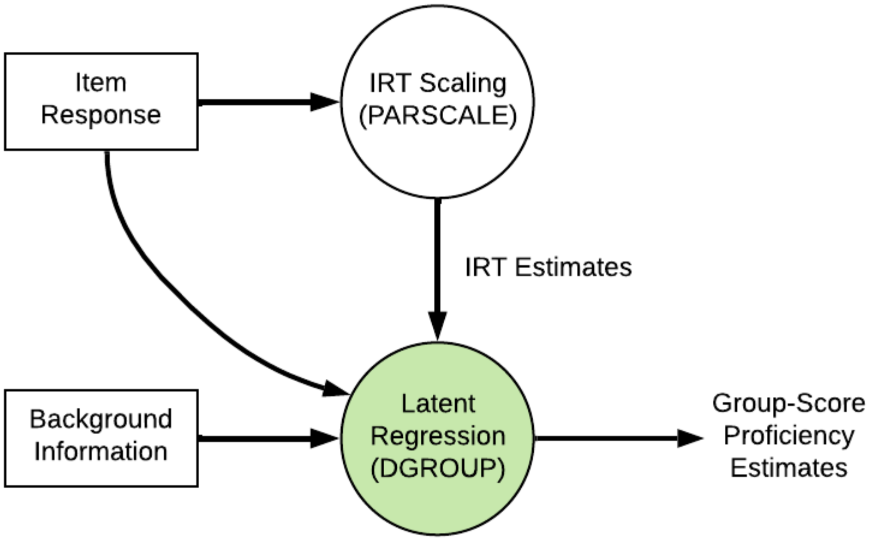
Abstract

The National Assessment of Educational Progress (NAEP) employs an item response theory (IRT)–based latent regression estimation procedure to recover group-score proficiency estimates. At this latent regression stage, item parameter estimates out of the IRT calibration stage are needed to compute the likelihood function, and the current practice is to take the item parameter estimates as fixed and known without considering the uncertainty associated with these estimates. To evaluate if and to what extent this practice introduces additional uncertainty in the estimation process, we conducted a series of simulation studies to calculate the bias, standard error, and root mean square error associated with the final group-score estimates. Results suggest that the amount of additional variation unaccounted for by the current practice is not very large compared to the estimates based on the true item parameter values. However, the estimation process is associated with some difficulty recovering the lower end of the proficiency distribution, especially when the sample size is small. A better understanding of how to accurately recover this portion of the distribution would be an important research direction for the program to follow.

Keywords: item response theory (IRT), IRT-based latent regression, calibration sample size, group-score estimates, proficiency estimates

In its operational practice, the National Assessment of Educational Progress (NAEP) program employs an item response theory (IRT)-based latent regression method (Mislevy et al., 1992) to estimate statistical properties of the overall population and subpopulations of interest. A high-level description of this estimation procedure is depicted in Figure 1. In this figure, the rectangles represent observed variables collected from the NAEP operations, such as students' responses to the assessed cognitive items and their background information. The circles represent the two estimation stages (i.e., the IRT scaling stage that is carried out in PARSCALE [Muraki & Bock, 1999] and the latent regression stage that is carried out in DGROUP [Rogers et al., 2006]). The IRT scaling stage characterizes the dependency of item responses on latent proficiencies and establishes a common scale across years upon which the latent proficiencies are expressed. The latent regression stage characterizes the dependency of latent proficiencies on a set of contextual variables extracted from the background information. Because the latent proficiencies are not directly observable at the latent regression stage, students' responses to the cognitive items are again involved as indicators of the latent proficiencies in the estimation process.

Figure 1. Item Response Theory–Based Latent Regression



Note. This item response theory (IRT)–based latent regression procedure was implemented in the National Assessment of Educational Progress (NAEP).

As indicated in Figure 1, the item parameter estimates out of the IRT scaling stage are needed in the latent regression stage, and the current practice is to take them as fixed and known in the later analysis stages where the latent regression parameters and the uncertainty of group-level results are estimated. As a result, the current operational estimation procedure does not account for the uncertainty associated with the item parameter estimates that could vary depending on the actual calibration sample size. Therefore, the standard error estimates of the reported group-score proficiency estimates are likely underestimated. With large sample sizes, such as the national and state combined assessments,[1] ignoring this additional uncertainty may not matter much. However, for national-only assessments and special studies, which cannot afford a very large calibration sample, it is necessary to obtain a better understanding of the effect of varying calibration sample size on the group-score estimates and the corresponding standard error estimates.

For national-only assessments, a minimum of 2,000 examinees per item is typically required by the NAEP assessment design guidelines. This convention was adopted in 1980s (Jia et al., 2010), and the required sample size was greater than the common guideline of 1,000 examinees per item required to get stable estimates of item parameters for three-parameter logistic (3PL) models (Hambleton, 1989; Hulin et al., 1982; Lord, 1968). Such guidelines (i.e., a minimum of between 1,000 and 2,000 examinees per item) were mainly established for accurate recovery of item parameters for 3PL models and often evaluated with a focus on reporting individual scores. The relationship between IRT calibration sample size and the amount of error in item parameter estimates, as well as how that error adds to the uncertainty in group-score statistics, has not been as thoroughly investigated.

For special studies, the sample size is frequently constrained by the budget and objectives of the study. As a result, the study sample size may not reach the NAEP operational guideline of 2,000 students per item. To investigate whether it is computationally feasible to scale NAEP data with a per-item sample size smaller than 2,000, Jia et al. (2010) drew random subsamples from the operational NAEP data and re-estimated several group-score statistics. Results from this pseudosimulation[2] study indicated the possibility of reducing the minimum per-item sample size but did not quantify the impact in terms of the standard errors of group-score statistics. Given that the results from NAEP special studies often guide the future assessment design, it is important for the NAEP program to appropriately quantify different error components on the group-score proficiency estimates and establish more detailed guidelines on special study sample size.

The current report documents a series of simulations designed to examine how uncertainty in item parameter estimates due to smaller per-item sample size may impact group-score estimates and corresponding standard errors. To control potential confounding factors, we started with a simplified population model that specifies the relationship between the latent proficiency and three contextual variables. The regression coefficients of the latent proficiency on these three contextual variables were estimated from the operational NAEP data and treated as the true values to evaluate the simulation results. Starting with this simplified condition makes the evaluation more straightforward in comparison to the situation where the true population model is not known. Per-item sample size is controlled as an experimental factor, and the objective is to see how the standard errors of the group-score estimates are affected as this experimental factor varies.

The rest of the report will unfold as follows. The Simulation Design section describes implementation details of the simulations, and the Simulation Results section summarizes the main findings observed from these simulations. The Summary section concludes this study report and makes suggestions on future research directions.

Simulation Design

The current study consists of simulations in which student’s background information collected from the NAEP operational data sets were kept but their responses to the assigned operational items were simulated. This section will discuss design features implemented in these simulations.

Matrix Sampling Design

The subject domains in NAEP are specified broadly. To reduce the burden of any individual student selected to participate in the assessment while still ensuring the content coverage, the NAEP instrument is configured and administered through a matrix sampling design (Johnson, 1992; Sirotnik, 1974). That is, the cognitive items are selected from a large item pool to form relatively short test booklets, each of which satisfies certain content requirements and statistical specifications. The number of items in each booklet is kept relatively modest (usually between 20 and 40 items, depending on the subject), and each sampled student takes only one booklet so that the individual testing time is kept within available limits (usually 60 minutes). The collection of all the booklets will cover the complete item pool.

The current study followed the matrix sampling design implemented in the 2017 NAEP mathematics digitally based assessment (DBA) at Grade 8, which involved a total of 177 items after scaling.[3] These items were distributed in 10 blocks with no overlap of items across blocks, and the 10 blocks were later assembled into 50 booklets[4] through balanced incomplete block (BIB) spiraling (see Allen et al., 2001, p. 20). Sampled Grade 8 students were administered one of the 50 booklets, and they had 1 hour to complete the assigned booklet. To maintain this operational booklet structure, the current study first sampled students at the booklet level and then aggregated the 50 booklet samples to form a single study sample. This sampling process ensured the same booklet structure was used in both the operational assessment and the simulations, regardless of the per-item sample size.

Obtaining True Item Parameters and True Ability Values

The original 2017 NAEP mathematics DBA data contained about 144,000 records across the 50 booklets of consideration. Because three of the 50 booklets were identified as special forms that could be assigned to students who required certain accommodations, the sample distribution of these three booklets was slightly different from the overall sample. To facilitate the booklet-level sampling procedure (which will be described in detail below), we removed some of the accommodated records from these three booklets to make them more representative of the overall sample. As a result, the total sample size was decreased to around 140,000. The reduced operational sample was taken as the student population in this study, from which samples of varying sizes were drawn.

The true population model was defined as a mixture of seven subpopulations based on three contextual variables (i.e., students' ethnicity, whether they were students with disabilities

[SDs], whether they were English language learners [ELLs]). That is, based on the background information collected in the 2017 operations, students were categorized into one of the following seven mutually exclusive subgroups:

- Test takers who were SDs but not ELLs, regardless of their race/ethnicity
- Test takers who were ELLs regardless of their race/ethnicity and SD status
- Five race/ethnicity subgroups of test takers who were neither SDs nor ELLs
    - White
    - Black
    - Hispanic
    - Asian
    - Other

These seven subgroups represent NAEP subgroups of different proficiency profiles. Note that the definition of these seven subgroups, especially the mutual exclusivity among them, is a deviation from operational analysis and simplifies the model structure implemented in the following analyses.

There was an additional modification on the 177 operational items used in this study: We collapsed[5] polytomous items down to binary items and used two-parameter logistic (2PL) models to estimate these modified items. This modification limited the analysis to 2PL and 3PL models and made the analysis a bit simpler. After these collapses, there were 104 3PL items and 73 2PL items. Besides, we assumed the 177 items all load on a single latent construct, simplifying the latent structure to be unidimensional.

Using the modified operational data (i.e., total sample size being reduced to around 140,000 and polytomous items being collapsed down to binary items), we conducted the IRT calibration in PARSCALE with a multigroup setup. Each student had a subgroup number as defined by the true population model specified above, and in total there were seven subgroups. Under the multigroup setup, PARSCALE estimated the mean and standard deviation for each of the seven subgroups, while the overall mean and standard deviation were constrained to be 0 and 1, respectively. Item parameter estimates obtained from this IRT calibration run were saved and used as the true item parameter values in the simulations. Table 1 summarizes the true *a*-, *b*-, and *c*-parameter values of these 177 items by item type.

Table 1. Summary of True Item Parameters, by Item Type

		2PL items		
Item type	N	Mean (*SD*)	Min	Max
a	73	0.79 (0.24)	0.28	1.45
b	73	0.20 (1.41)	-3.40	4.44
		3PL items		
a	104	1.04 (0.32)	0.25	2.27
b	104	0.53 (0.73)	-1.79	2.29
c	104	0.17 (0.08)	0.04	0.49

Note. 2PL = two-parameter logistic model; 3PL = three-parameter logistic model; *SD* = standard deviation; Min = minimum; Max = maximum.

The computer program, PARSCALE, estimates item parameters by marginal maximum likelihood estimation procedure (Bock & Aitkin, 1981) and uses the expectation-maximization (EM) algorithm to solve the maximum likelihood equations. Prior distributions are imposed on item parameters with the following starting values: *b*-parameter, N(0, 2^2); *a*-parameter, lognormal(0, 0.5^2); and *c*-parameter, beta(11, 41).[6] The locations of the prior distributions (but not the dispersion) are updated after each estimation cycle to reflect changes in the provisional estimates of the item parameters (Allen et al., 2001, Chapter 12). Omitted responses are treated as fractionally correct, and not-reached responses are treated as not presented in the calibration process.

Then, for the seven subgroups, we conducted an intercept-only latent regression run (using DGROUP[7]) for each subgroup separately. Under the NAEP context, the latent regression model is specified as (Allen et al., 2001, Chapter 12),

$$\boldsymbol{\theta} = \Gamma' \boldsymbol{y} + \boldsymbol{\epsilon}\,, \tag{1}$$

where $\boldsymbol{\theta}$ denotes a vector of latent proficiencies, $\boldsymbol{y}$ denotes a vector of contextual variables collected from the NAEP operations, Γ is the matrix consisting of the regression coefficients, and $\boldsymbol{\epsilon}$ represents the error terms that are assumed to be multivariate normally distributed with mean zero and variance-covariance matrix Σ. In the current study, we assumed univariate latent proficiency and conducted an intercept-only latent regression run for each subgroup. This means that the latent regression model (see Equation 1) was applied to each subgroup separately and a single intercept term was estimated for each subgroup. A separate variance term was estimated for each subgroup, which is different from the operational setup that assumes homogeneous variance across different subgroups. This intercept-only latent regression model was consistent with the multigroup setup used in the IRT calibration stage but much simplified from the operational model.

As discussed before, the latent regression stage characterizes the dependency of latent proficiencies on the contextual variables. Therefore, the selection of contextual variables to be included in the estimation process is a critical decision. The current study, however, focused on a best case scenario in which a much-simplified true population model is known and specified in the estimation process. This simplification eliminates the effect of model misspecification and focuses the investigation on per-item sample size involved at the IRT calibration stage.

The first set of plausible values[8] (PVs) obtained from these seven intercept-only latent regression runs were taken as the true ability values in the following simulations. Table 2 summarizes the true ability values by subgroups (on the θ scale established by the IRT calibration) and also lists the mean, standard deviation, and percentage of each subgroup at or above the three NAEP achievement levels.[9] Corresponding statistics of the student population (i.e., the combination of the seven subgroups) are listed on the last row of Table 2. We will examine the recovery of these distributional statistics in the Simulation Results section.

Table 2. Summary of True Ability Values, by Subgroup

Subgroup	Group size	Mean	SD	Basic	Prof	Adv
White	65,081 (46%)	0.374	0.842	85.2%	46.7%	13.4%
Black	20,547 (15%)	-0.477	0.839	50.9%	13.9%	2.0%
Hispanic	21,557 (15%)	-0.091	0.822	68.9%	25.8%	4.7%
Asian	6,489 (5%)	0.654	1.002	87.8%	58.3%	26.1%
Other	5,928 (4%)	0.028	0.908	72.3%	32.5%	8.1%
SD	12,505 (9%)	-0.899	0.990	33.4%	9.3%	1.7%
ELL	8,230 (6%)	-0.954	0.866	29.8%	5.6%	0.6%
Total	140,337	-0.015	0.996	69.4%	32.9%	9.0%

Note. SD = standard deviation; Prof = proficient; Adv = advanced; SD = student with disabilities; ELL = English language learner.

Per-Item Sample Size

Six levels of per-item sample size were considered in this study (i.e., 500, 1,000, 2,000, 4,000, 10,000, and Everyone).[10] At each per-item sample size level, students were sampled from the population and their responses were simulated using the true ability values and the "true" item parameter values. The simulated item responses together with the three contextual variables retained from the 2017 operation were used to estimate item parameters (i.e., IRT calibration stage) and corresponding group-score proficiencies (i.e., latent regression stage). By repeating this process 500 times, we obtained an estimate of the uncertainty/variation in the group-score estimates at each per-item sample size level.

To help interpret the observed uncertainty, we included another estimation approach that does not involve the IRT calibration stage. That is, the simulated data were directly fitted to the latent regression stage while using the true item parameter values. This alternative estimation approach corresponds to the current practice that assumes the item parameters are known and ignores the additional uncertainty associated with the calibration process. By comparing the uncertainty estimates associated with this alternative approach that does not involve IRT calibration to the approach that involves both IRT calibration and latent regression, we could quantify the additional uncertainty that is due to using a finite calibration sample. We expected these two approaches (i.e., the one using true item parameter values and the other using estimated item parameters) to generate close estimation results at a large per-item sample size level. But at a small, more realistic per-item sample size, the approach of using the estimated item parameters was expected to be associated with larger uncertainty. The simulation process is summarized in the next subsection.

Simulation Steps

Step 1. Draw a stratified random sample[11] with replacement from the student population at one of the six specified per-item sample sizes. The stratification variable is "booklet," and the sampling process was carried out in this way to ensure a data structure mimicking the operational BIB design.

Step 2. Simulate sampled students' responses to the assigned items using their true ability values and the true item parameter values.

Step 3. Using the data generated from Step 2, conduct IRT calibration in PARSCALE with the previously described multigroup setup.

Step 4. Transform the item parameter estimates from Step 3 to the common metric defined by the student population. Transformation detail is specified after the description of the simulation steps.

Step 5. Using the data generated from Step 2 and the transformed item parameter estimates from Step 4, conduct seven latent regression runs with an intercept-only model, one for each subgroup. Conduct another set of latent regression runs with the data generated from Step 2 and the true item parameter values. The first set of latent regression runs generates estimation results based on the estimated item parameters, and the second set generates estimation results based on the true item parameter values. The same sample is used in these two sets of latent regression runs.

Repeat Steps 1–5 500 times at each per-item sample size.

To resolve the indeterminacy in the θ scale, PARSCALE standardizes the item parameter estimates by setting the mean and standard deviation of the estimated θ scale to be 0 and 1, respectively. This standardization convention results in arbitrarily different scales for a single test if the test is calibrated with two different samples (Mislevy & Stocking, 1989). Step 4 was therefore implemented to convert the IRT scaling results across different replications to a common metric—in this case the metric on which the true item parameter values are expressed. A linear transformation based on lining up the means of the *a*- and *b*-parameters was employed,

$$\mathrm{A} = \frac{\exp\,(\mu(\log(\hat{a})))}{\exp\,(\mu(\log(a)))}, \tag{2}$$

$$\mathrm{B} = \mu(b) - A\mu(\hat{b}), \tag{3}$$

where $\mu(\log(a))$ and $\mu(b)$ represent the mean of the logarithm of the true *a*-parameters and the mean of the true *b*-parameters over the 177 items, and $\mu(\log(\hat{a}))$ and $\mu(\hat{b})$ represent the mean of the logarithm of the estimated *a*-parameters and the mean of the estimated *b*-parameters in a given replication. The transformed *a*- and *b*-parameters for this replication are

$$\hat{a}^{*} = \frac{\hat{a}}{A} \text{ and} \tag{4}$$

$$\hat{b}^{*} = A\hat{b} + B. \tag{5}$$

Equation 2, albeit different from the transformation constant A used in the usual mean/mean method (Kolen & Brennan, 2004), lines up the mean of *a*-parameters and results in more stable results than the mean/sigma method that uses the mean and standard deviation of the *b*-parameter estimates (T. C. Davey, personal communication, October 7, 2020).

Evaluation Criteria

The original objective of this study was to evaluate the uncertainty due to using a finite sample at the IRT calibration stage and to investigate how this uncertainty propagates through the latent regression stage and affects the uncertainty of final group-score estimates. The evaluation focus was on the change in standard errors of these group-score estimates as we varied the per-item sample size. However, the simulation results suggested that the size of the calibration sample not only affects the variability of the group-score estimates but also their accuracy. Based on this observation, we also considered estimation accuracy as one of the evaluation criteria.

Let *t* represent the population value[12] of a group-score statistic. At each per-item sample size, we computed simulation-based estimates of bias, standard error, and root mean square error to evaluate the accuracy and precision of an estimator of *t*

$$\text{Bias} = \frac{1}{500}\sum_{i=1}^{500}\hat{t}_i - t, \tag{6}$$

$$SE = \sqrt{\frac{1}{499}\sum_{i=1}^{500}(\hat{t}_i - \bar{t})^2}, \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{499}\sum_{i=1}^{500}(\hat{t}_i - t)^2}, \tag{8}$$

where $\hat{t}_i$ represents the estimate of *t* from the *i*th replication and $\bar{t}$ represents the average of $\hat{t}_i$ over the 500 replications at a specific per-item sample size. By following the NAEP operational procedure, the estimator $\hat{t}_i$ is an average of the estimated statistics over 20 sets of PVs. For more details on the PV methodology, refer to Allen et al. (2001, Chapter 12). Also notice that for a group-score assessment like NAEP, the bias of an estimator is usually more impactful in comparison to its variability, as the bias affects the location of the group proficiency estimates and could interact in complicated ways with the linking process.

Simulation Results

The focus of the current study is to compare two estimation approaches: one using the true item parameter values in the latent regression stage (IRT Pop) and the other using the estimated item parameters in the latent regression stage (IRT Est). To provide another reference point in the comparison, we also included the group-score statistics estimates based on each sample's true proficiency values and labeled it "Sample." That is, after a sample of students was drawn at each replication, the sample estimator simply calculated the group-score proficiency estimates such as means and percentiles based on this sample's true ability values without going through any estimation. In comparison, IRT Pop and IRT Est involved an estimation process using the simulated responses from this same sample. IRT Pop involved one additional estimation step (estimating the latent regression model using true item parameter values), and IRT Est involved two additional estimation steps (estimating the IRT parameters and estimating the latent regression parameters using the item parameter estimates). Therefore, including the sampled value as a reference point may help isolate the uncertainty due to the sampling of students in the simulation process from the additional uncertainty that could be attributed to the estimation processes, which is the focus of the current study. A summary of the definition of IRT Pop, IRT Est, and Sample, together with the population value, is provided in Table 3.

The following subsections present groups of three subplots for each group-score statistics of interest:

- comparison on the bias (top subplot),
- comparison on the standard error (middle subplot), and
- comparison on the RMSE (bottom subplot).

Table 3. Comparison of the Three Estimators Discussed in This Study, Plus Population Value

Estimator	Sampling	IRT calibration	Latent regression	Summary
Population value	Not involved, based on population	Not involved	Not involved	No sampling, no estimation steps
Sample	Yes	Not involved	Not involved	Sampling, no estimation steps
IRT Pop	Yes	Not involved	Yes	Sampling, one estimation step
IRT Est	Yes	Yes	Yes	Sampling, two estimation steps

Note. IRT = item response theory; IRT Pop = IRT population; IRT Est = IRT estimate.

In each subplot in the figures that follow (i.e., Bias, *SE*, or RMSE), results corresponding to the six levels of per-item sample size are presented, and "PIN" at the top indicates the per-item sample size. The *x*-axis labels the subgroups under comparison. The estimator based on true item parameter values is labeled as "IRT Pop" and marked with a red x, the estimator based on estimated item parameters is labeled as "IRT Est" and marked with a solid blue circle, and the sampled value is labeled as "Sample" and marked with a black triangle. The black horizontal line marks the value of zero on each plot, indicating no difference from the population value. The population values can be found in Table 2. The focus is to compare the accuracy and precision of IRT Pop and IRT Est at various per-item sample sizes. A brief description of the observations from each figure is provided in the next section. Observed patterns consistent with already discussed observations will not be repeated.

Recovery of Population/Subgroup Means

Bias Subplot

In general, IRT Pop is an unbiased estimator of the student population (marked as "Total" in the subplots) and subgroup means, even at small per-item sample sizes (Figure 2). In contrast, IRT Est overestimates the student population and subgroup means at small per-item sample sizes and behaves more like the IRT Pop estimator as the per-item sample size increases. When the per-item sample size is 500, on average, the IRT Est estimator overestimates the population mean by an amount close to 3% of its standard deviation and overestimates the ELL mean by an amount close to 6% of its standard deviation. Observing this bias due to using estimated item parameters, we conducted some further investigation on the IRT parameter estimates, which is summarized in the Recovery of IRT Parameters subsection.

SE Subplot

As expected, for both IRT Pop and IRT Est, the standard error decreases as the per-item sample size increases. The standard error of IRT Est is very close to IRT Pop, and both are only slightly larger than the standard error of Sample, even when the per-item sample size is small, suggesting that the additional variation due to the item parameter estimation process is not very concerning. Smaller subgroups, such as Asian, Other, SD, and ELL, are associated with somewhat larger additional variation compared to the larger subgroups.

RMSE Subplot

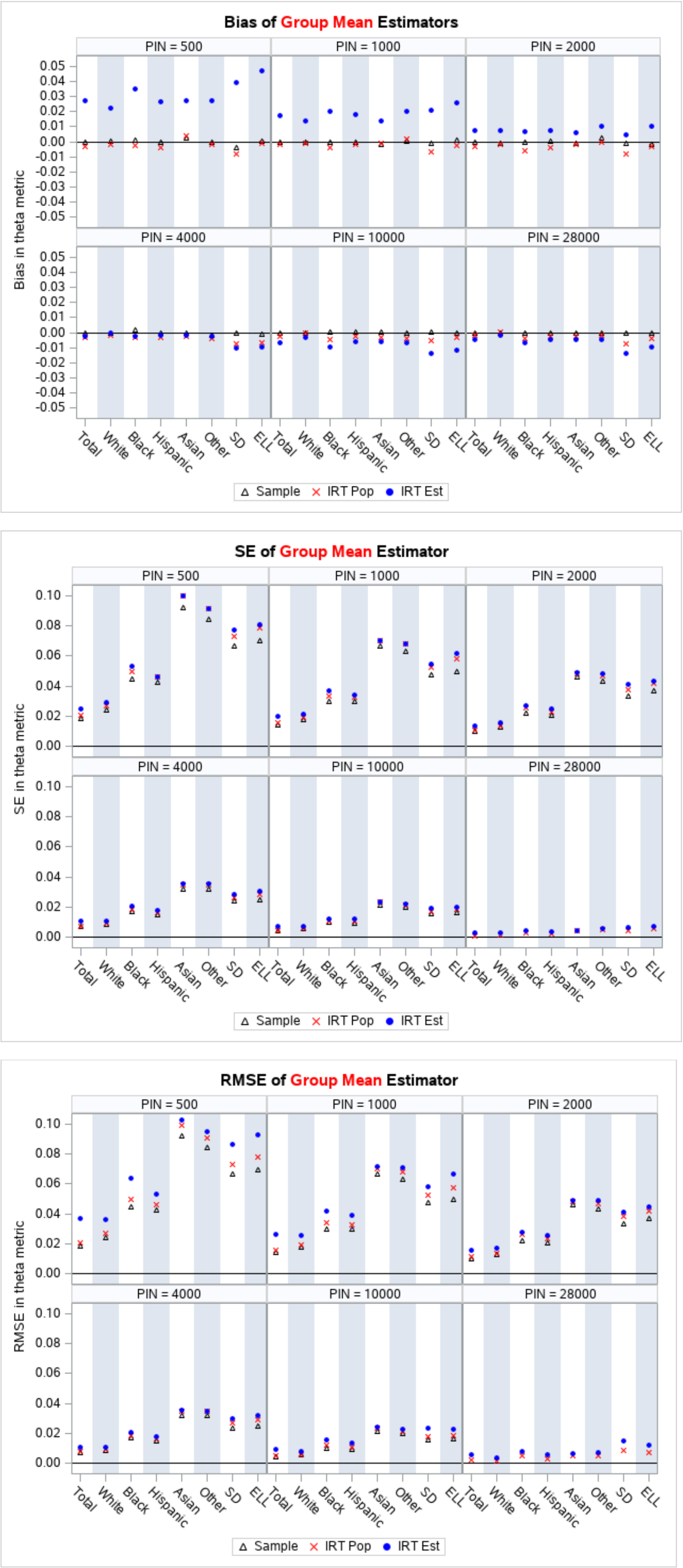
Similar to the SE subplot, for both IRT Pop and IRT Est, the RMSE decreases as the per-item sample size increases, and the estimates for smaller subgroups are associated with larger RMSE. Based on the observations from the Bias and SE subplots, the difference in RMSE between IRT Pop and IRT Est can be largely attributed to the additional bias due to using item parameter estimates at the latent regression stage.

Recovery of Population/Subgroup Standard Deviations

Bias Subplot

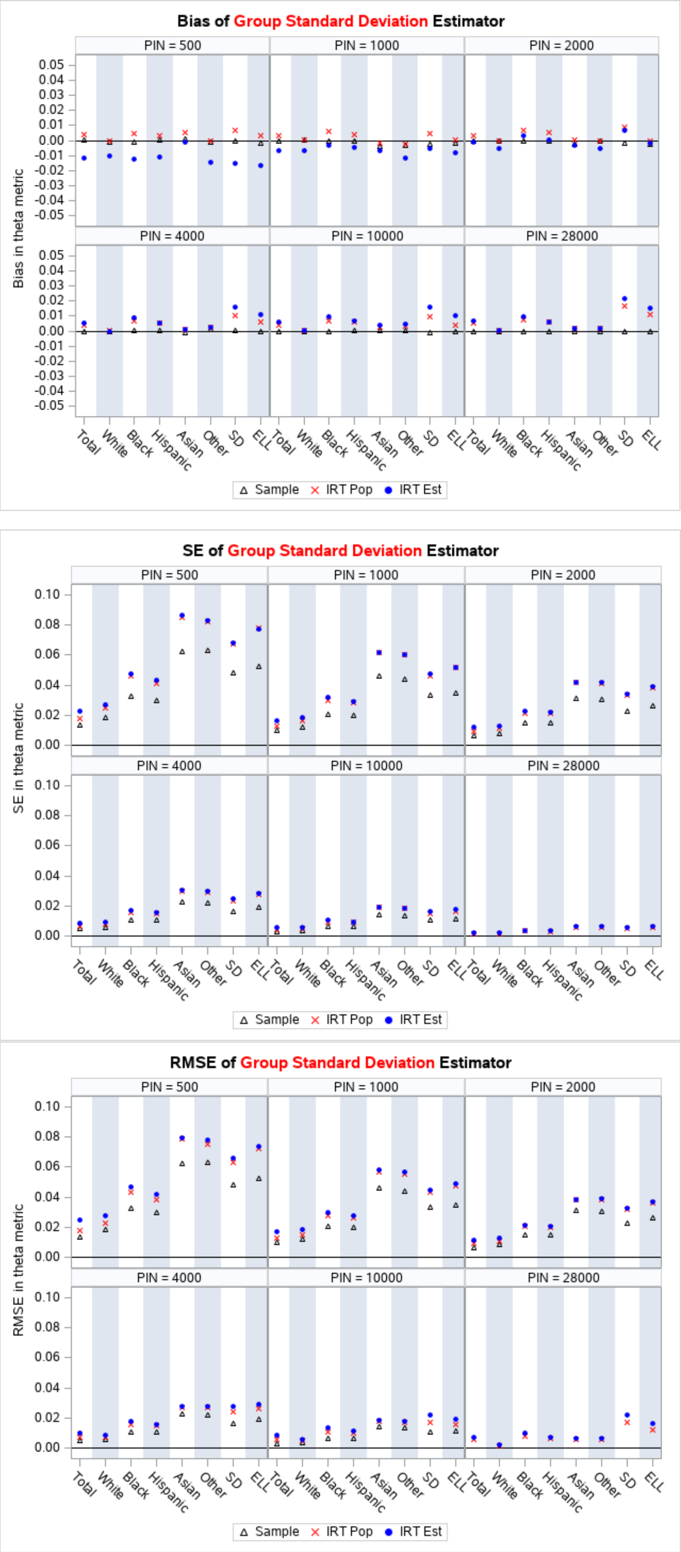
IRT Pop slightly overestimates lower performing subgroups' standard deviation (such as SD and ELL subgroups) even at very large per-item sample size. IRT Est changes from underestimate to overestimate and behaves more like IRT Pop as the per-item sample size increases (Figure 3).

Figure 2. Recovery of Population and Subgroup Means



Note. PIN = per-item sample size; IRT = item response theory; SD = students with disability; ELL = English language learner; RMSE = root mean square error; *SE* = standard error.

Figure 3. Recovery of Population and Subgroup Standard Deviations



Note. PIN = per-item sample size; IRT = item response theory; SD = students with disability; ELL = English language learner; RMSE = root mean square error; *SE* = standard error.

SE Subplot

The standard error of IRT Est is very close to that of IRT Pop, and both are greater than the standard error of Sample at smaller per-item sample size. The impact is more prominent for the smaller subgroups (i.e., Asian, Other, SD, and ELL) than the larger subgroups. The observation appears to suggest that there is additional variation due to the estimation process in general, but the additional uncertainty introduced is very small.

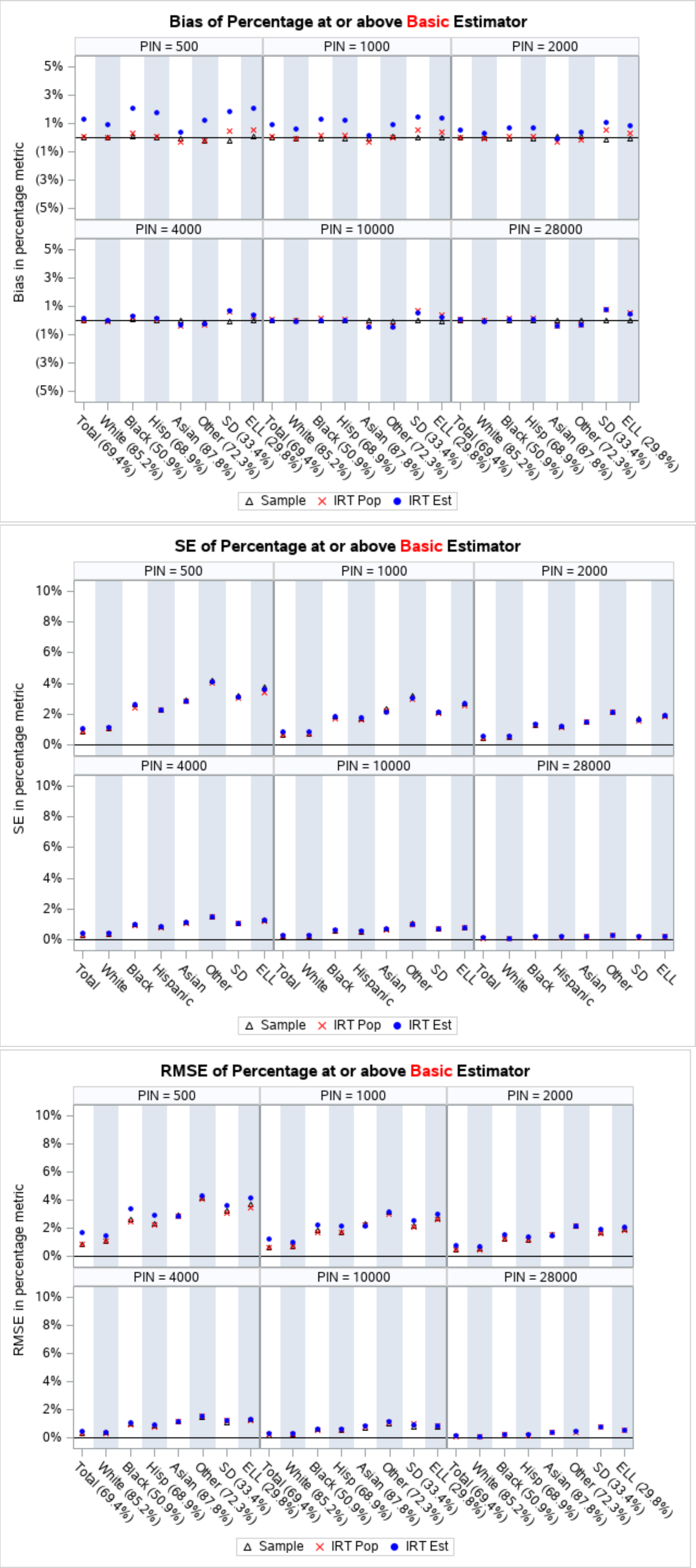
RMSE Subplot

Again, the difference in RMSE between IRT Pop and IRT Est can be largely attributed to the additional bias due to using estimated item parameters at the latent regression stage.

Recovery of Percentage at or Above Achievement Levels

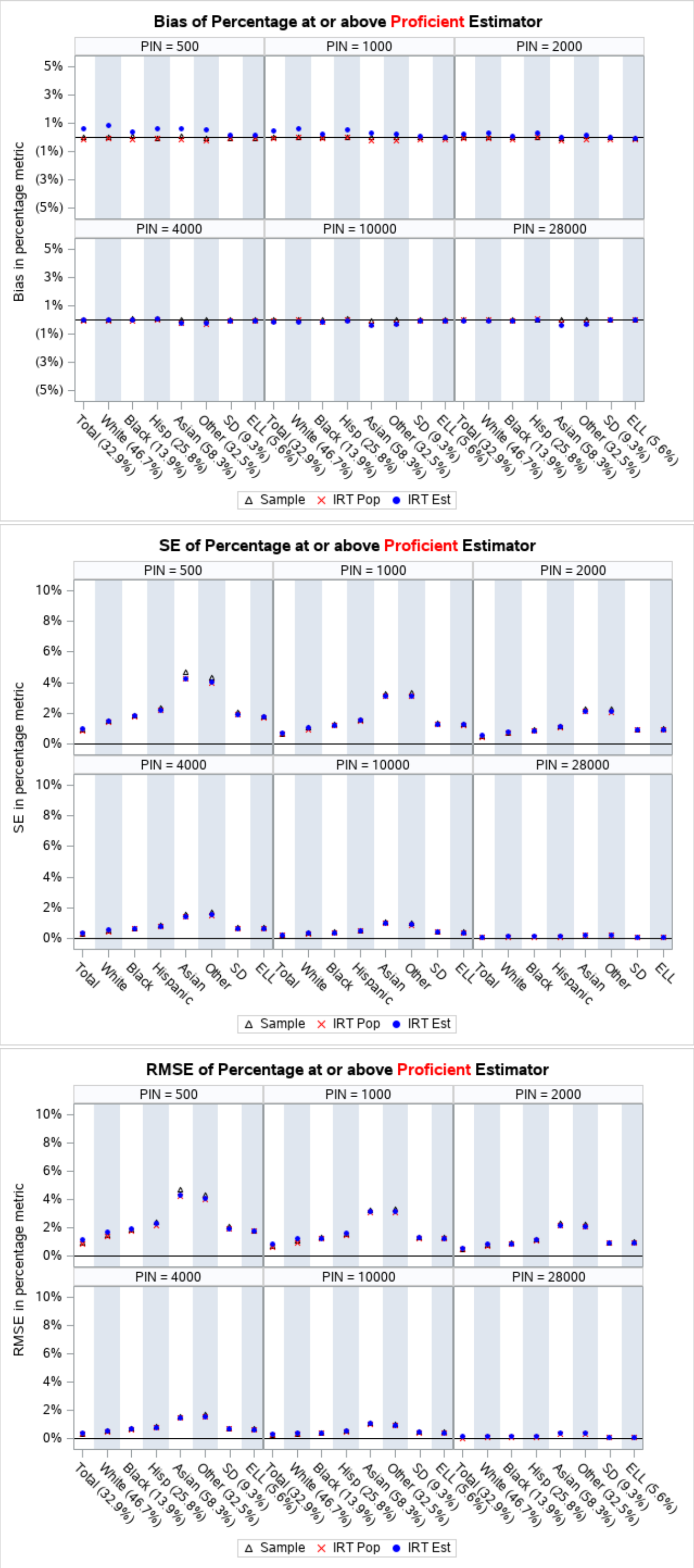
In these plots, the *x*-axis labels subgroup names, as well as the population value (in parenthesis) of the percentage at or above certain achievement level for each subgroup (Figures 4, 5, and 6).

Figure 4. Recovery of Percentage at or Above the Basic Achievement Level



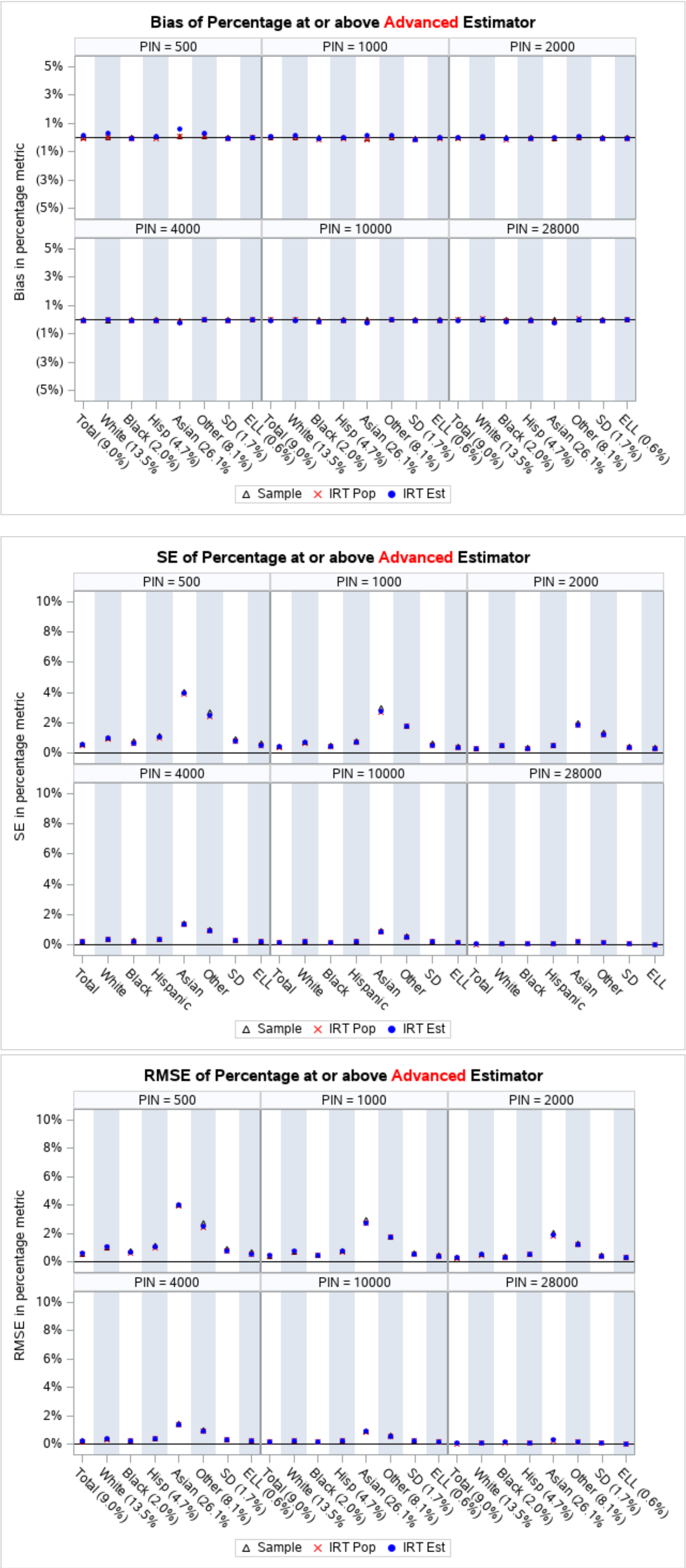
Note. PIN = per-item sample size; IRT = item response theory; SD = students with disability; ELL = English language learner; RMSE = root mean square error; *SE* = standard error.

Figure 5. Recovery of Percentage at or Above the Proficient Achievement Level



Note. PIN = per-item sample size; IRT = item response theory; SD = students with disability; ELL = English language learner; RMSE = root mean square error; *SE* = standard error.

Figure 6. Recovery of Percentage at or Above the Advanced Achievement Level



Note. PIN = per-item sample size; IRT = item response theory; SD = students with disability; ELL = English language learner; RMSE = root mean square error; *SE* = standard error.

Bias Subplot

In general, IRT Pop can unbiasedly recover the percentage at or above the three achievement levels for the population and subgroups. IRT Est has larger bias for the percentage at or above the basic achievement level when the per-item sample size is small, and the direction of the bias is consistent with the observation that IRT Est overestimates the population and subgroup means at smaller per-item sample size. However, IRT Est can unbiasedly recover the percentage at or above the other two achievement levels, suggesting that the lower performing area of the ability distribution is where the estimation bias happens.

SE Subplot

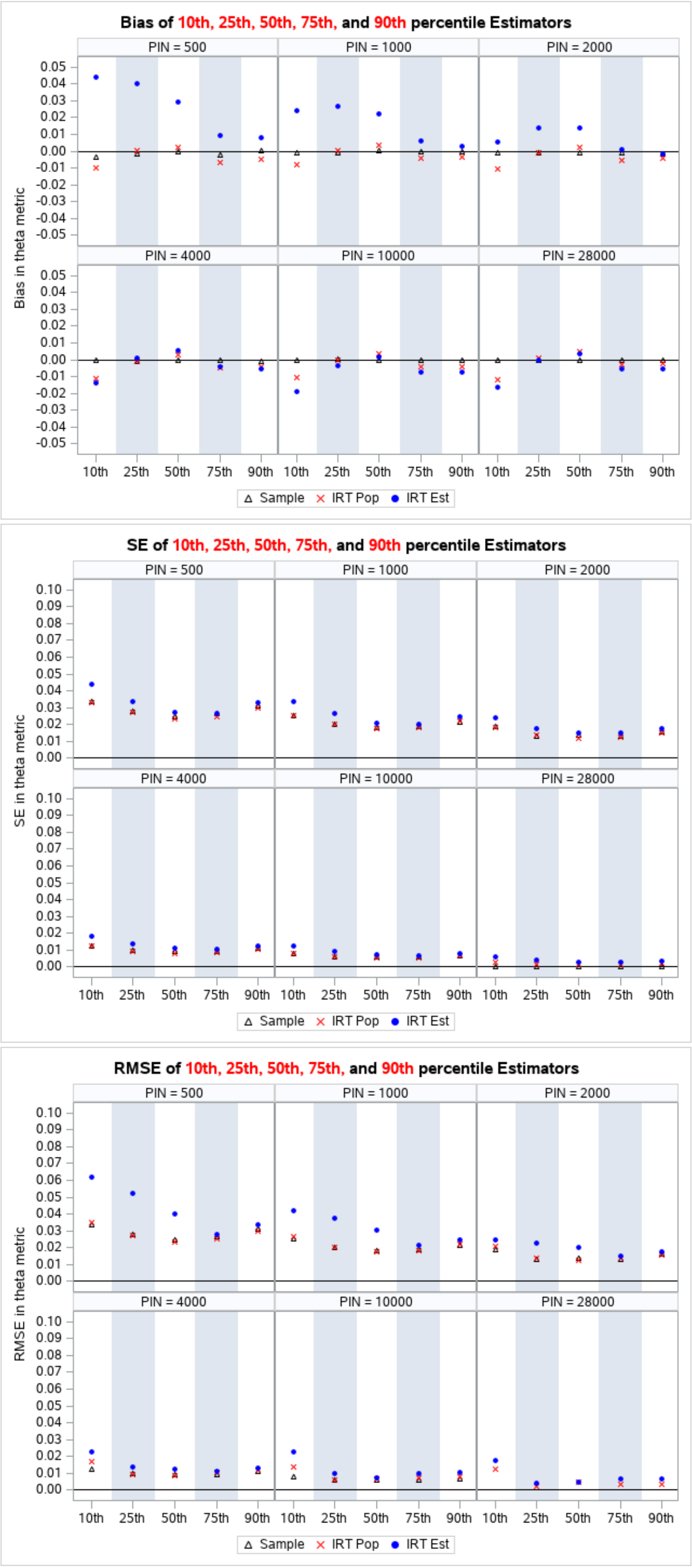
The standard error of IRT Est is very close to that of IRT Pop, and both are very close to the standard error of Sample, suggesting no additional variation on top of the variation due to sampling.

Recovery of Population’s 10th, 25th, 50th, 75th, and 90th Percentiles

In addition to the population and subgroup means, standard deviations, and percentages at or above the three achievement levels, we also inspected the recovery of the population’s 10th, 25th, 50th, 75th, and 90th percentiles (Figure 7). The results are also summarized in a set of three subplots—Bias, *SE*, and RMSE—with the *x*-axis indicating the percentage numbers associated with these five percentile estimates. The most interesting pattern is found in the Bias subplot. As it shows, IRT Est overestimates all five percentiles at smaller per-item sample size, but the bias is much smaller for the 75th and 90th percentiles. As the per-item sample size increases, IRT Est becomes less biased but still underestimates the 10th percentile. On the other hand, IRT Pop very slightly underestimates the 10th percentile, even at the largest per-item sample size (i.e., when everyone is used).

This observation, in combination with the prior observations, suggests that the estimation procedure has some difficulty in accurately locating the position of lower performing individuals. As this could be related to whether the *c*-parameter of the 3PL items were accurately recovered or not, we examined the recovery of the IRT parameters at the considered per-item sample size levels by item type (i.e., 3PL vs. 2PL).

Figure 7. Recovery of Population’s 10th, 25th, 50th, 75th, and 90th Percentiles



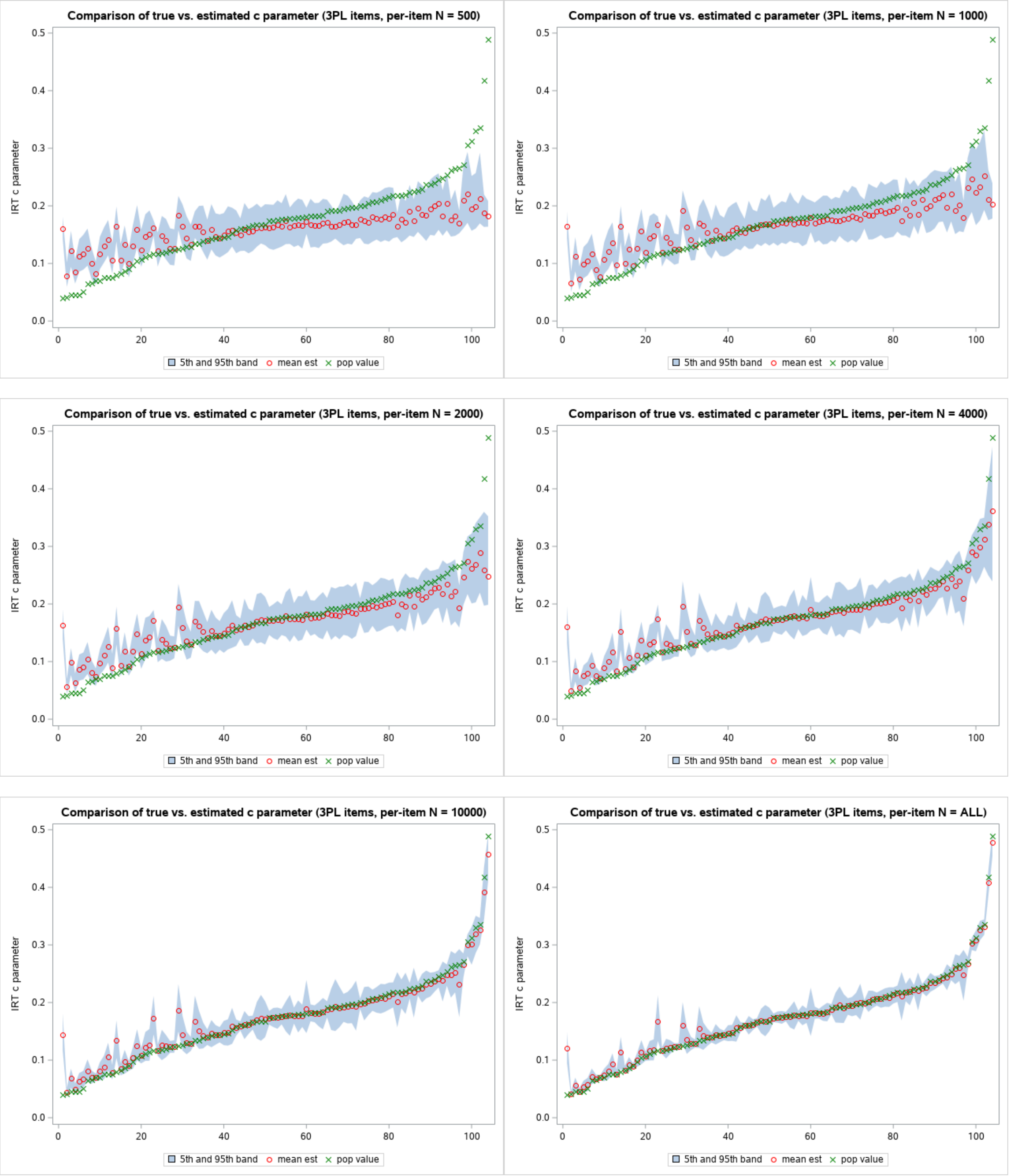
Note. PIN = per-item sample size; IRT = item response theory; RMSE = root mean square error; *SE* = standard error.

Recovery of IRT Parameters

Figure 8, Figure 9, and Figure 10 plot the true *c*-parameter, *b*-parameter, and *a*-parameter values against their estimates at the six per-item sample size levels under consideration for the 104 3PL items. Figure 11 and Figure 12 plot the true *b*-parameter and *a*-parameter values against their estimates for the 73 2PL items. In these plots, the green cross represents the true item parameter value, the red circle represents the average of the item parameter estimates over 500 replications, and the blue band shows the range between the 5th and 95th percentiles of these 500 estimates. The blue band can be seen as a 90% confidence interval around the mean estimate (i.e., the red circle). And the focus of the comparison is to see whether this blue band covers the true value represented by the green cross and to evaluate the difference between the true values (i.e., the green cross) and the estimates (i.e., the red circle). The items are ordered by the corresponding true parameter values from the smallest to the largest, and the numerical order is reflected by the *x*-axis.

Figure 8 shows that at a smaller per-item sample size level, the *c*-parameter estimates are less variable than the true values, as the red circles are varying around the posterior mean of about 0.17 and deviating from the green crosses. This is also confirmed by the standard deviation estimate of the *c*-parameters being smaller than its true value, which is shown in Table 5. Therefore, for items with more extreme true *c*-parameter values, there is larger estimation bias, and the 90% confidence interval (i.e., the blue band) might not cover the true values. As the per-item sample size increases, the *c*-parameter estimates (i.e., the red circle) move closer to their true values (i.e., the green cross), except for a few items that are associated with low *b*-parameters. It is known that for very easy items, it is difficult to estimate the pseudoguessing parameter very accurately, because there are very few observations low enough to provide good information to locate the *c*-parameter (Lord, 1968, 1983; Thissen & Wainer, 1982).

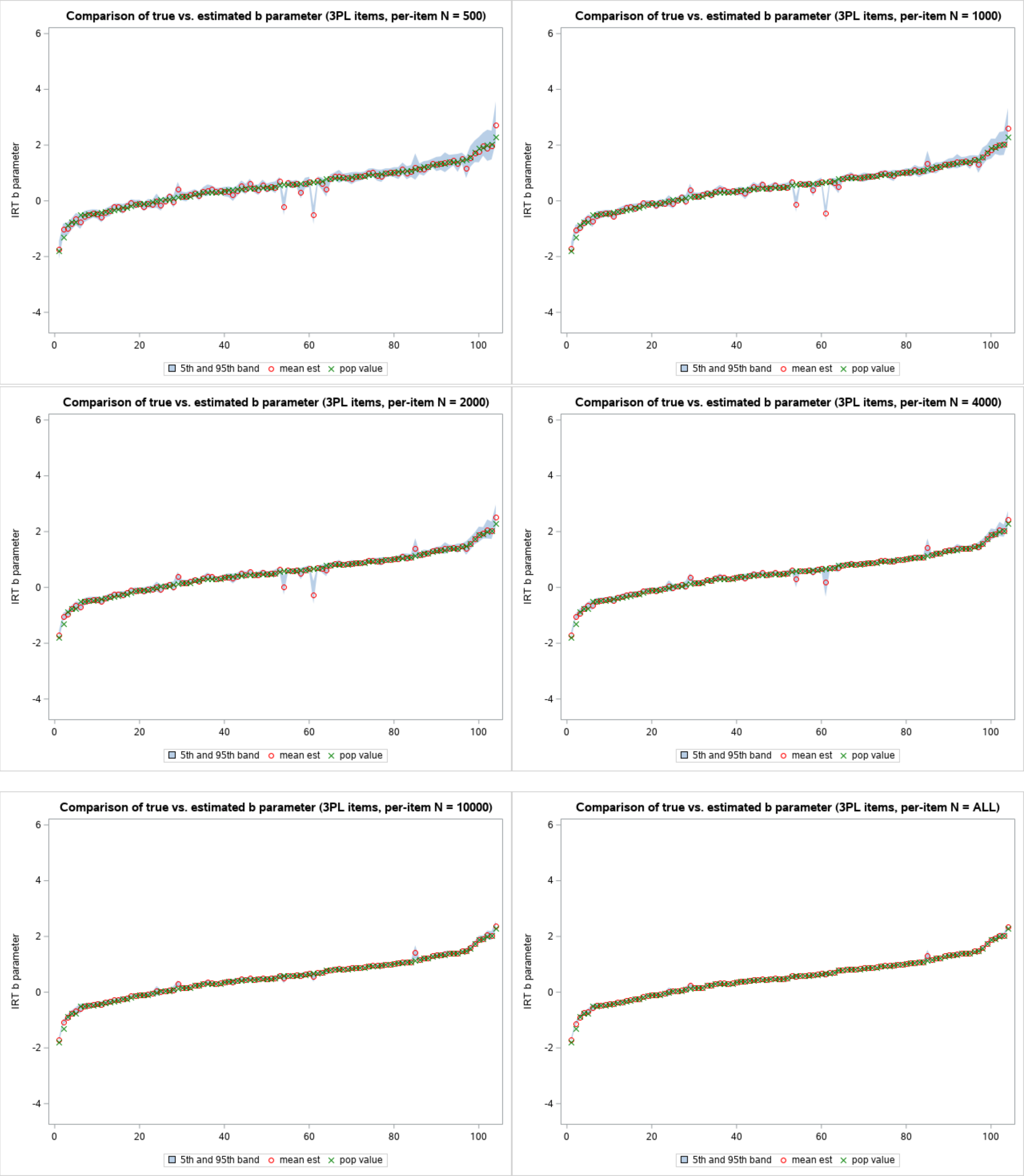
Figure 8. Recovery of the *c*-Parameter of the 104 Three-Parameter Logistic (3PL) Items



Note. IRT = item response theory; est = estimate; pop = population.

Figure 9 and Figure 10 show that for the 104 3PL items, the difference between the true item parameter values and the corresponding estimates decreases as the per-item sample size increases, but the magnitude of the difference is much more noticeable compared to the 73 2PL items (Figure 11 and Figure 12), especially for the *a*-parameter. The difficulty in estimating the *c*-parameter could impact the estimation of the corresponding *a*- and *b*-parameter when the per-item sample size is small. For future studies, it would be interesting to follow this direction and investigate how to improve the estimation process to obtain accurate *c*-parameters. Researchers (Baker, 1967; Lord, 1968; Lord & Wingersky, 1985) have suggested considering ability distributions with more weights on the two ends (e.g., uniform or U-shaped distribution) for parameter estimation purposes. Directly implementing a sampling procedure to generate such ability distributions might be difficult, but NAEP could consider more targeted testing in its future assessments to reduce guessing and nonresponses, which will increase the information collected from the cognitive items and help estimate the proficiency level of lower performing groups.

Figure 9. Recovery of the *b*-Parameter of the 104 Three-Parameter Logistic (3PL) Items



Comparison of true vs. estimated b parameter (3PL items, per-item N = 500)

Comparison of true vs. estimated b parameter (3PL items, per-item N = 1000)

Comparison of true vs. estimated b parameter (3PL items, per-item N = 2000)

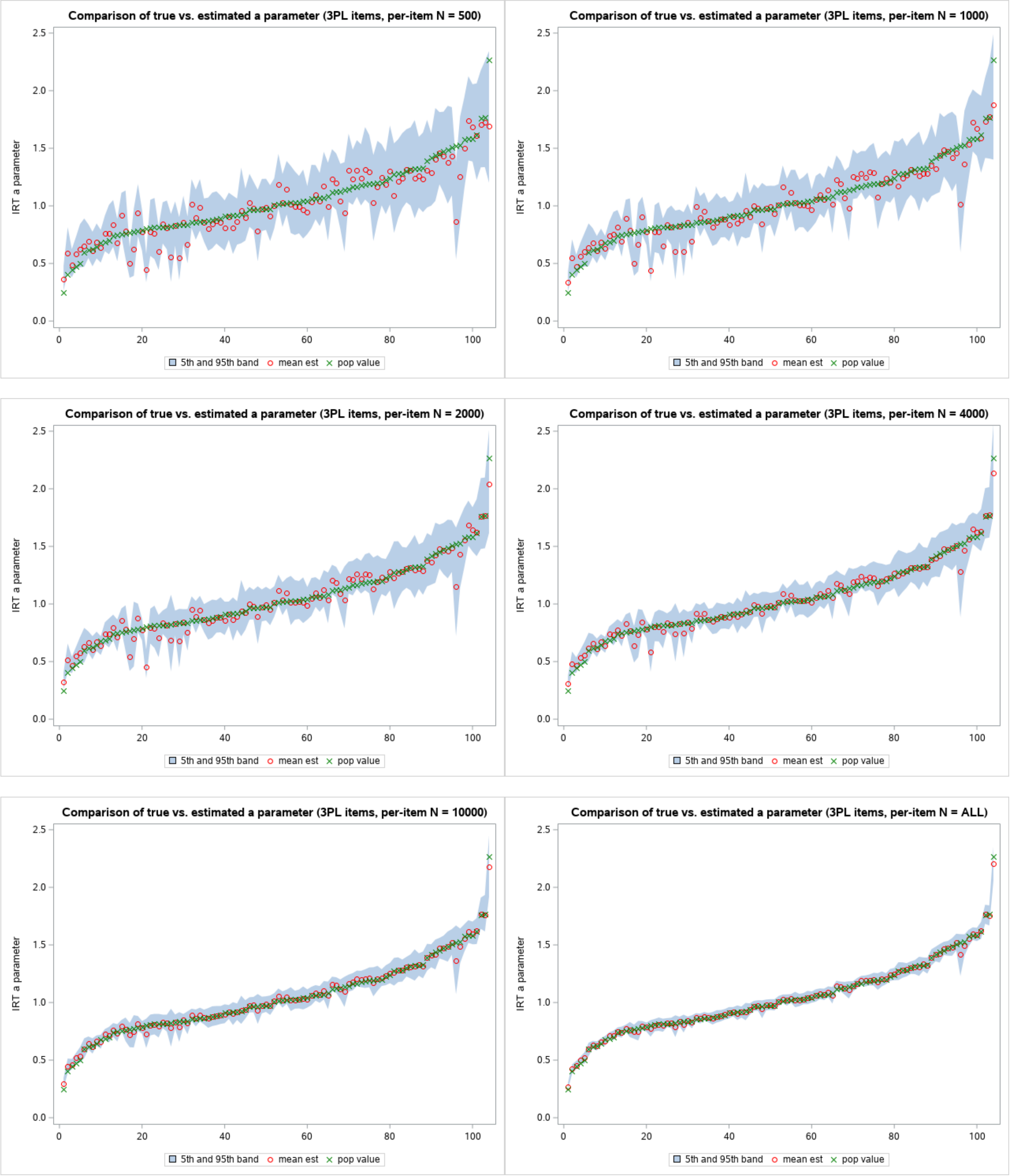
Comparison of true vs. estimated b parameter (3PL items, per-item N = 4000)

Comparison of true vs. estimated b parameter (3PL items, per-item N = 10000)

Comparison of true vs. estimated b parameter (3PL items, per-item N = ALL)

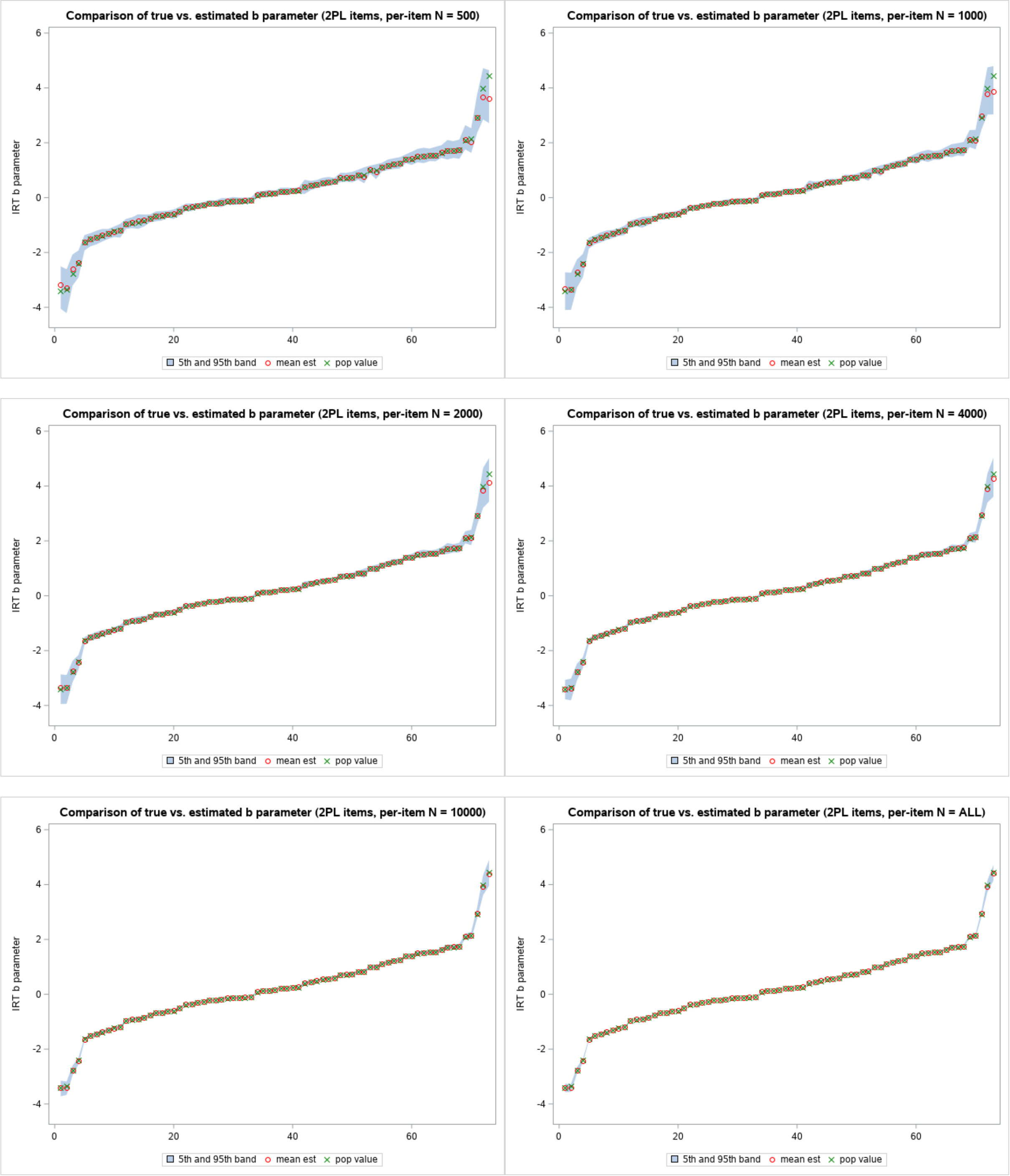
Note. IRT = item response theory; est = estimate; pop = population.

Figure 10. Recovery of the *a*-Parameter of the 104 Three-Parameter Logistic (3PL) Items



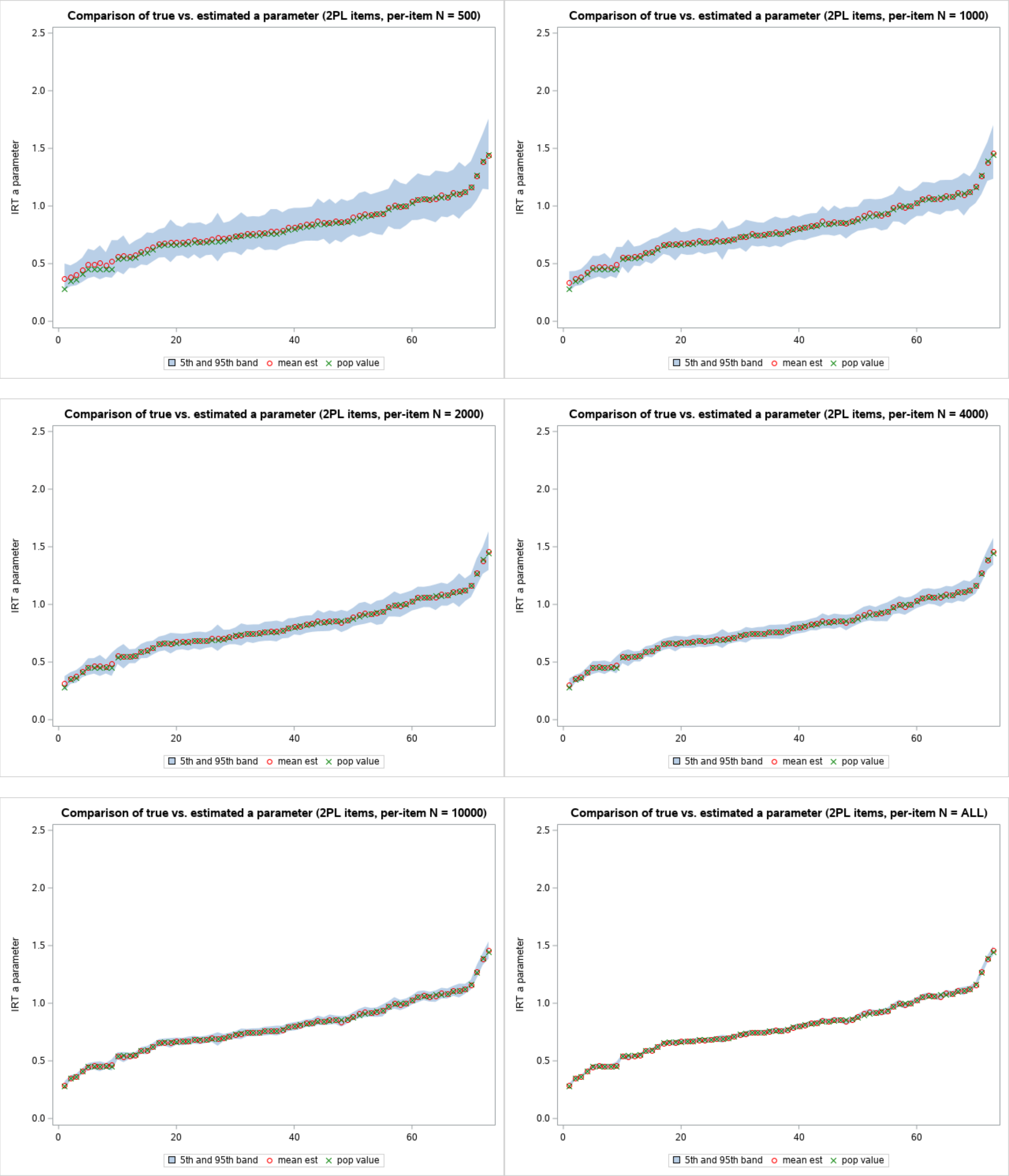
Note. IRT = item response theory; est = estimate; pop = population.

Figure 11. Recovery of the *b*-Parameter of the 73 Two-Parameter Logistic (2PL) Items



Note. IRT = item response theory; est = estimate; pop = population.

Figure 12. Recovery of the *a*-Parameter of the 73 Two-Parameter Logistic (2PL) Items



Note. IRT = item response theory; est = estimate; pop = population.

To help evaluate the comparison, Table 4 and Table 5 summarize the true values as well as the mean, standard deviation, minimum, and maximum of the item parameter estimates across the 500 replications at each per-item sample size level by item type.

Table 4. Summary of Two-Parameter Logistic (2PL) Item Parameter Estimates at Different per-Item Sample Size Levels

2PL Items (73 items)				
IRT parameter	Sample size	Mean (*SD*)	Min	Max
a	True value	0.79 (0.24)	0.28	1.45
	500	0.81 (0.25)	0.22	2.07
	1000	0.80 (0.24)	0.21	1.98
	2000	0.80 (0.24)	0.19	1.93
	4000	0.80 (0.24)	0.22	1.68
	10000	0.79 (0.24)	0.23	1.60
	Everyone	0.79 (0.24)	0.23	1.55
b	True value	0.20 (1.41)	-3.40	4.44
	500	0.20 (1.35)	-5.00	5.96
	1000	0.20 (1.38)	-5.64	5.93
	2000	0.20 (1.39)	-4.91	6.33
	4000	0.20 (1.40)	-4.26	5.80
	10000	0.20 (1.40)	-3.97	5.59
	Everyone	0.20 (1.41)	-3.72	5.20

Note. SD = standard deviation; Min = minimum; Max = maximum.

Table 5. Summary of Three-Parameter Logistic (3PL) Item Parameter Estimates at Different per-Item Sample Size Levels

3PL ITEMS (104 items)				
IRT parameter	Sample size	Mean (*SD*)	Min	Max
a	True value	1.04 (0.32)	0.25	2.27
	500	1.01 (0.35)	0.21	2.97
	1000	1.02 (0.34)	0.21	3.14
	2000	1.03 (0.34)	0.23	3.35
	4000	1.04 (0.33)	0.22	2.91
	10000	1.04 (0.32)	0.23	3.01
	Everyone	1.04 (0.32)	0.24	2.63
b	True value	0.53 (0.73)	-1.79	2.29
	500	0.51 (0.75)	-2.55	4.54
	1000	0.52 (0.75)	-2.17	4.95
	2000	0.53 (0.74)	-2.14	3.80
	4000	0.54 (0.73)	-1.97	3.39
	10000	0.54 (0.72)	-1.89	2.74
	Everyone	0.54 (0.72)	-1.84	2.54
c	True value	0.17 (0.08)	0.04	0.49
	500	0.16 (0.04)	0.05	0.47
	1000	0.16 (0.04)	0.04	0.42
	2000	0.17 (0.05)	0.03	0.52
	4000	0.17 (0.06)	0.03	0.52
	10000	0.17 (0.07)	0.03	0.52
	Everyone	0.17 (0.07)	0.03	0.52

Note. SD = standard deviation; Min = minimum; Max = maximum.

Summary

This report summarizes a simulation study conducted to evaluate the impact of varying calibration sample size on the group-score proficiency estimates. The study focus was on the uncertainty (i.e., the additional variation in the final estimates that could be attributed to having a finite calibration sample). This uncertainty is not accounted for in the current NAEP practice as the item parameter estimates out of the IRT scaling stage are taken as fixed and known in the latent regression estimation stage. The simulation results suggested that when the per-item sample size is small, using the estimated item parameters does not greatly impact the standard error. It is the bias, the difference between the true ability values and the estimates, that is more concerning. The estimation process seems to have difficulty in accurately recovering the lower end of the distribution.

An examination of the item parameter estimates shows that for 3PL items, when the per-item sample size is small, the *c*-parameter estimates are less variable and could be more biased for items with more extreme true *c*-parameter values. Compared to 2PL items, the *a*- and *b*-parameter estimates of the 3PL items are associated with larger bias at smaller per-item sample size levels. Further, for very easy items, even with very large per-item sample size, it could still be difficult to accurately estimate the pseudoguessing parameters.

Given that it is more challenging to accurately estimate the lower end of the proficiency distribution, the program should pay more attention to the lower performing subgroups at various estimation stages. The program could consider sampling plans that emphasize the lower end and collect more data in this area to improve the IRT estimates. Another possibility is to implement more targeted/customized instruments in the field to reduce guessing and nonresponses, which will help improve the quality of the proficiency estimates.

To extend the current study, we would suggest introducing the proportion of 3PL items as a factor in the simulations to study the effect of having more 3PL items in the instrument versus having fewer 3PL items. We conducted a preliminary follow-up study in this direction by creating two tests of the same length: one with 2PL items exclusively and the other with 3PL items exclusively. At a per-item sample size of 500, we observed that for the test with 3PL items exclusively its group-score estimates are associated with larger bias and larger standard error.

There are a few limitations of the current study. First, the current study was built upon a much-simplified population model that is very different from the models used in the NAEP operations. This simplification controls the confounding factors and helps establish a baseline for understanding the effects of varying calibration sample size on the group-score proficiency estimates. Future studies should consider introducing model complexity as an experimental factor. Second, in the current study, the results were estimated following the two-step procedure implemented in the NAEP operations (see Figure 1) and using PARSCALE for IRT calibration and DGROUP for latent regression estimation. Somewhat different results might be obtained if programs other than PARSCALE/DGROUP were used to carry out the two-step estimation process. Therefore, interpretation of the current study's results should be limited to the context of the NAEP operational estimation procedure.

Another limitation is that the current study did not consider the weight factor associated with the multistage sampling procedure. In NAEP operations, each sampled student carries a weight that reflects the probability of being selected by the sampling procedure. These weights are used in the IRT calibration stage, the latent regression stage, the computation of group-score statistics, and the corresponding standard errors. In addition, NAEP utilizes a replication method known as "jackknife" to estimate the variance due to sampling (Allen et al., 2001). To simplify the comparison, the current study did not use weights in the estimation process and did not apply the jackknife method to estimate the sampling variance. The conceptualization of standard error is slightly different in the current study than in the NAEP operations. These factors (i.e., applying weights and jackknife procedures) should be considered in the simulation and estimation process in the future.

To summarize, the program should consider better understanding the effect of having easy 3PL items in the instrument on the recovery of item parameters and proficiency estimates for low-performing subgroups. This investigation should be conducted under a simplified population model as the current study did. Once this effect is better gauged, the program could consider gradually introducing more complex models to help interpret the results in a context that is more similar to the NAEP operations.

References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report* (NCES 2001-509). U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/pdf/main1998/2001509.pdf

Baker, F. B. (1967). The effect of criterion score grouping upon item parameter estimation. *British Journal of Mathematical and Statistical Psychology*, *20*(2), 227–238. https://doi.org/10.1111/j.2044-8317.1967.tb00389.x

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/bf02293801

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (pp. 147–200). Macmillan Publishing; American Council on Education.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, *6*(3), 249–260. https://doi.org/10.1177/014662168200600301

Jia, Y., Moran, R., Qian, J., & Lin, M.-J. (2010). *Task 2.2.8.1.2: Reporting targets*. National Center for Education Statistics.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement*, *29*(2), 95–110. https://doi.org/10.1111/j.1745-3984.1992.tb00369.x

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practice*. Springer.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, *28*(4), 989–1020. https://doi.org/10.1177/001316446802800401

Lord, F. M. (1983). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika, 48*(3), 425–435. https://doi.org/10.1007/bf02293684

Lord, F. M., & Wingersky, M. S. (1985). Sampling variances and covariances of parameter estimates in item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 69–88). University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*(1), 57–75. https://doi.org/10.1177/014662168901300106

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Muraki, E., & Bock, R. D. (1999). *PARSCALE: Parameter scaling of rating data* [Computer software]. Scientific Software International.

Rogers, A. M., Tang, C., Lin, M.-J., & Kandathil, M. (2006). *DGROUP* [Computer software]. ETS.

Sirotnik, K. (1974). An introduction to matrix sampling for the practitioner. In J. W. Popham (Ed.), *Evaluation in education: Current applications* (pp. 453–529). McCutchen Publishing Company.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*(4), 397–412. https://doi.org/10.1007/bf02293705

Notes

[1] The NAEP reading and mathematics assessments at Grades 4 and 8 are usually national and state combined assessments, and the per-item sample size for these assessments can reach 20,000 students.

[2] In a pseudosimulation, data are not generated "from scratch." Instead, random samples are drawn from real testing data with some properties retained and some other properties simulated. It is a way to maintain certain structures/features of the real data that are difficult to simulate. Although the data-generating process in the current study follows the definition of pseudosimulation, we will use "simulation" in the rest of the report to describe the data generating process because the difference between the two terms is not meaningful for the current study.

[3] The 2017 NAEP mathematics Grade 8 assessment involved 179 items before scaling. Two items were removed in the scaling process due to content or model-data fit issues. These two items were not included in this study, making the total number of items 177.

[4] Each booklet consists of two blocks.

[5] In the NAEP operational analysis, polytomous items are fitted with generalized partial credit models (GPCM, Muraki, 1992). In the current study, nonzero categories of a polytomous item were collapsed into one category, while the original zero category remained the zero category.

[6] The prior on the *c*-parameter is a two-parameter beta distribution with its parameter values determined by the number of response options for an item and a weight factor of 50 (see Allen et al., 2001, Chapter 12). Single-selection multiple-choice items in Grade 8 mathematics assessments have five options, and the corresponding prior is beta (11, 41).

[7] The computer program DGROUP has several versions for researchers to choose depending on the latent structure they specify and the estimation method they want to use. For the current study, we used the version named BGROUP as the latent structure is unidimensional.

[8] For a detailed description of PVs, see Allen et al. (2001, Chapter 12).

[9] NAEP achievement levels are performance standards that describe what students should know and be able to do. Results are reported as percentages of students performing at or above three NAEP achievement levels (i.e., NAEP Basic, NAEP Proficient, and NAEP Advanced).

[10] In the Everyone condition, there was no sampling process and every record in the student population was used. Students' item responses were simulated in each replication. The Everyone condition is labeled as per-item sample size equals 28,000 in the plots for the convenience of creating ordered labels.

[11] NAEP employs a multistage probability sampling design that is more complex than the sampling design used in the current study. Besides, to properly reflect the difference in sampling probabilities, the NAEP operational estimation process involves the usage of sampling weights, while the current study made another simplification and did not use sampling weights.

[12] A population value is defined as the value of a group-score statistics calculated from the student population, using the true ability values. The population values of the group-score statistics under consideration are listed in Table 2.