Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 61

Effects of Calibration Approach and Ability Distribution on IRT Observed-Score Equating

Hyung Jin Kim[†]
Won-Chan Lee

May, 2026

[†]Hyung Jin Kim is Psychometrician, National Council of State Boards of Nursing (NCSBN) (email: hyungjin-kim@uiowa.edu). Won-Chan Lee is Director, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: won-chan-lee@uiowa.edu).

Disclaimer

The opinions expressed herein are solely those of the authors and do not necessarily represent those of National Council of State Boards of Nursing.

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Fax: 319-384-0505
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

When items are calibrated, a central question concerns the metric on which final item parameter estimates (IPEs) are expressed. One assumes that IPEs are expressed on the scale of the population (prior) distribution, typically $N(0, 1)$. Another perspective assumes that this metric is defined by the scale of the posterior ability distribution (PD) produced by the calibration process. While this distinction may have minimal impact in many standard applications of Item Response Theory (IRT), it can have meaningful implications in contexts such as test equating. In IRT observed-score equating, for instance, the choice between a $N(0, 1)$ prior and a PD can influence the shape of fitted score distributions and, ultimately, the resulting conversion tables. This study investigates the issue by (a) comparing several widely used calibration methods and software programs, and (b) examining how resulting differences in IPEs and PDs, combined with different assumptions about the ability distribution, affect fitted score distributions and equating outcomes.

1 Two Perspectives on the Scale of Item Parameter Estimates

The Expectation-Maximization (EM) algorithm for estimating item parameters is an iterative process consisting of two steps: the E-step and the M-step. Briefly, the E-step uses a prior ability distribution to compute the expected values of the sufficient statistics for the complete data. The M-step, then, updates item parameter estimates (IPEs) and the posterior ability distribution (PD) by maximizing the complete data likelihood based on the sufficient statistics obtained from the preceding E-step. The mean and SD of the PD might be the same as, at least very similar to, those of the prior distribution, but not skewness or kurtosis suggesting that the shape can differ

The scale on which final IPEs are expressed depends on which step of the EM algorithm is emphasized. In the E-step, an ability distribution of the intended population is used as a prior. Since item parameters are calibrated using a sample assumed to be representative of that population, one could argue that the scale of the final IPEs should align with this population ability distribution. Alternatively, the focus can shift to the M-step where IPEs are jointly updated with the PD. Given that the IPEs and PD function as a pair within the complete-data likelihood, the scale of the final IPEs can be considered to align with the metric of the PD.

Consequently, there are two prevailing perspectives regarding the scale on which final IPEs are expressed: (1) the scale of the prior for the population ability distribution (e.g., $N(0, 1)$), and (2) the scale of PD obtained along with IPEs (Baker, 1990; Baker & Kim, 2004, pp. 174-175).

2 Calibration Approaches

When a standard normal distribution with a mean of 0 and a standard deviation of 1 (i.e., $N(0,1)$) is assumed as the prior for the population ability distribution in the EM algorithm, there are two options for specifying item parameters at the beginning of each E-step. The first option is to use the IPEs from the previous M-step. The second option involves transforming the IPEs from the previous step using the same scale transformation applied to rescale the posterior distribution (PD) to have a mean of 0 and a standard deviation of 1. This procedure is referred to as scaling to the (0, 1) metric.

Under the standard normal (SN) assumption, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003), by default, updates both IPEs and the PD to the (0, 1) metric after each EM cycle, although this rescaling can be suppressed via a program option. ICL (Hanson, 2002) does not provide a built-in rescaling option but can be modified to perform the rescaling procedure. The flexMIRT program (Cai, 2020) does not implement rescaling.

For the empirical histogram (EH) method where each E-step uses the PD from the previous cycle as the prior, both the IPEs and PDs are rescaled to the (0, 1) metric at the end of each iteration. However, it is important to note that flexMIRT's implementation to EH method (Woods, 2007) differs from the versions implemented in BILOG-MG and ICL.

Therefore, there are three possible calibration procedures: (1) the SN as the prior (SN-as-prior) without rescaling (SN-N), (2) the SN-as-prior with rescaling (SN-Y), and (3) the EH methods, which inherently includes rescaling (EH-Y). Table 1 summarizes these three calibration procedures along with the programs capable of implementing each approach.

Table 1
Calibration Approaches and Programs

Calibration Procedure	**Estimation Method**	**Scaling to (0, 1)**	**Programs**
SN-N	SN-as-Prior	No	BILOG-MG, flexMIRT, ICL
SN-Y	SN-as-Prior	Yes	BILOG-MG, ICL
EH-Y	EH	Yes	BILOG-MG, flexMIRT, ICL

3 Two Types of Ability Distributions for IRT Fitted Score Distributions

When IRT observed-score equating is conducted, fitted score distributions for both Form X and Form Y must be constructed based on IPEs and an assumed ability distribution. The choice of an ability distribution can influence the fitted score distributions and potentially affect the final equating results.

Table 2
Approaches for Constructing IRT Fitted Score Distributions

	Estimation Method	**Scaling to (0, 1)**	**Calibration Procedure**	**Ability Distribution**
#1	SN-as-Prior	Yes	SN-Y	SN
#2				PD
#3		No	SN-N	SN
#4				PD
#5	EH	Yes	EH-Y	SN
#6				PD

For each set of IPEs from the three calibration procedures, two types of ability distributions were used to construct fitted distributions: (a) the default PD obtained alongside the set of IPEs (i.e., paired PD) and (b) $N(0, 1)$, commonly used as the prior population distribution. These two options correspond to the two competing perspectives regarding the appropriate metric for IPEs. Table 2 summarizes the six possible approaches for constructing fitted score distributions. It is important to note that the SN distribution is used in two distinct ways in this study. The first refers to its use as a prior for the population ability distribution during item parameter estimation; and the second refers to its use as the assumed ability distribution when constructing fitted score distributions.

Although many studies have compared the performance of different estimation methods and programs, most have focused primarily on differences in IPEs. There is limited research examining how such differences influence the fitted score distributions used in IRT observed-score equating. Therefore, this study investigates the extent to which differences in IPEs affect the fitted score distributions. Taking this one step further, it also explores how the choice of ability distribution impacts equating results.

4 Methodology

This section consists of four subsections. The first subsection describes datasets used as the new and old forms, followed by a subsection that briefly outlines the procedures for creating a raw-to-scale conversion table for the old form. The third subsection presents the study factors and their associated conditions, and the last subsection explains how results were evaluated.

4.1 Data

Three real data sets were selected from one large-scale assessment, each representing a different subject area. Each dataset included two data files, one for the new form and the other one for the old form. The new and old forms are referred to as Form X and Form Y, respectively. The groups of examinees administered each form were assumed to be randomly equivalent. For each dataset, the study randomly selected 30, 45, or 50 items. In addition, 3,000 examinees were randomly sampled for each form, with

Table 3
Summary Statistics for Three Data Sets

Data Set	Form	Number of Items	Min	Max	Mean	SD
1	X	30	2	30	18.053	4.996
	Y	30	2	30	16.973	5.614
2	X	45	1	45	23.874	9.524
	Y	45	3	45	24.008	8.204
3	X	50	2	49	28.417	8.598
	Y	50	3	50	29.514	8.937

responses corresponding to the selected items. Table 3 provides summary statistics for the datasets.

4.2 Creating Raw-to-Scale Score Conversion Table for Form Y

Since rounded scale scores are typically reported to examinees, the study created a separate raw-to-scale score conversion table for Form Y in each dataset. First, the Kelley's rule of thumb (Kolen and Brennan, 2014) was applied to determine the number of distinct scale-score points. Then, the arcsine transformation method (Kolen, 1988) was used to achieve approximately equal conditional standard errors of measurement across the scale-score points. As a result, for Form Y in the first dataset, the rounded scale scores ranged from 20 to 55 in increments of one. This conversion table was used to obtain the rounded scale scores for the new form through IRT observed-score equating.

4.3 Study Factors

All items were calibrated using dichotomous IRT models with a scaling constant of 1, specifically the two-parameter logistic (2PL) and three-parameter logistic (3PL) IRT models. To estimate item parameters, the study used three calibration programs: BILOG-MG, flexMIRT, and ICL programs. Calibration settings were kept as consistent as possible across the three programs. For the EM algorithm, the maximum number of EM cycles was set to 3,000 with a convergence criterion of 0.0001 . Within each M-step, the maximum number of iterations and the convergence criterion were also set to 3,000 and 0.0001, respectively. The population ability distribution was specified using 49 equally spaced points ranging from -6 to 6 in increments of 0.25. The study did not impose any prior distributions on the item parameters. Table 4 summarizes the calibration settings.

Two calibration methods were considered: (1) the SN-as-Prior method and (2) the EH method. For the SN-as-Prior method, two options were considered for rescaling to the (0, 1) metric: (a) with rescaling and (b) without rescaling. The EH method, by definition, always applies rescaling at the end of M-steps. For constructing fitted score distributions, the study considered two types of ability distributions: the paired PD obtained with IPEs, and the SN distribution. Table 5 presents the study factors and

Table 4
Calibration Settings

Description	**Setting**
Number of EM Cycles	3,000
Number of Iterations within M-Step	3,000
Converge Criteria for EM Cycles and within M-Steps	0.0001
Quadrature Points	49 equally-spaced points from -6 to 6
Prior Distribution for Item Parameters	No Prior for a and b

Table 5
Study Factors and Conditions

Study Factors	**Conditions**
Program	BILOG-MG, flexMIRT, ICL
Estimation Method	SN-as-Prior, EH
Scaling to the (0, 1) Metric	Yes, No
Ability Distribution	PD, SN

their conditions.

For simplicity, calibration procedures are referenced using abbreviation codes. A two-word abbreviation (i.e., XX-X) refers to a calibration procedure where the first word indicates the estimation method (SN for SN-as-Prior for EH for EH), and the second word indicates the rescaling option (Y for rescaling and N for no rescaling). For example, the SN-N approach refers to the calibration procedure where item parameters were estimated using the SN-as-Prior method without rescaling. Table 1 presents all two-word abbreviations used in this study. In cases where three-word abbreviations are used (e.g., X-XX-X), the first word denotes calibration program (B for BILOG-MG, F for flexMIRT, and I for ICL), and the remaining two words follow to the same structure as the two-letter codes. Table 6 lists all three-word abbreviations used in this study.

4.4 Evaluation

To evaluate results, for both Form X and Form Y, the study examined absolute differences in IPEs across the three calibration procedures and three programs. IPEs were compared in three ways: (1) across the three programs using the SN-as-Prior estimation method, (2) across the three programs using the EH method, and (3) between the SN-as-Prior and EH methods within the same calibration program. For the first comparison, IPEs from BILOG-MG with rescaling (i.e., B-SN-Y) served as the baseline against which IPEs from flexMIRT and ICL were compared. Similarly, for the second comparison, IPEs from the B-EH-Y approach were used as the reference for comparisons across programs using the EH method. In the third comparison, since the EH method inherently implements rescaling, IPEs from the SN-as-Prior method with rescaling (i.e., SN-Y) method served as the baseline. In addition to IPEs, the study also examined and compared the PDs obtained alongside the IPEs.

Using each combination of IPEs and an ability distribution (paired PD or SN), the

Table 6
Three-Letter Abbreviations for Calibration Approach

Calibration Procedure	Computer Program	Estimation Method	Scaling to (0, 1)
B-SN-Y	BILOG-MG	SN-as-Prior	Yes
B-SN-N			No
B-EH-Y		EH	Yes
F-SN-N	flexMIRT	SN-as-Prior	No
F-EH-Y		EH	Yes
I-SN-Y	ICL	SN-as-Prior	Yes
I-SN-N			No
I-EH-Y		EH	Yes

study constructed fitted score distributions for Form X and Form Y. Based on these distributions, the study conducted IRT observed-score equating from Form X to Form Y. For each type of ability distribution, the study computed differences between Form Y equivalents and Form X raw scores and compared patterns in the difference-scores across calibration procedures and programs. Rounded scale scores were also compared across all possible calibration conditions.

Assuming that the groups of examinees taking each form adequately represent the target population, it is reasonable to expect that, in the context of IRT observed-score equating, appropriate combinations of IPEs and ability distributions should yield fitted distributions that are smooth and closely match the observed score distributions. Furthermore, IRT observed-score equating should produce equated scores that do not deviate substantially from unsmoothed equivalents. Accordingly, for both Form X and Form Y, the fitted score distributions were compared against the observed score distributions, and the equated scores using IRT observed-score equating were compared against the unsmoothed equivalents across the full range of Form X raw-score points.

5 Results

The results are organized and presented in four subsections. The first subsection reports results for the IPEs. The second and third subsections compare PDs and fitted score distributions, respectively. Equating results are discussed in the final subsection. Since the results were virtually identical across the three datasets and consistent across both the 2PL and 3PL IRT models, only the results for the 45-item datasets under the 2PL model are presented here.

5.1 Differences in Item Parameter Estimates

When IPEs were compared across the three programs using the SN-as-Prior estimation method, IPEs from flexMIRT and ICL were compared against those from BILOG-MG with rescaling (i.e., B-SN-Y). Table 7 summarizes the absolute differences in IPEs across

Table 7
Absolute Differences in Item Parameter Estimates among Different Programs for the SN-as-Prior Method

Program	Scaling to (0, 1)	Parameter	Form X			Form Y		
			Min	Max	Mean	Min	Max	Mean
BILOG-MG	Yes	a	-	-	-	-	-	-
		b	-	-	-	-	-	-
	No	a	0.001	0.007	0.003	0.000	0.006	0.003
		b	0.000	0.006	0.002	0.000	0.015	0.003
flexMIRT	No	a	0.001	0.008	0.004	0.001	0.007	0.004
		b	0.000	0.007	0.003	0.000	0.016	0.004
ICL	Yes	a	0.001	0.007	0.003	0.000	0.006	0.003
		b	0.000	0.006	0.002	0.000	0.015	0.003
	No	a	0.001	0.008	0.004	0.000	0.007	0.003
		b	0.000	0.007	0.002	0.000	0.015	0.003

Note. - refers to a baseline used for comparing IPEs.

Table 8
Absolute Differences in Item Parameter Estimates among Different Programs for the EH Method

Program	Scaling to (0, 1)	Parameter	Form X			Form Y		
			Min	Max	Mean	Min	Max	Mean
BILOG-MG	Yes	a	-	-	-	-	-	-
		b	-	-	-	-	-	-
flexMIRT		a	0.000	0.006	0.002	0.002	0.072	0.032
		b	0.000	0.005	0.001	0.001	0.086	0.021
ICL		a	0.001	0.006	0.003	0.006	0.148	0.070
		b	0.000	0.005	0.002	0.000	0.158	0.044

Note. - refers to a baseline used for comparing IPEs.

the three programs in terms of minimum, maximum, and mean values. Based on Table 7, all summary numbers were close to zero, indicating that the IPEs were highly consistent across the three programs, regardless of how rescaling was handles.

Table 8 presents the results of comparing IPEs across the three programs when the EH method was used for calibration. As with the SN-as-Prior method, absolute differences in IPEs between flexMIRT/ICL and BILOG-MG were computed and summarized using the same three descriptive statistics. Table 8 shows that, although the IPEs were still quite similar across programs, the differences were somewhat larger than those observed under the SN-as-Prior method.

Lastly, for each calibration program, IPEs based on the EH method were compared to those obtained using the SN-as-prior method. As shown in Table 9, the differences between the two calibration methods were relatively large–larger than the program-to-

Table 9
Absolute Differences in Item Parameter Estimates between the SN-as-Prior and EH methods within Each Program

Program	Estimation Method	Scaling to (0, 1)	Parameter	Form X			Form Y		
				Min	Max	Mean	Min	Max	Mean
BILOG-MG	SN-as-Prior	Yes	a	-	-	-	-	-	-
			b	-	-	-	-	-	-
	EH		a	0.000	0.356	0.080	0.007	0.662	0.196
			b	0.001	0.153	0.036	0.003	1.335	0.126
flexMIRT	SN-as-Prior	No	a	-	-	-	-	-	-
			b	-	-	-	-	-	-
	EH	Yes	a	0.001	0.369	0.077	0.005	0.727	0.223
			b	0.004	0.145	0.035	0.003	1.405	0.138
ICL	SN-as-Prior	Yes	a	-	-	-	-	-	-
			b	-	-	-	-	-	-
	EH		a	0.000	0.359	0.080	0.004	0.804	0.258
			b	0.001	0.152	0.036	0.003	1.478	0.155

Note. - refers to a baseline used for comparing IPEs.

program differences observed under the EH method (see Table 8). These results suggest that the choice of calibration method (EH vs. SN-as-Prior) has a greater impact on IPEs than the choice of calibration program.

Based on the Table 7, IPEs obtained using the SN-as-Prior method were similar regardless of whether rescaling to the (0, 1) metric was applied. Moreover, results for the SN-N approach were similar to those for the SN-Y procedure in terms of PDs, fitted score distributions, and equating outcomes. Therefore, results for the SN-N procedure are not reported in the remainder of the results section. Additionally, since IPEs obtained using the SN-Y procedure were consistent across BILOG-MG, flexMIRT, and ICL, the results for the B-SN-Y procedure are used to represent the SN-Y approach more generally. As a result, the remaining sections compare and discuss results from the B-SN-Y, B-EH-Y, F-EH-Y, and I-EH-Y approaches.

5.2 Posterior Distributions

Figure 1 presents the PDs obtained using four different calibration approaches (B-SN-Y, B-EH-Y, F-EH-Y, and I-EH-Y) for both Form X and Form Y. Note that PDs obtained using the SN-as-Prior method were very similar across the three programs and across the two rescaling options.

Based on Figure 1(a), for Form X, the PD from the B-SN-Y approach displayed a unimodal, bell-shaped curve. All PDs derived from the SN-as-Prior method exhibited similar shapes, regardless of the calibration program. The PDs estimated using the EH method (i.e., B-EH-Y, F-EH-Y, and I-EH-Y) also followed a comparable pattern across the three programs, though they appeared bumpier than thos from the B-SN-

Y approach. Under the EH method, the differences in PDs across the three different programs tended to be larger than those observed with the SN-as-Prior method.

Similar patterns were found for Form Y. As shown in Figure 1(b), the SN-as-Prior method continued to yield smooth, unimodal, bell-shaped PDs for Form Y. In contrast, the PDs produced using the EH method were also bumpy and showed consistent patterns across the three programs, though the bumpiness was more pronounced than that observed for Form X. It is worth noting that Form Y had two items with extreme difficulty parameter estimates–one with $b > 4$ and another with $b <$ -3.8. The instability of these item parameter estimates may have propagated into the PDs, given that the IPEs and PDs are updated jointly in the M-step. Additionally, it is possible that the sample drawn for Form Y comprised subgroups with subtly different ability distributions, which the EH method may have reflected as irregular bumps rather than clean modes. Despite the more pronounced bumpiness in Form Y, discrepancies in PDs across programs were more evident under the EH method than under the SN-as-Prior method, consistent with the pattern observed for Form X.

5.3 IRT Fitted Score Distributions

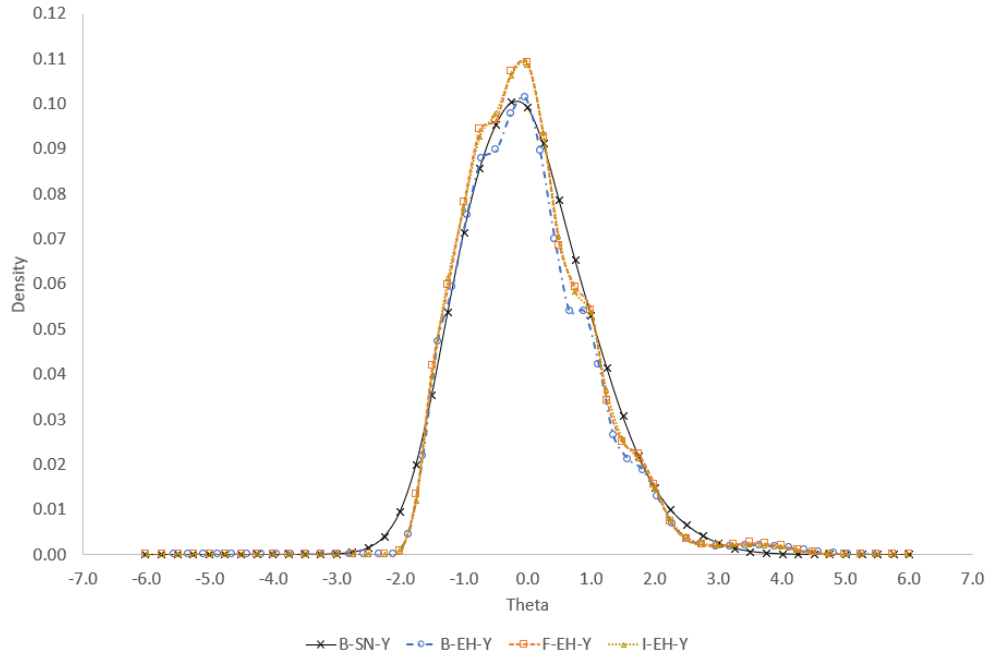
Figure 2 presents IRT fitted score distributions for Form X, constructed using the four calibration approaches (B-SN-Y, B-EH-Y, F-EH-Y, and I-EH-Y) in combination with two ability distributions (paired PD and SN). Figure 2(a) shows the fitted score distributions based on the IPEs and their paired PDs, while Figure 2(b) shows the distributions based on the SN distribution. In both plots, the solid gray line represents the observed score distribution for Form X.

In Figure 2(a), the fitted distributions were highly similar across the four calibration approaches. However, the distribution based on the B-SN-Y approach showed a slightly different pattern compared to those based on the EH method. Figure 2(b) also shows consistent results when the SN distribution was used to construct fitted distributions. The fitted score distributions from the EH method were similar across the three programs but differed somewhat from the distribution produced by the B-SN-Y approach. In terms of how well the fitted distributions recovered the observed score distribution, those based on the paired PDs tended to align more closely with the observed distribution than those based on the SN distribution.

Similarly, Figure 3 displays fitted score distributions for Form Y based on the two types of ability distributions. The results followed a similar pattern. When the distributions were constructed using paired PDs (Figure 3(a)), all four calibration approaches closely recovered the observed distribution. The fitted distributions from the EH method were similar across the three calibration programs but showed somewhat larger deviations compared to the distribution from the B-SN-Y approach. When the SN distribution was used to construct the fitted distributions (Figure 3(b)), the distribution from the B-SN-Y calibration approach appeared noticeably different from those produced using the EH method. For Form Y, the fitted score distributions based on the paired PDs provided a better match to the observed score distribution than those based on the SN distribution. Moreover, when the SN distribution was used, the deviations from the

Figure 1. Comparisons in Posterior Distributions

(a) Form X



(b) Form Y

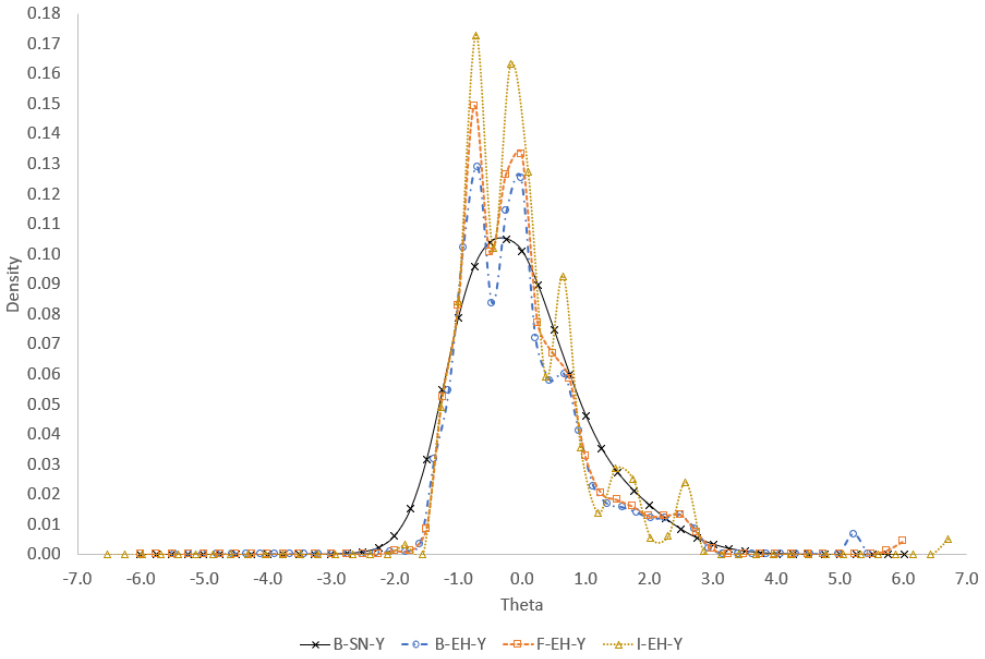
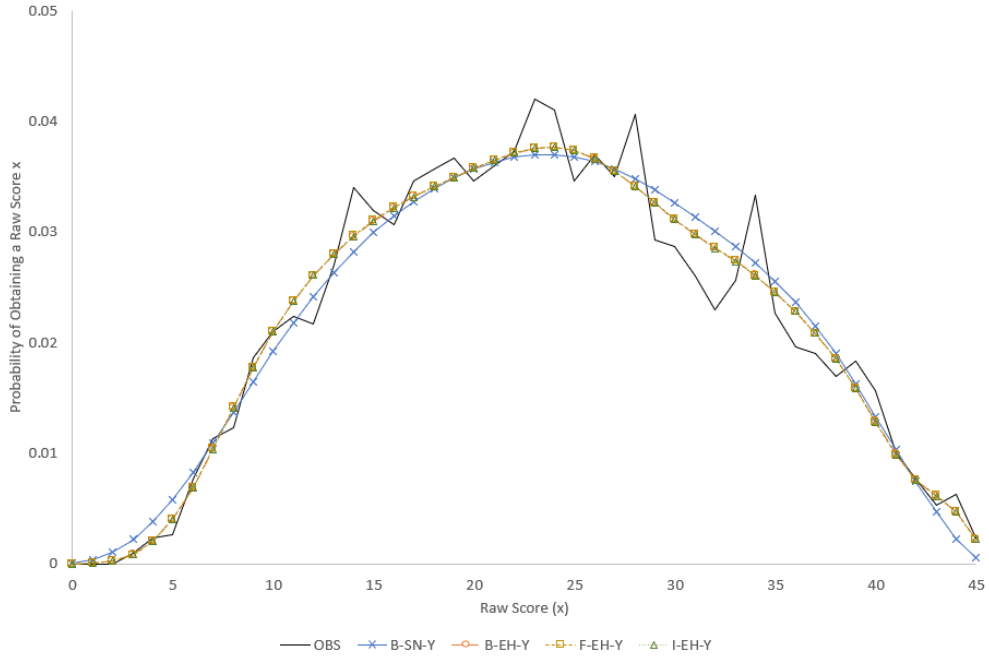


Figure 2. Comparisons in IRT Fitted Score Distributions: Form X

(a) Using Paired PD



(b) Using SN Distribution

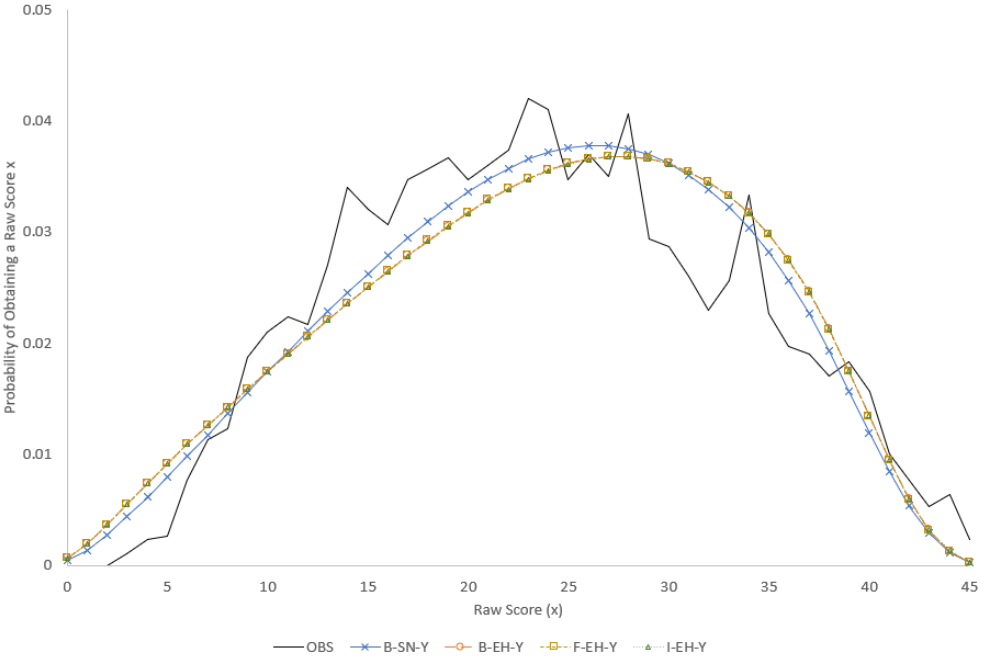
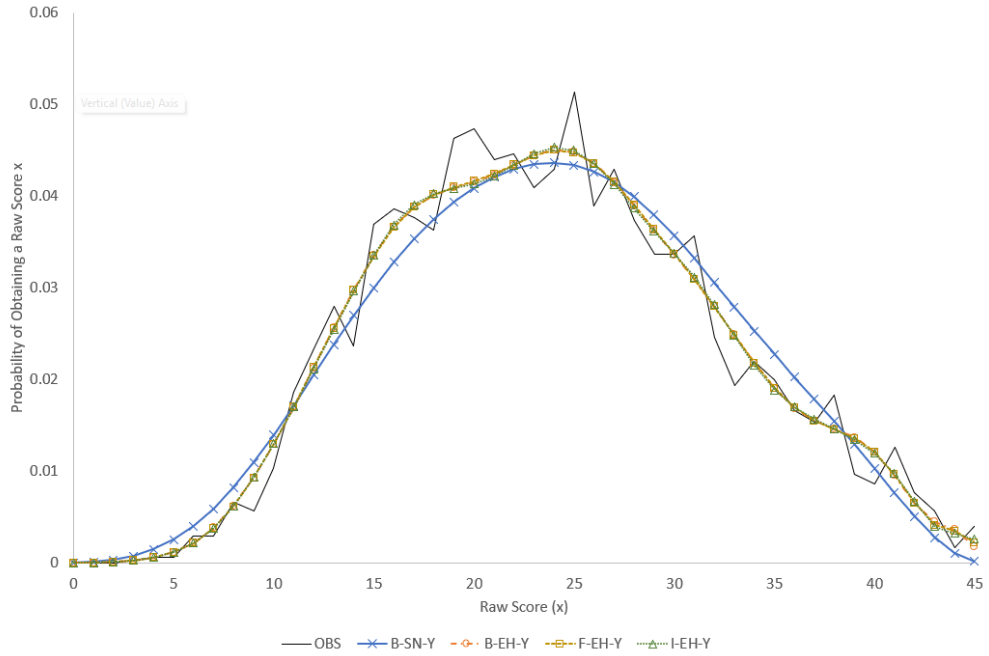
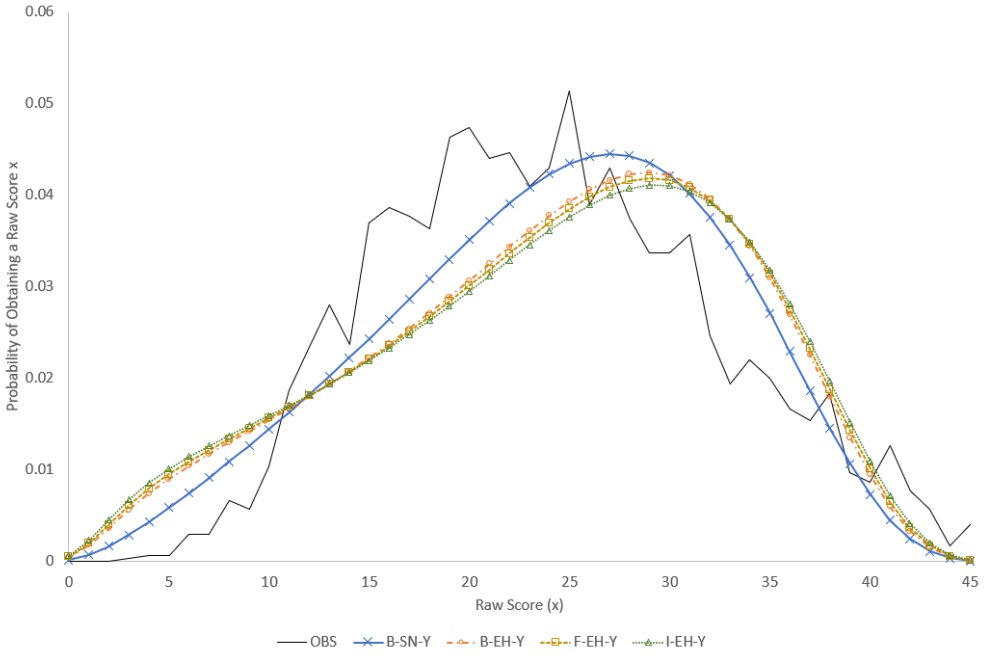


Figure 3. Comparisons in IRT Fitted Score Distributions: Form Y

(a) Using Paired PD



(b) Using SN Distribution



observed distribution were more pronounced for Form Y than for Form X.

5.4 Equating Results

Based on the two types of fitted score distributions (paired PD and SN), IRT observed-score equating was conducted from Form X to Form Y. Figure 4(a) presents a difference plot using fitted score distributions constructed from the paired PDs for both forms. Note that, since the PDs differ between Form X and Form Y, separate PDs were used for each form in equating process. In the plot, the horizontal axis represents the raw scores on Form X, and the vertical axis shows the difference between Form Y equivalents and the Form X raw score. The solid light gray line represents the unsmoothed equivalents, and the two dashed gray lines indicate the 1SE band around those equivalents. The other four lines represent the equated scores derived from item parameters estimated using the B-SN-Y, B-EH-Y, F-EH-Y, and I-EH-Y approaches.

As shown in Figure 4(a), all four approaches yielded equivalents that were generally close to the unsmoothed equivalents and remained within the 1SE band for most raw score points. The results based on the B-SN-Y approach appeared somewhat different from those based on the EH method, while the EH-based results were highly consistent across the three calibration programs.

When the SN distribution was used to construct the fitted score distributions, the results differed noticeably from those based on the use of paired PDs. As shown in Figure 4(b), among the four approaches, the approaches associated with the EH method produced relatively similar equating results across the different calibration programs. In contrast, the results from the B-SN-Y approach were noticeably different from those obtained using the EH method. Across all four approaches, the equated scores tended to deviate more from the unsmoothed equivalents compared to those shown in Figure 4(a). Among them, the B-SN-Y approach gave equivalents that were somewhat closer to the unsmoothed results than those based on the EH method.

In addition to the difference plots, the study also compared rounded scale scores across the different calibration approaches and the three programs. Table 10 summarizes these comparisons when separate PDs were used for Form X and Form Y in IRT observed-score equating. The first column lists Form X raw scores, ranging from 0 to 45. The next two columns provide the rounded scales scores for BILOG-MG and ICL when item calibration was performed using the SN-as-Prior method with rescaling (i.e., the B-SN-Y and I-SN-Y). The following three columns show results for BILOG-MG, flexMIRT, and ICL using the SN-as-Prior method without rescaling (i.e., the B-SN-N, F-SN-N, and I-SN-N). The final three columns present the rounded scale scores for BILOG-MG, flexMIRT, and ICL using the EH method (i.e., the B-EH-Y, F-EH-Y, and I-EH-Y). Highlighted cells in the table indicate instances where the rounded scale scores differ from those in the immediately preceding column.

Based on Table 10, when the SN distribution was used as the prior for the population ability distribution during calibration (i.e., SN-Y and SN-N), the rounded scale scores were identical across the three programs, regardless of whether rescaling was applied. However, when the EH method was used for calibration, raw scores were converted to

different rounded scale scores depending on the programs.

Table 11 compares rounded scale scours when the SN distribution was used to construct the fitted score distributions and consequently, for equating. The overall patterns were similar to those observed when separate PDs were used for constructing the fitted score distributions. However, under the EH method, differences in rounded scale scores across the programs were more evident when the SN distribution was used than when the separate PDs were used for Form X and Form Y. Furthermore, it is worth noting that, for the same calibration procedure and program, the rounded scale scores varied depending on the choice of ability distribution used to construct the fitted score distribution.

6 Summary and Discussion

When items are calibrated, there are two prevailing perspectives regarding the metric on which final IPEs are expressed: (1) the scale of PD obtained along with IPEs, and (2) the scale of the prior for the population ability distribution (e.g., $N(0, 1)$). As an initial step toward addressing this question, the present study examined the extent to which the two perspectives produce differences in IPEs, PDs, fitted score distributions, and equating relationships. Additionally, the study explored the impact of calibration programs on IPEs and PDs, such as differences could propagate through subsequent procedures.

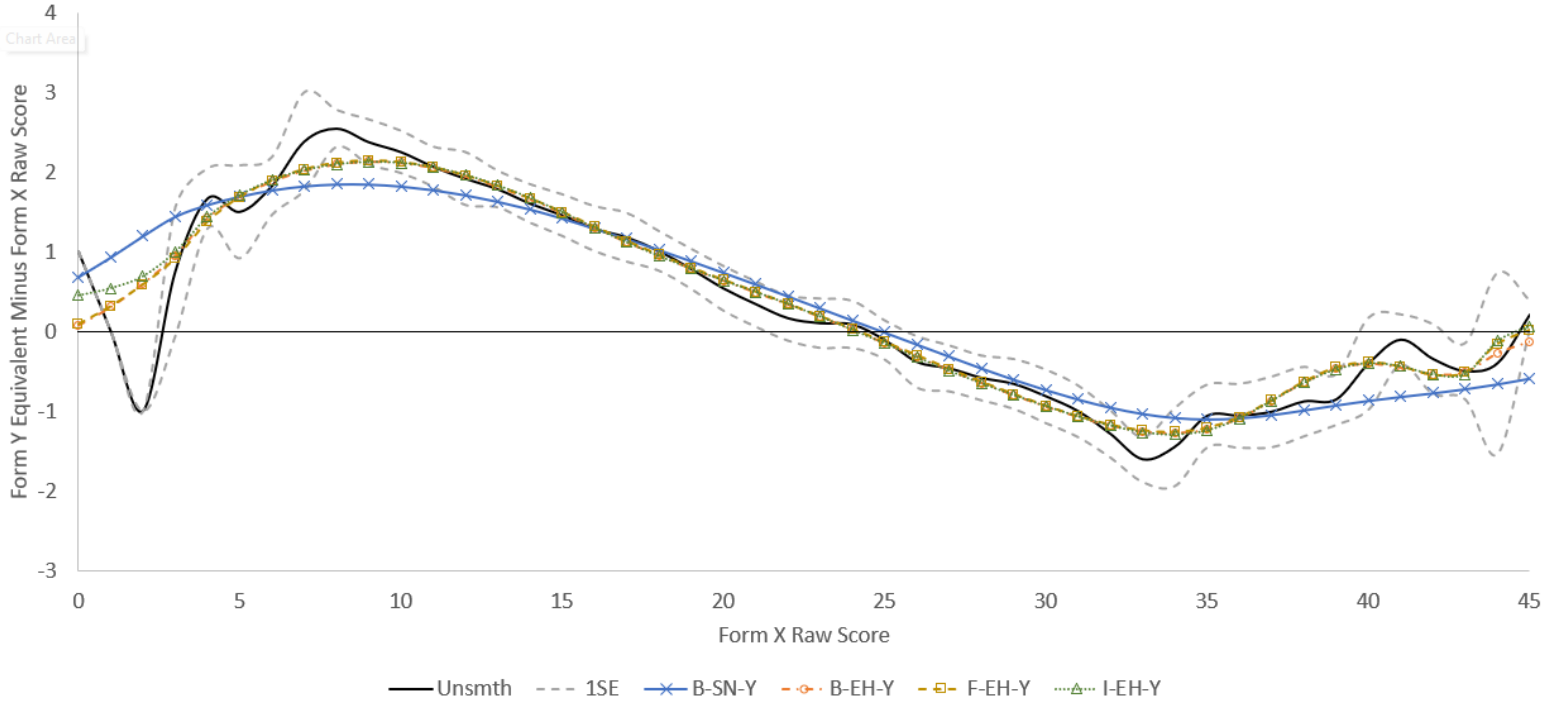
Using three programs (BILOG-MG, flexMIRT, and ICL), the study estimated item parameters using two calibration methods: SN-as-Prior and EH methods. Two rescaling options were considered for the SN-as-Prior method, while the EH method always includes a rescaling step. For constructing fitted score distributions, the study considered two ability distributions: the SN distribution and the PD obtained alongside IPEs. To disentangle potential confounding effects, the study compared results in a stepwise manner, starting with IPEs and PDs, then fitted score distributions, followed by difference plots and final conversion tables from equating.

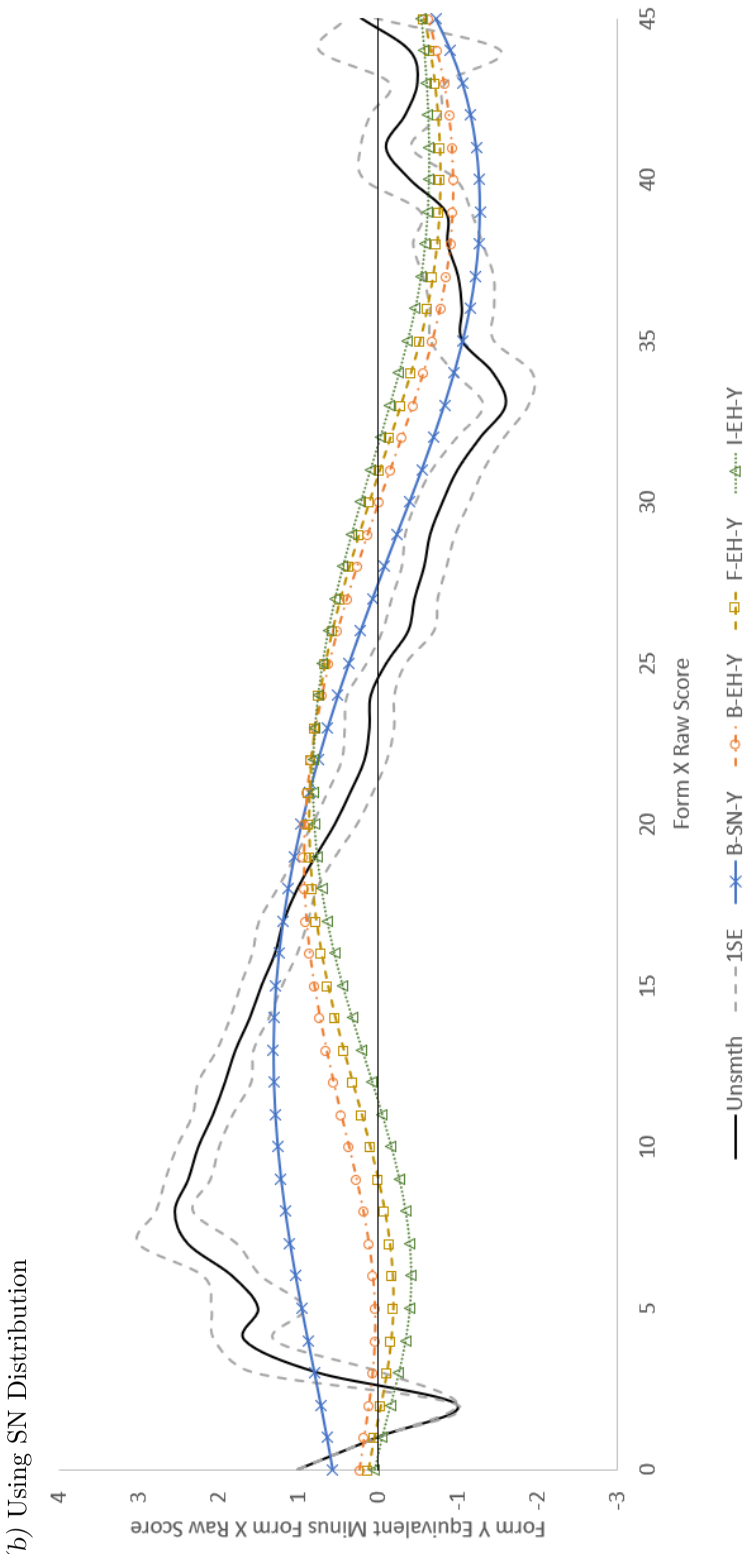
Based on the study results, rescaling to the (0, 1) metric had minimal impact on both IPEs and PDs. Within the SN-as-Prior method, IPEs were similar regardless of whether rescaling was applied at the end of the M-steps. Moreover, IPEs were more sensitive to the choice of calibration methods (SN-as-Prior vs. EH) than to the choice of calibration program. When the SN distribution was imposed as a prior, IPEs were highly consistent across the three programs. The EH method also yielded similar IPEs across programs, though the differences were somewhat larger than those observed with the SN-as-Prior method. Notably, flexMIRT appears to implement a more recent version of the EH method (Cai, 2020; Woods, 2007), while BILOG-MG and ICL do not. Despite the procedural differences in implementing the EH method, the differences in IPEs across programs were still smaller than those observed between the SN-as-Prior and EH methods, indicating that the choice of calibration method has a greater effect on IPEs than the choice of computer program.

Similar findings were observed for PDs. For the SN-as-Prior method, the PDs were

Figure 4. Difference Plots

(a) Using Separate PDs





(b) Using SN Distribution

sample size, across prescaling options and programs. The EH method also produced similar PDs across the three programs, but slightly greater variation than the SN-as-Prior method. Again, the effect of calibration method on PDs was more pronounced than the effect of calibration program.

For IRT observed-score equating, the study considered two types of ability distribution to construct fitted score distributions for Form X and Form Y: the SN distribution and the paired PD. The results showed that the fitted score distributions based on the paired PDs were similar across calibration methods and programs, and closely resembled the observed score distributions for both forms. In contrast, fitted distributions constructed using the SN distribution deviated more substantially from the observed score distributions, particularly when the EH method was used for calibration. This finding is not surprising, as the EH method does not involve the SN distribution in its estimation procedure, raising questions about the appropriateness of using the SN distribution for constructing fitted distributions when EH calibration is used.

In terms of equating performance, difference plots showed that using paired PDs for both forms resulted in equated scores that were close to the unsmoothed equivalents. When the SN distribution was used instead, the equated scores deviated more from the unsmoothed results, especially when the EH method was used to estimate item parameters. Regarding rounded scale scores, the SN-as-Prior method produced nearly identical results across all three programs, regardless of whether rescaling was applied. However, when the EH method was used, some raw scores were mapped to different rounded scale scores depending on the calibration program. It is important to note that results for rounded scale scores depend on the specific raw-to-scale conversion table used. While a different conversion table might produce different outcomes, the relative degree of similarity or difference across programs and methods would likely remain consistent.

If the groups of examinees who take the test forms adequately represent the population, then two expectations should be met in the context of IRT observed-score equating: (1) the fitted distributions should be smooth and closely match the observed score distributions, and (2) the equated scores should not deviate substantially from the unsmoothed equivalents. The study found that both expectations were better satisfied when paired PDs were used in equating, as compared to using the SN distribution. Fitted score distributions based on paired PDs more closely aligned with the observed distributions, and the resulting equated scores were more consistent with the unsmoothed equivalents. From this perspective, it appears reasonable to assume that IPEs are expressed on the scale of the posterior distribution in the context of IRT observed-score equating.

However, this study was limited to three real datasets and two dichotomous IRT models. It did not include absolute criteria for evaluating the accuracy of the resulting IPEs, PDs, fitted score distributions, or equating relationships. Future research should consider extending the study to include polytomous IRT models to assess whether the observed patterns hold across item types, calibration methods, and ability distributions. In addition, simulation studies should be conducted to evaluate the accuracy of results using statistical criteria such as conditional and overall bias, standard error, and root mean squared error. Such studies could explore additional factors, including test length,

7 References

Baker, F. M. (1990). Come observations on the metric of PC-BILOG results. *An upper asymptote for the three-parameter logistic item-response models*, *14*, 139-150.

Baker, F. M., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York, NY: Dekker.

Cai, L. (2020). *flexMIRT® version 3.6: Flexible multilevel multidimensional item analysis and test scoring*. [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Hanson, B. A. (2002). *IRT command language*. [Computer software]. Iowa City, IA.

Kolen, M. J. (1988). Defining scale scores in relation to measurement error. *Journal of Educational Measurement*, *25*, 97-110.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

Woods, C. (2007). Empirical histogram in item response theory with ordinal data. *Educational and Psychological Measurement*, *67*, 73-87.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3: Item analysis and test scoring with binary logistic models*. [Computer software]. Chicago, IL: Scientific Software.

Table 10
Comparisons in Rounded Scale Scores Using Separate PDs for Ability Distributions

	SN-Y		SN-N			EH-Y		
X	**BILOG-MG**	**ICL**	**BILOG-MG**	**flexMIRT**	**ICL**	**BILOG-MG**	**flexMIRT**	**ICL**
0	22	22	22	22	22	20	20	21
1	24	24	24	24	24	23	23	23
2	25	25	25	25	25	24	24	25
3	26	26	26	26	26	26	26	26
4	27	27	27	27	27	27	27	27
5	28	28	28	28	28	28	28	28
6	29	29	29	29	29	29	29	29
7	30	30	30	30	30	30	30	30
8	30	30	30	30	30	30	30	30
9	31	31	31	31	31	31	31	31
10	32	32	32	32	32	32	32	32
11	32	32	32	32	32	32	32	32
12	33	33	33	33	33	33	33	33
13	33	33	33	33	33	33	33	33
14	34	34	34	34	34	34	34	34
15	34	34	34	34	34	34	34	34
16	35	35	35	35	35	35	35	35
17	35	35	35	35	35	35	35	35
18	36	36	36	36	36	36	36	36
19	36	36	36	36	36	36	36	36
20	37	37	37	37	37	36	36	36
21	37	37	37	37	37	37	37	37
22	37	37	37	37	37	37	37	37
23	38	38	38	38	38	38	38	38
24	38	38	38	38	38	38	38	38
25	39	39	39	39	39	39	39	39
26	39	39	39	39	39	39	39	39
27	40	40	40	40	40	40	40	40
28	40	40	40	40	40	40	40	40
29	41	41	41	41	41	41	41	41
30	41	41	41	41	41	41	41	41
31	42	42	42	42	42	42	42	42
32	42	42	42	42	42	42	42	42
33	43	43	43	43	43	43	43	43
34	43	43	43	43	43	43	43	43
35	44	44	44	44	44	44	44	44
36	45	45	45	45	45	45	45	45
37	45	45	45	45	45	45	45	45
38	46	46	46	46	46	46	46	46
39	47	47	47	47	47	47	47	47
40	48	48	48	48	48	48	48	48
41	48	48	48	48	48	49	49	49
42	49	49	49	49	49	50	50	50
43	50	50	50	50	50	51	51	51
44	52	52	52	52	52	52	52	52
45	54	54	54	54	54	55	55	55

: indicates that the rounded scale score is different from the rounded scale score located on the left.

Table 11
Comparisons in Rounded Scale Scores Using SN Distribution for Ability Distributions

	SN-Y		SN-N			EH-Y		
X	**BILOG-MG**	**ICL**	**BILOG-MG**	**flexMIRT**	**ICL**	**BILOG-MG**	**flexMIRT**	**ICL**
0	21	21	21	21	21	20	20	20
1	23	23	23	23	23	23	23	22
2	25	25	25	25	25	24	24	24
3	26	26	26	26	26	25	25	25
4	27	27	27	27	27	26	26	26
5	28	28	28	28	28	27	27	26
6	28	28	28	28	28	28	27	27
7	29	29	29	29	29	28	28	28
8	30	30	30	30	30	29	29	29
9	31	31	31	31	30	30	30	30
10	31	31	31	31	31	31	30	30
11	32	32	32	32	32	31	31	31
12	32	32	32	32	32	32	32	32
13	33	33	33	33	33	33	32	32
14	34	34	34	34	34	33	33	33
15	34	34	34	34	34	34	34	34
16	35	35	35	35	35	34	34	34
17	35	35	35	35	35	35	35	35
18	36	36	36	36	36	36	36	35
19	36	36	36	36	36	36	36	36
20	37	37	37	37	37	37	37	37
21	37	37	37	37	37	37	37	37
22	38	38	38	38	38	38	38	38
23	38	38	38	38	38	38	38	38
24	39	39	39	39	39	39	39	39
25	39	39	39	39	39	39	39	39
26	40	40	40	40	40	40	40	40
27	40	40	40	40	40	40	40	40
28	40	40	40	40	40	41	41	41
29	41	41	41	41	41	41	41	41
30	41	41	41	41	41	42	42	42
31	42	42	42	42	42	42	42	42
32	42	42	42	42	42	43	43	43
33	43	43	43	43	43	43	43	43
34	43	43	43	43	43	44	44	44
35	44	44	44	44	44	44	44	44
36	45	45	45	45	45	45	45	45
37	45	45	45	45	45	45	45	46
38	46	46	46	46	46	46	46	46
39	46	46	46	46	46	47	47	47
40	47	47	47	47	47	47	48	48
41	48	48	48	48	48	48	48	49
42	49	49	49	49	49	49	49	49
43	50	50	50	50	50	50	50	51
44	51	51	51	51	51	52	52	52
45	53	53	53	53	53	54	54	54

: indicates that the rounded scale score is different from the rounded scale score located on the left.