Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 16, July 2015 ISSN 1531-7714

Effect of Adjusting Pseudo-Guessing Parameter Estimates on Test Scaling When Item Parameter Drift Is Present

Kyung T. Han
Graduate Management Admission Council

Craig S. Wells & Ronald K. Hambleton
University of Massachusetts Amherst

In item response theory test scaling/equating with the three-parameter model, the scaling coefficients A and B have no impact on the c-parameter estimates of the test items since the c-parameter estimates are not adjusted in the scaling/equating procedure. The main research question in this study concerned how serious the consequences would be if c-parameter estimates are not adjusted in the test equating procedure when item-parameter drift (IPD) is present. This drift is commonly observed in equating studies and hence, has been the source of considerable research. The results from a series of Monte-Carlo simulation studies conducted under 32 different combinations of conditions showed that some calibration strategies in the study, where the c-parameters were adjusted to be identical across two test forms, resulted in more robust equating performance in the presence of IPD. This paper discusses the practical effectiveness and the theoretical importance of appropriately adjusting c-parameter estimates in equating.

Multiple-choice items remain popular with educational tests, despite advances being made with many new types of item formats. Multiple-choice items still have much to offer the assessment field because they permit wide content coverage and allow for automated scoring. but the chance of candidates guessing correct answers introduces an element of random error that detracts from the measurement accuracy of any tests that include this item format.

Birnbaum (1968) developed the three-parameter logistic model (3PLM) to account, statistically at least, for the guessing behavior of low-performing candidates, and since then many studies have followed that show how model fit is improved with the inclusion of a "guessing parameter" in the statistical model (e.g., Hambleton, Swaminathan, & Rogers, 1991). This parameter, denoted as "c" in Birnbaum's representation of the model, is sometimes called the "guessing parameter." It was introduced in the statistical model to improve model fit for lower-performing candidates. Because it is common for c-parameter estimates to be smaller than the value that would result if examinees answered an item correctly by a pure random guess (in reality, the performance of lower-performing candidates is a combination of guessing and misinformation), Lord (1974) argued that it would be better to refer to the parameter as the "pseudo-guessing parameter."

Handling the *c*-Parameter in Scaling/Equating

A potentially serious problem regarding *c*-parameters arises when the 3PLM is used in the linking item set for test scaling and/or equating. Except when using concurrent calibration or the fixed common-item-parameter (FCIP; Kolen & Brennan, 2004) estimation for scaling/equating, item response theory (IRT) scaling/equating methods use a linear transformation for moving item statistics—and eventually person estimates—from one scale to another. With the mean-sigma scaling method (Marco, 1977), for example, only *b*-parameters are used to compute the scaling coefficients *A* and *B* associated with the linear transformation for mapping item statistics from one scale to another. Thus, change in *c*-parameter estimates across linking items in two forms of a test would not be accounted for directly with those scaling methods. When the mean-mean method (Loyd & Hoover, 1980) is used to compute the scaling coefficients, *a*-parameter statistics and *b*-parameter statistics are utilized, but *c*-parameter information is still not taken into account.

Even when using test characteristic curve (TCC) methods (Stocking & Lord, 1983; Haebara, 1980), which use all available item statistics, two unsolved problems remain. First, the scaling coefficients *A* and *B* determined by the TCC equating method are based on a limited range of scores across the θ scale because the loss function is computed only where there are observed scores. Score points below chance scores are not included in the scaling process. Thus, a change in the lower asymptote between two test forms may be hard to capture in the equating of scores. Second, even if the scaling coefficients *A* and *B* were appropriately computed, reflecting any differences in the lower asymptotes of the test items in the linking set, they would have no impact, ultimately, on the *c*-parameter estimates of the test items in the two forms.

Traditionally, *c*-parameter estimates are not transformed in the scaling/equating procedure. Considering the large standard errors inherent in *c*-parameter estimates, in fact, any potential gain realized in measurement precision from attempting to scale the *c*-parameters may not outweigh the additional computational complications. Given that the *c*-parameter is considered a mathematical adjustment for a chance score, there is no theoretical reason to suppose that *c*-parameter estimates vary much across groups, which may explain why adjusting the *c*-parameter in the test scaling/equating procedure has not received serious attention in the field.

Potential Issues With *c*-Parameter in Scaling When Item Parameter Drift Exists

Item response theory (IRT) assumes that item parameter values are invariant across subpopulations. A population could change over time, however, for a variety of reasons, such as, changes in curriculum, practice effects, and item exposure, and when this happens, item parameter values also can change. This is known as *item parameter drift* (IPD). When only the *b*-parameter has drifted from its original value, it is referred to as uniform IPD; when the *a*-parameter has changed, it is referred to as nonuniform IPD (Wells, Subkoviak, & Serlin, 2002).

Eliminating IPD items from the linking item set before the scaling/equating procedure commences would be considered the best strategy for preventing possible deterioration in measurement accuracy due to the IPD. To date, an extensive number of studies have been conducted to develop IPD detection methods (Donoghue & Isham, 1998; DeMars, 2004) and to resolve the issues with IPD (Bock, Muraki, & Pfeiffenberger, 1988). In fact, IPD detection, which often is based on analyses for detecting differential item functioning (DIF), has since become a standard process for many testing programs that involve test scaling/equating. It is important to understand, however, that even when using a variety of IPD detection techniques, it is nearly impossible to screen out every single IPD item from the linking set. For example, Donoghue and Isham (1998) examined the performance of several widely used IPD detection methods, including Lord's chi-square measure (1980), Raju's area measures (1988, 1990), and the Mantel-Haenszel method (Holland & Thayer, 1988). In their simulation study, they found that the power of the IPD detection methods (i.e., successful detection rate) often went to below 0.50 when the false-positive (Type I error) rate was controlled at about 0.05. The IPD detection rate could be improved by increasing the false-positive rate (that is, by moving the type I error rate from .01 to .05). Generally, however, test

contractors and test publishers are reluctant to do so because they do not want to flag too many non-IPD items due to the cost of items and content considerations. Not surprisingly then, when it comes to setting type I error rates for DIF detection studies, the .01 level is often used. This choice of level reduces the number of linking items that are deleted, including IPD items. It seems clear, then, that plenty of reasons for possible inclusions of IPD items in the link item set exist, which makes it even more critical to understand the possible impact of IPD on test scaling/equating.

Many studies have been conducted for the purpose of understanding and evaluating IPD and its consequences (Wells et al., 2002; Rupp & Zumbo; 2006; Han & Wells, 2007; Han & Guo, 2011). Most of these studies, however, focused only on IPD with *b*- and/or *a*-parameters; the effect of IPD with respect to the *c*-parameter has not been seriously considered. It is quite reasonable to expect chance scores to be relatively stable against changes in population. It is, however, important to realize that a chance score is not the only factor contributing to *c*-parameter estimates—the reason why the *c*-parameter is called the 'pseudo' guessing parameter. In fact, in most parameter estimation methods (e.g., marginal maximum likelihood estimation, MMLE), the estimation procedures for *a*-, *b*-, and *c*-parameters essentially are interdependent except when one of the parameters is fixed. As a result, if IPD with *b*- and/or *a*- parameters occurs, it can influence not only the *b*- and/or *a*-parameter estimates but also the *c*-parameter estimates. Eventually, this may result in *c*-parameter estimates that are significantly different across different times (or occasions). When this occurs in the test scaling/equating process, the consequence for ignoring the adjustment procedure for the *c*-parameter estimates is unknown. Determining which *c*-parameter estimate to use for each item after scaling also remains ambiguous, especially when the *c*-parameter difference among multiple test occasions is more than negligible.

The main research question in this study was, therefore, to investigate the resulting consequences if *c*-parameter estimates are not transformed in the test scaling/equating procedure when IPD occurs in one or more of the linking items. To address these issues, the researchers investigated a series of specific questions: (a) how stable are *c*-parameter estimates when IPD exists for one or more of the *b*-parameters, (b) how much difference would it make if the *c*-parameter estimates were adjusted, and (c) what is the effect across adjustment methods? In addressing these questions, we attempted to provide both researchers and practitioners with guidance about how they should treat the c-parameter in the presence of IPD.

Method

Data

A Monte Carlo simulation study was used to examine the effect of IPD in the *b*-parameter on *c*-parameter estimates and its consequences for linking scales between two testing administrations. We simulated two years of test administrations to model an external linking design. To make the study as realistic as possible, we based the simulation on a large-scale statewide assessment. Item bank values from a seventh-grade reading test administered in a statewide large-scale assessment were used as the generating item parameter values for the simulated tests of Years 1 and 2, respectively.

For Year 1, responses for 30 items were generated to represent the external linking items for Year 1 that were used to place the item parameter estimates for Year 2 onto the scale of Year 1. For Year 2, responses were generated for 40 scoring/operational items unique to Year 2 and 30 external linking items that were not used for scoring in Year 2 but that were common across both years. The real statewide test program that served as the basis for our study had about 40 scoring items with 3~4 linking items per individual test taker. About 30 linking items with the balanced-incomplete-block (BIB) design were administered to all examinees. In our simulation study, we chose not to implement the BIB design to simplify the research design, and instead, we simulated 30 linking items for each simulee. Regarding the sample size for the linking items, the actual statewide test was administered to more than 50,000 examinees each year, while this study had only 5,000 examinees—meaning that the sample size per linking item in our study was comparable to the real test administration. We used the 3PLM to simulate data that would represent multiple-choice item responses. Table 1 displays the descriptive statistics for the generating item parameter values.

Table 1. Descriptive Statistics for the Item Sets of the Year 2 Test

Item Set	*n*	*a*		*b*		*c*	
		Mean	SD	Mean	SD	Mean	SD
Scoring item set	40	1.000	0.173	–0.028	0.797	0.195	0.043
Linking item set (not scored)	30	1.024	0.226	–0.034	0.873	0.191	0.042

Proficiency parameter values for 5,000 simulated examinees were drawn from the standard normal distribution for Year 1: $\theta \sim N(0,1)$. For Year 2, proficiency parameter values for 5,000 examinees were sampled from a distribution to represent an improvement of 4% in terms of the proportion of examinees at or above the Proficient level: $\theta \sim N(0.10,1)$. We used the computer program WinGen (Han, 2007) to generate the item response data. PARSCALE (Muraki & Bock, 2003) was used to calibrate the item parameters via marginal maximum likelihood estimation (MMLE) and to estimate the proficiency parameters via the expected a posteriori (EAP) method. The examinees at Year 2 were classified into one of four achievement levels: Below Basic (BB), Basic (B), Proficient (P), and Advanced (A) based on the cut scores, –1.0, –0.1, and 0.9, respectively.[1]

Conditions

To examine how IPD on the *b*-parameter influences *c*-parameter estimates and the test scaling/equating process, we manipulated two different factors. First, the magnitude of IPD introduced on the *b*-parameter was manipulated to be small, medium, or large (δ = –0.1, –0.3, or –0.5, respectively)[2]. To simplify the research design and make the results more intuitively generalizable, the IPD in this study was in the direction of making test items easier on the second administration. A zero-IPD condition (δ = 0) was examined to provide baseline results from which to judge IPD's impact. IPD was introduced on 33.3% of the linking items (i.e., 10 items). Second, we conducted four distinct strategies for calibrating the item parameters, with each strategy varying with respect to how the *c*-parameter was treated and adjusted.

In the first strategy, the item parameters for Year 1 and 2 were calibrated separately—a common strategy in practice. This strategy produced two separate *c*-parameter estimates for each linking item between Year 1 and 2. Typically in practice, only the *c*-parameter estimates from Year 1 are used in the equating process (Kolen & Brennan, 2004) and the estimates from Year 2 are ignored assuming the difference in the *c*-parameter estimates across years is negligible.

In the second strategy, we performed the item parameter calibration for each year by fixing the *c*-parameters of all linking items to 0.20. This strategy is used sometimes in practice, especially for items that exhibit technical problems in *c*-parameter estimation. The value of 0.20 was chosen to represent the probability of an examinee correctly answering a multiple-choice item with five options via a random guess. Some literature suggested that actual *c*-parameter estimates might be slightly smaller than the probability of answering an item correctly by random guessing because low proficiency examinees tend to respond to attractive distracters (Lord, 1974). A recent study by Han (2012), however, investigated the *c*-parameter estimates from various real test programs and found that the *c*-parameter estimates were often very close to *k* with *k* being one divided by the number of options for multiple-choice items.

In the third strategy, the *c*-parameters of the linking items in Year 1 were freely estimated; those *c*-parameter estimates from Year 1 were then used to fix the *c*-parameters of the linking items in Year 2 (*a*- and *b*-parameters were freely estimated). One can view this strategy as a partial fixed common-item-parameter (FCIP) scaling because only the *c*-parameter estimate is fixed to the previous value. Linking/scaling is still required following item parameter estimation.

Finally, in the fourth strategy, the item parameters were calibrated separately for each year. Upon completion of the linear transformation, each of the linking items would have two different values (as in Strategy 1). In the fourth strategy, however, the two *c*-parameter estimates are averaged and used for each of the rescaled linking items, which effectively is what occurs with a concurrent calibration solution to equating. This strategy has the advantage of capitalizing on data provided by two samples of examinees and is

[1] The cut scores in the study were close to the actual cut scores (–0.89, –0.11 and 0.88) being used for the state assessment that we attempted to mimic.

[2] These values were based on previous studies, which suggested that 0.25 and 0.50 represented small and moderate amounts of IPD, respectively (Wells, Subkoviak, & Serlin, 2002; Han & Wells, 2007; Rupp & Zumbo 2006, Wollack, Sung, & Kang, 2006).

common in practice. When the sample sizes in the two administrations differ, a weighted average of the *c*-parameter estimates could be considered.

The four item calibration strategies, each distinct from the others in terms of how they compute the *c*-parameter estimates, were, in fact, not altogether new in the field. In practice one commonly observes Strategies 2, 3, or 4 (often partially) being used as alternatives to Strategy 1 when Strategy 1 alone is infeasible (for example, when some linking items fail to converge unless fixing some item parameters, usually the *c*-parameters). In other words, the main motivation for using Strategies 2, 3, or 4, in practice, is not necessarily because they outperform Strategy 1 or are theoretically superior but simply that Strategy 1 sometimes does not work. Consequences of using Strategies 2, 3, or 4 as an alternative to Strategy 1, however, have not been thoroughly studied until now. Table 2 provides a summary of the four calibration strategies.

Table 2. Four Strategies for Calibrating Linking Item Parameters

	Linking Items in Year 1	Linking Items in Year 2
Strategy 1	Freely estimate *a*, *b*, and *c*	Freely estimate *a*, *b*, and *c*
Strategy 2	Estimate *a* and *b*, fixing *c* = 0.20	Estimate *a* and *b*, fixing *c* = 0.20
Strategy 3	Freely estimate *a*, *b*, and *c*	Estimate *a* and *b*, fixing *c* to *c* estimates from Year 1
Strategy 4	Freely estimate *a*, *b*, and *c*	Freely estimate *a*, *b*, and *c*. Then, average the two *c*-parameter estimates for each item

After we estimated the item parameters, we used two different scaling methods (mean-sigma and TCC) to rescale the item parameter estimates for the Year 2 test onto the Year 1 test scale. Using the scoring items from the Year 2 that were rescaled back to the Year 1 test scale, examinees' proficiency estimates on the θ scale were then computed. Because the scoring items from Year 2 already had been rescaled to be on the same scale as the Year 1 test, the proficiency estimates from the Year 2 were automatically equated back to the θ scale of Year 1. This approach was essentially equivalent to the IRT true score equating described in Kolen and Brennan (2004). The statewide assessment that our study imitated employed additional nonlinear transformation procedures to compute the reporting scores from the proficiency estimates on the θ scale. This study did not implement an additional score transformation for reporting the score scale, but, instead, used the proficiency estimates on the θ scale as the final scores in order to improve the generalizability of the study results.

This study involved 32 conditions (4 IPD magnitudes x 4 estimation strategies x 2 scaling methods) which were each replicated 100 times.

Data Analysis

The *a*-, *b*-, and *c*-parameter estimates across the 100 replications for the linking items were summarized by their mean values. Change in average *a*-, *b*-, and *c*-estimates, as the amount of IPD changed, was visually investigated with line graphs for each item parameter estimation strategy.

We used the mean of the scaling coefficients *A* and *B* over the 100 replications to evaluate the impact of IPD and parameter estimation strategy on test equating. Once the Year 2 test items were rescaled onto the Year 1 test scale via the linear transformation with the scaling coefficients *A* and *B*, we also computed the bias on the expected scores for the linking only, scoring only, and linking + scoring item sets due to the IPD and its impact on the *c*-parameter estimates based on the TCC with the zero-IPD condition.

We evaluated the consequences of IPD, the choice of estimation strategy, and the scaling method two different ways. First, we assessed the score consistency across conditions using *RMSE* computed as:

$$RMSE_\theta = \sqrt{\frac{\sum_{i=1}^{N}(\theta_i^* - \theta_i)^2}{N}} \quad (1)$$

where θ_i^* is the IRT proficiency score of examinee *i* in Year 2 that has been equated to the Year 1 test scale, and θ_i is the true proficiency score of examinee *i*. To evaluate the systematic bias in the proficiency estimation, we computed the bias statistic, in which the signed differences between the true parameter values and estimates were averaged. Second, to evaluate the consequences of the IPD proficiency classification, we classified the equated proficiency estimates into one of four typical achievement levels—Below Basic (BB), Basic (B), Proficient (P), and Advanced (A)—based on the cut scores –1.0, –0.1, and 0.9, and then assessed the

number of misclassifications due to IPD and other conditions. This last criterion was especially important because it provided an opportunity to evaluate the practical consequence of IPD on the ultimate use of the test results, namely, making proficiency classifications.

Results

Scaling Coefficients A and B

The scaling coefficients *A* and *B* were computed using both the mean-sigma and TCC methods to place the item parameter estimates for Year 2 onto the same scale as Year 1. Figure 1 displays the average *A* and *B* values observed for each of the studied conditions across 100 replications, as well as the empirical standard errors for *A* and *B*. As observed in Figure 1 (left), the scaling coefficient *A* was affected by the IPD introduced on the *b*-parameter. There was an apparent effect due to scaling method in that the mean-sigma method was affected more in comparison with the TCC method. This likely was due to the fact that the mean-sigma method uses the standard deviation of the *b*-parameter estimates, which were directly influenced by the type of IPD simulated. We did not see any meaningful differences on *A* among the calibration strategies.

The empirical standard error for *A* (Figure 1, right) was influenced by the item calibration strategies. The item calibration Strategies 2 and 3, where the *c*-parameters for the linking items were either fixed to 0.2 or to the *c*-parameter estimates from the previous year, respectively, showed the lowest standard error regardless of scaling method. This implies that when the *c*-parameters were not estimated (as in item calibration Strategies 2 and 3), these strategies yielded more stable *a*- and *b*-parameter estimates, which in turn caused the scaling coefficient *A* to became more stable. When the mean-sigma method was used, we observed no difference in standard error of scaling coefficient *A* between the item calibration Strategies 1 and 4 where *c*-parameters for the linking items were freely estimated without fixing item parameter values. This was because averaging the *c*-parameter estimates for each linking item (in item calibration Strategy 4) had no effect on their *a*- and *b*-parameter estimates and because the mean-sigma method only used the *a*- and *b*-parameter

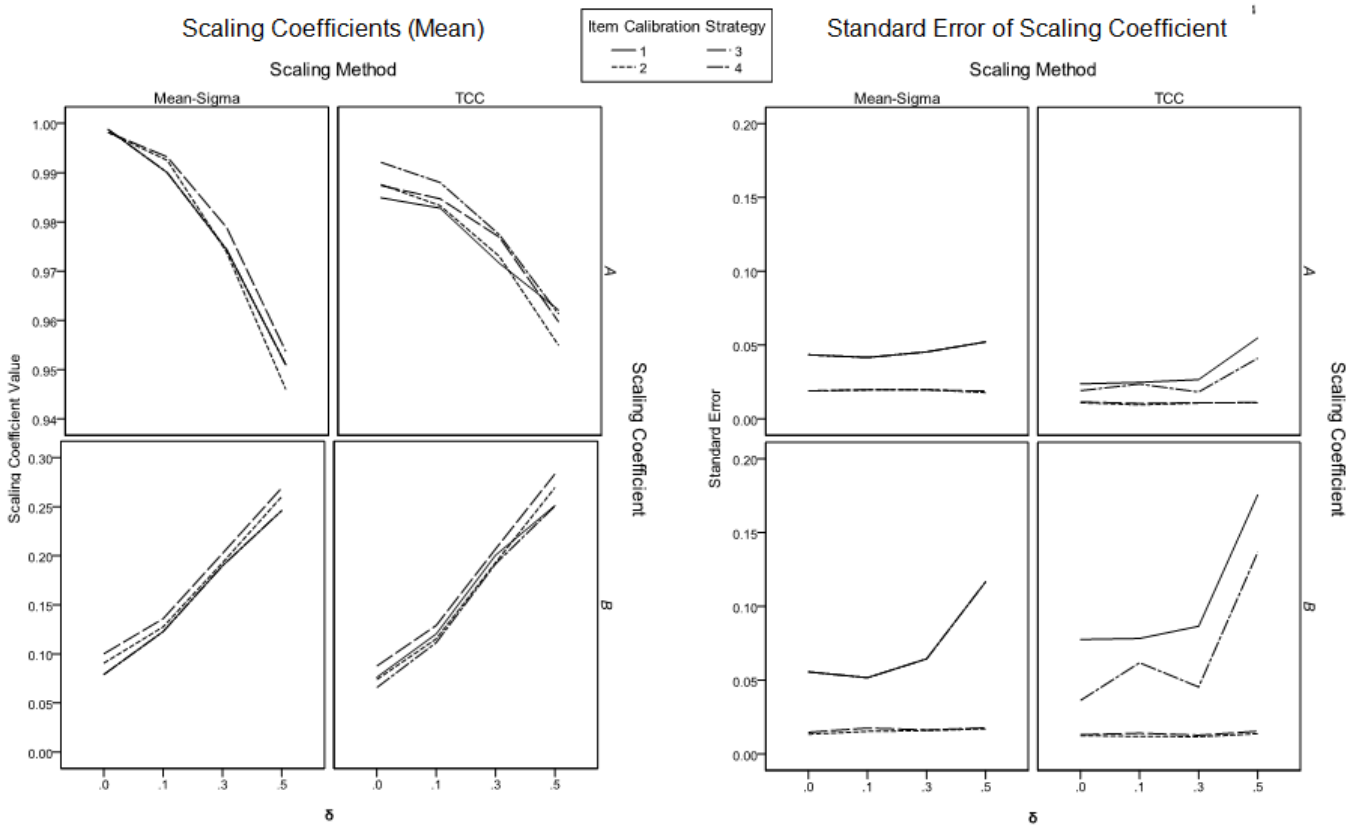

Figure 1. Change in mean (left) and standard error (right) of scaling coefficients due to IPD

estimates to compute scaling coefficient *A*.

On the other hand, with the TCC method, we did find differences in the standard error of scaling coefficient *A* between the item calibration Strategies 1 and 4. The averaged *c*-parameter estimate for each linking item in calibration Strategy 4 was reflected in the test characteristic curve computation, and it affected the scaling coefficient *A*. Item calibration Strategies 2 and 3 still exhibited more stable scaling coefficient estimates compared to Strategy 1, however. Overall, with the TCC method, the scaling coefficient *A* was more stable than it was with the mean-sigma method. Furthermore, IPD did not seem to affect the standard error.

Figure 1 also illustrates the effect of the factors on the *B* scaling coefficient. As shown in Figure 1 (left), the scaling coefficient *B* was heavily influenced by IPD regardless of item calibration strategy and scaling method. The calibration strategy apparently had no meaningful effect on the average *B*. A choice of the item calibration strategy, however, made a substantial difference in the standard error of scaling coefficient *B* estimation. In shown in Figure 1 (right), calibration Strategies 2 and 3 had the smallest standard error, whereas calibration Strategies 1 and 4 resulted in substantially larger standard errors. There were small differences in the standard error of the *B* scaling coefficient between the two scaling methods. The mean-sigma method performed better with calibration Strategies 1 and 4 than with the TCC method, but the calibration Strategies 2 and 3 performed better with the TCC method in terms of the standard error of scaling coefficient *B*. The standard error seemed to be somewhat affected by the amount of IPD, but it was hard to conclude what effect IPD had on the standard error since the pattern of change in standard error due to IPD fluctuated.

Rescaled Linking Item Parameter Estimates

After the linking items of Year 2 were transformed onto the scale of the Year 1 test, using the scaling coefficients *A* and *B*, we evaluated the impact of IPD on the item parameter estimates using *RMSE* and *BIAS* statistics. As shown in Figure 2, the scaling method did not have a meaningful effect on the *RMSE* for the *a*-parameter within each condition. The item calibration Strategies 2 and 3 tended to show slightly smaller *RMSE* values on the *a*-parameter than calibration Strategies 1 and 4 did. The *RMSE* for the *a*-parameters, however, did not seem to be affected by IPD. On the other hand, as seen in Figure 3 (left), the bias of *a*-parameter estimates was influenced by IPD. The bias of the *a*-parameter estimates for the scoring items was moderately influenced by IPD via the scaling coefficient *A*, which was already influenced by IPD. The linking items showed more stable patterns of bias, however, even with the large IPD ($\delta = 0.5$). This occurred partly because the *a*-parameter estimates—initially affected by IPD introduced on the *b*-parameters—were recovered by the scaling coefficient *A*, which reflected the influence of IPD on the item parameter scales. At the aggregated item level (linking items + scoring items), the influence of IPD on the bias was minimal.

The *RMSE* of *b*-parameter estimates for the linking items directly, heavily, and monotonically increased as IPD increased no matter which item calibration strategy or scaling method was used, as shown in Figure 2 (middle). On the other hand, the *RMSE* for the scoring items did not show meaningful changes even with large IPD. The calibration Strategies 2 and 3, again, tended to result in slightly smaller *RMSE*s of *b*-parameter estimates.

The choice of scaling method did not seem to make a meaningful difference in *RMSE*; however, it did make a considerable difference in the bias for *b*-parameter estimation, as displayed in the middle of Figure 3. With the TCC method, all (linking + scoring) items showed practically no bias, even when the amount of IPD was large ($\delta = 0.50$). The mean-sigma method resulted in slightly biased (positively) *b*-parameter estimates even when IPD was not introduced, with the bias increasing slightly as IPD increased. On the other hand, when we examined only the scoring items, the IPD was slightly less influential on the bias for *b*-parameter with the mean-sigma method compared with the TCC method. In addition, calibration strategy did not seem to make any practical difference in bias of *b*-parameter estimates.

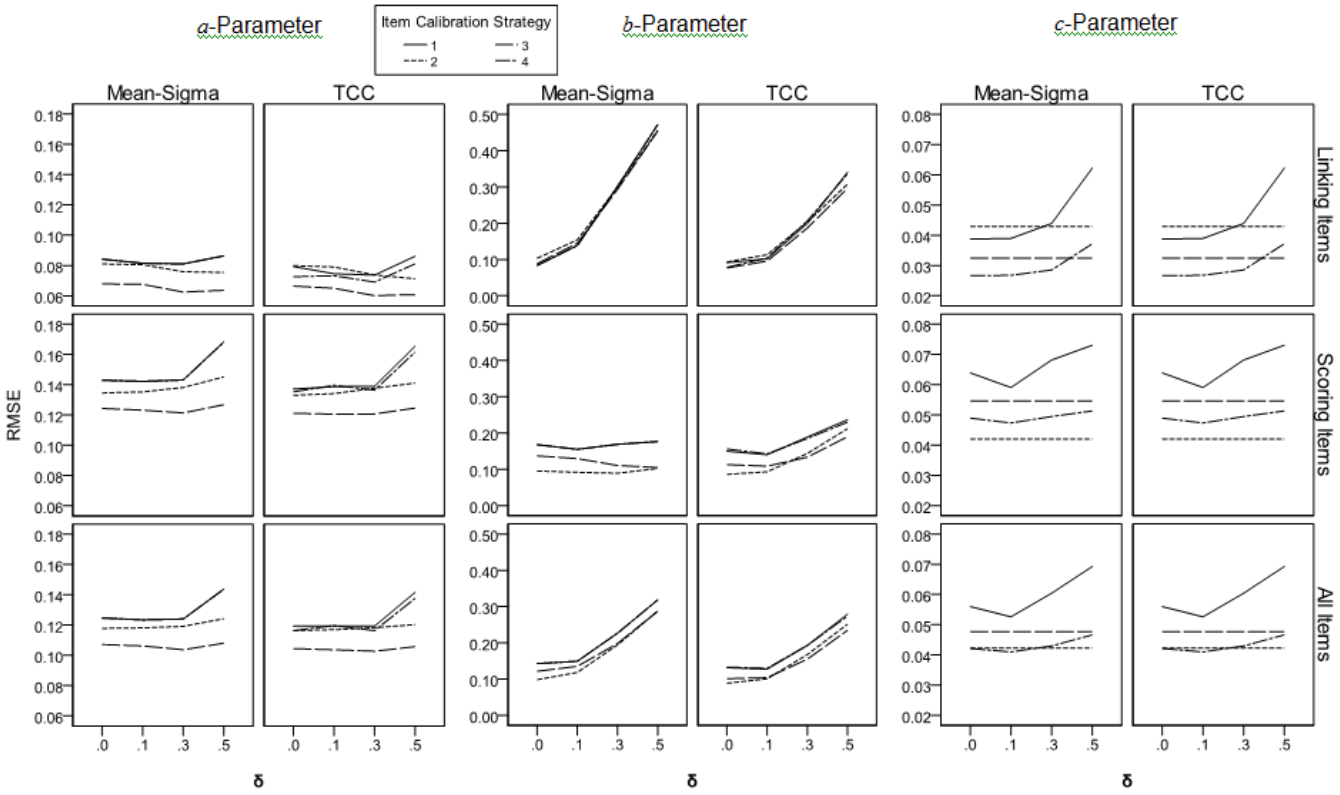

Figure 2. Root mean squared error of year 2 item parameter estimates after scaling

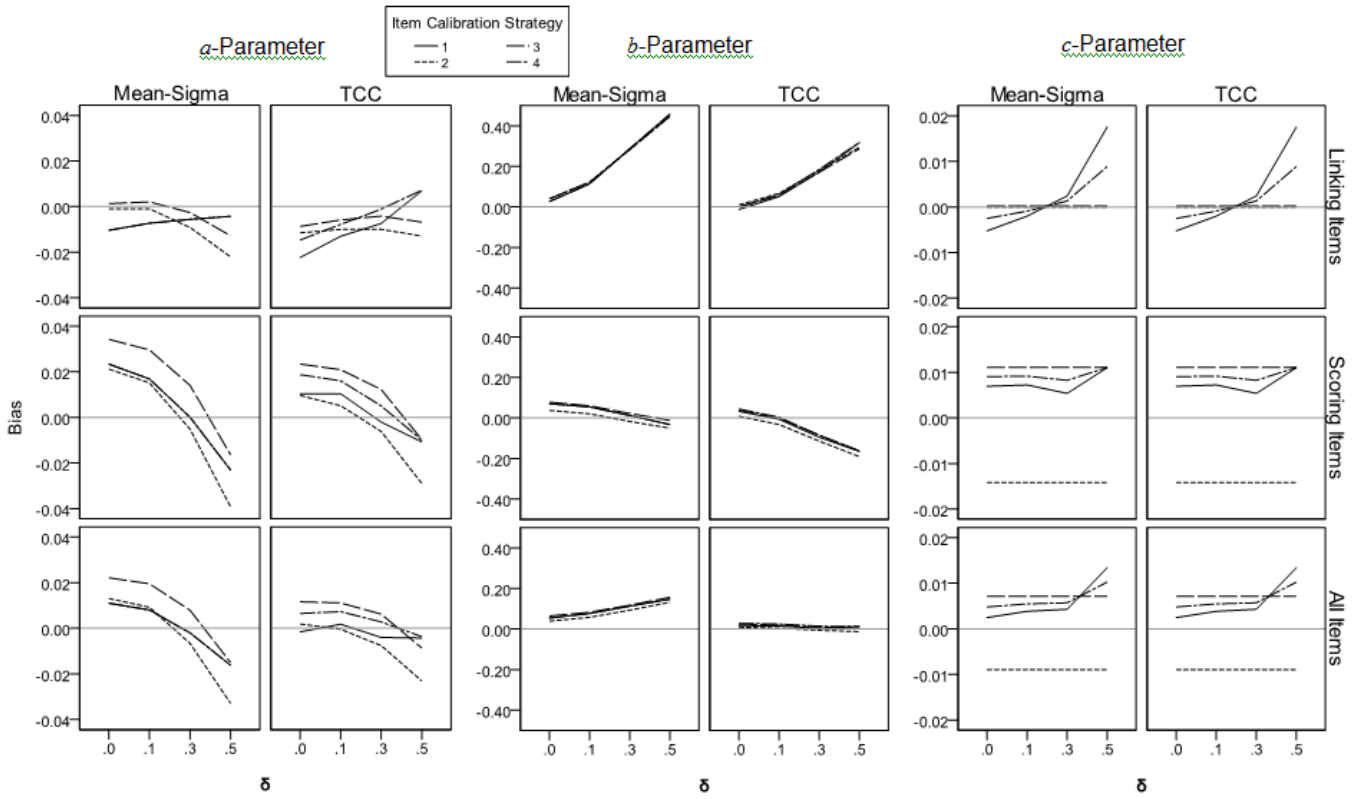

Figure 3. Bias of year 2 item parameter estimates after scaling

We also investigated the *c*-parameter estimates using *RMSE* and bias statistics (Figures 2 (right) and 3 (right)). Since *c*-parameter estimates were not rescaled with the scaling coefficients *A* and *B* (although *c*-parameters estimates were partially reflected in the computation of the scaling coefficients *A* and *B* with the TCC method), there was no effect due to choice of scaling methods. Rather *c*-parameters at Year 2 were the same as those seen in the Year 1 test with the calibration Strategies 2 and 3, achieved by fixing the estimated values for *c*-parameters. Thus, with calibration Strategies 2 and 3, the *RMSE* indicated that the *c*-parameter estimates were not influenced by the presence of IPD. Calibration Strategy 2 resulted in the smallest *RMSE*; calibration Strategy 1 tended to show the largest *RMSE*.

Most interesting was the fact that averaging *c*-parameter estimates for the linking items across years effectively reduced *RMSE*, and rendered it less affected by IPD. In terms of the bias for the *c*-parameter estimate see in Figure 3 (right), it may be inappropriate to evaluate the effectiveness of item calibration Strategy 2 by the amount of bias because the bias is directly affected by the value fixed for the *c*-parameter. Rather, the consistency of bias against IPD would be more interesting unless the bias was substantial. In fact, all four calibration strategies resulted in bias that was practically near zero even with large IPD. As expected, calibration Strategies 2 and 3 (fixing the estimates) resulted in a consistent bias that was not affected by IPD. With Strategy 1, the bias of *c*-parameter estimates for the linking items was substantially affected by IPD, but the influence of IPD on the scoring item bias was minimal. Strategy 4 (averaging) showed a pattern similar to Strategy 1, but the degree to which the bias was influenced by IPD was between that observed in Strategies 1 and 3.

Bias of Expected Score Due to IPD

The differences in TCCs between each IPD condition and non-IPD condition were compared to examine the bias of expected scores (on the theta scale) due to IPD. Figure 4 shows the bias due to IPD across the θ scale for the linking and scoring items. For the linking items seen in Figure 4 (left), it was apparent that for the mean-sigma method, the bias on the expected score dramatically increased as the magnitude (δ) of IPD increased. Moreover, the location of the maximum bias on the θ scale was near the original difficulty of the IPD linking items with IPD. In comparison, the bias for the TCC method was much smaller than that seen with the mean-sigma method. Calibration strategy had an apparent effect as well. For example, when the *c*-parameters were not controlled over time (i.e., Strategy 1), the lower tail of the bias, as seen in the first row of Figure 4 (left), increased negatively by roughly one point as the magnitude of IPD increased to 0.5, further illustrating the effect that IPD on the *b*-parameters had on the *c*-parameter estimates. On the other hand, with calibration Strategies 2 and 3, in which the *c*-parameters were fixed or controlled, the bias due to the IPD was minimized at the lower end. Strategy 4 also showed a slightly reduced impact of IPD on the expected score compared with the Strategy 1, but less than that observed for Strategies 2 and 3.

The middle section of Figure 4 also illustrates the bias on the expected scores for the scoring items due to the IPD, which may be more important and consequential in an external linking design. Overall, the results showed that examinees with a proficiency level near zero on the θ scale exhibited negative bias up to 6 score points under the worst IPD condition ($\delta = 0.5$). Unlike the case of the linking items seen in the left side of Figure 4, the mean-sigma method appeared to be slightly more robust against IPD compared with the TCC method, but the difference was not meaningful considering the total number of the scoring items, which was 40. The difference among the four calibration strategies was barely noticeable at the lower tails of Figure 4 (middle).

Although the linking items were not used for scoring with the external linking design of this study, we also evaluated the bias of expected scores on the aggregate item set (linking items + scoring items), which served as an analysis for an internal linking design. An interesting result shown in Figure 4 (right) is that the impact of IPD on the expected score was reduced by nearly half with the mean-sigma method compared with the TCC method. This was because the IPD influenced the expected scores for the linking items and the scoring items in the opposite directions with the mean-sigma method (Figure 4,left and middle), effectively canceling out a large portion of the bias due to IPD. Thus when using the internal linking design, the mean-sigma method appears more robust against IPD than the TCC method in terms of accuracy of expected score. There also were clear differences among the calibration strategies in the internal linking

case (Figure 4, right). Regardless of the choice of scaling method, the bias in the lower range of the θ scale (less than –2.0) was at least more than one point with Strategy 1. On the other hand, when the c-parameters were fixed over time (i.e., Strategies 2 and 3), the bias of expected scores in the lower range was minimized even under the large IPD condition. In addition, the impact of IPD on the bias was slightly reduced with Strategy 4 in comparison to the Strategy 1; however, Strategy 4 was not as robust against IPD as Strategies 2 and 3.

proficiency estimates with the *RMSE* and bias statistics. As far as what we expected to see from the bias of the expected score for the scoring items (Figure 4, middle), the *RMSE* and bias statistics for the proficiency estimates showed similar patterns across various scaling methods, item calibration strategies, and sample sizes (Figure 5). The mean-sigma method with calibration Strategies 1 and 4 tended to show slightly larger *RMSE* values than other combinations did. Calibration Strategies 2 and 3 resulted in consistently smaller *RMSE* values than those seen in the other strategies.

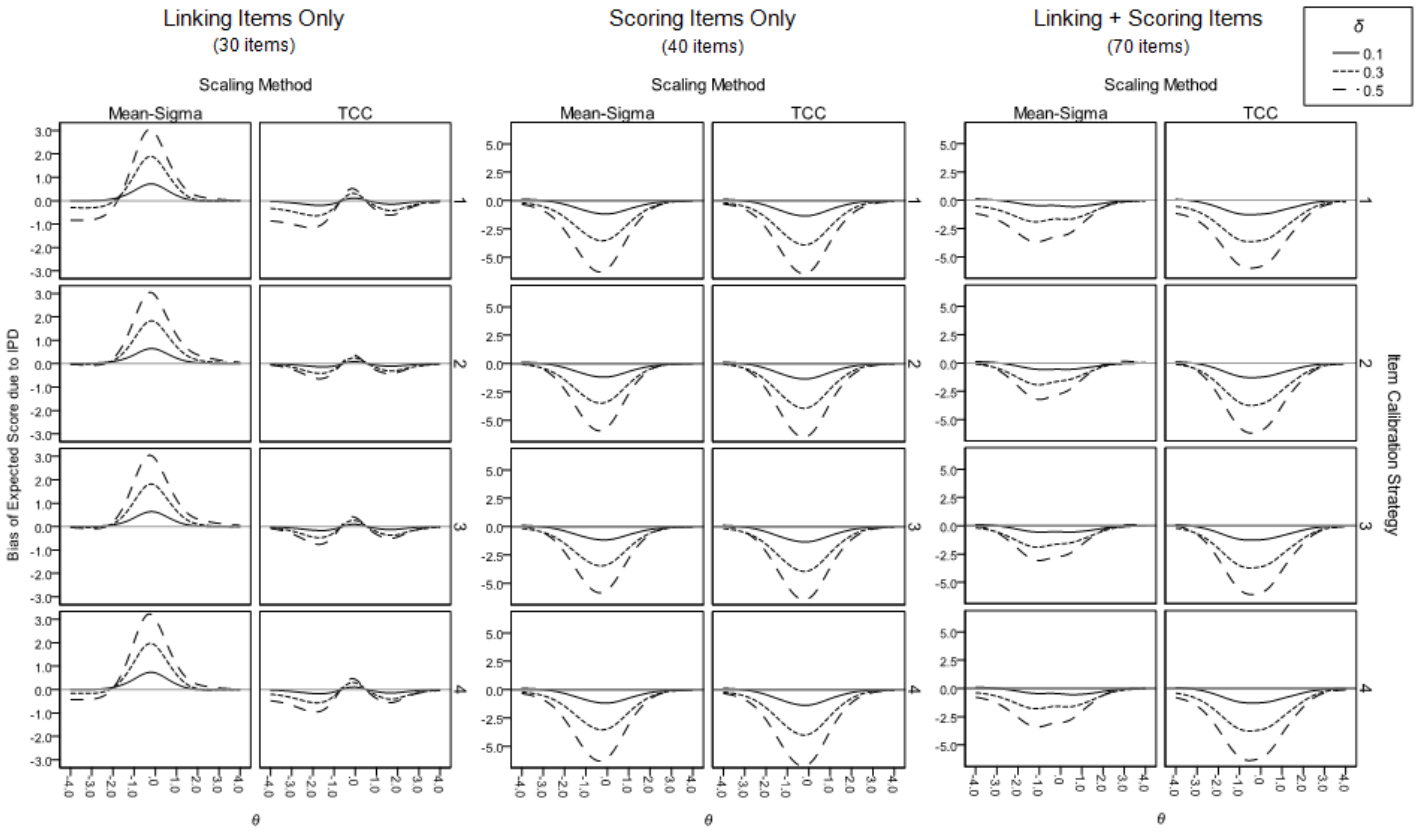

Figure 4. Conditional bias of expected score due to IPD for linking (left), scoring (middle), and linking + scoring (right) items

Proficiency Estimates and Consequences

Using the 40 scoring items, we estimated 5,000 examinee proficiency scores and evaluated the

The *RMSE* did not change with small IPD ($\delta = 0.10$), but increased moderately as the amount of IPD exceeded 0.10. In terms of bias, the proficiency estimates were biased even with the smallest IPD among the studied conditions.

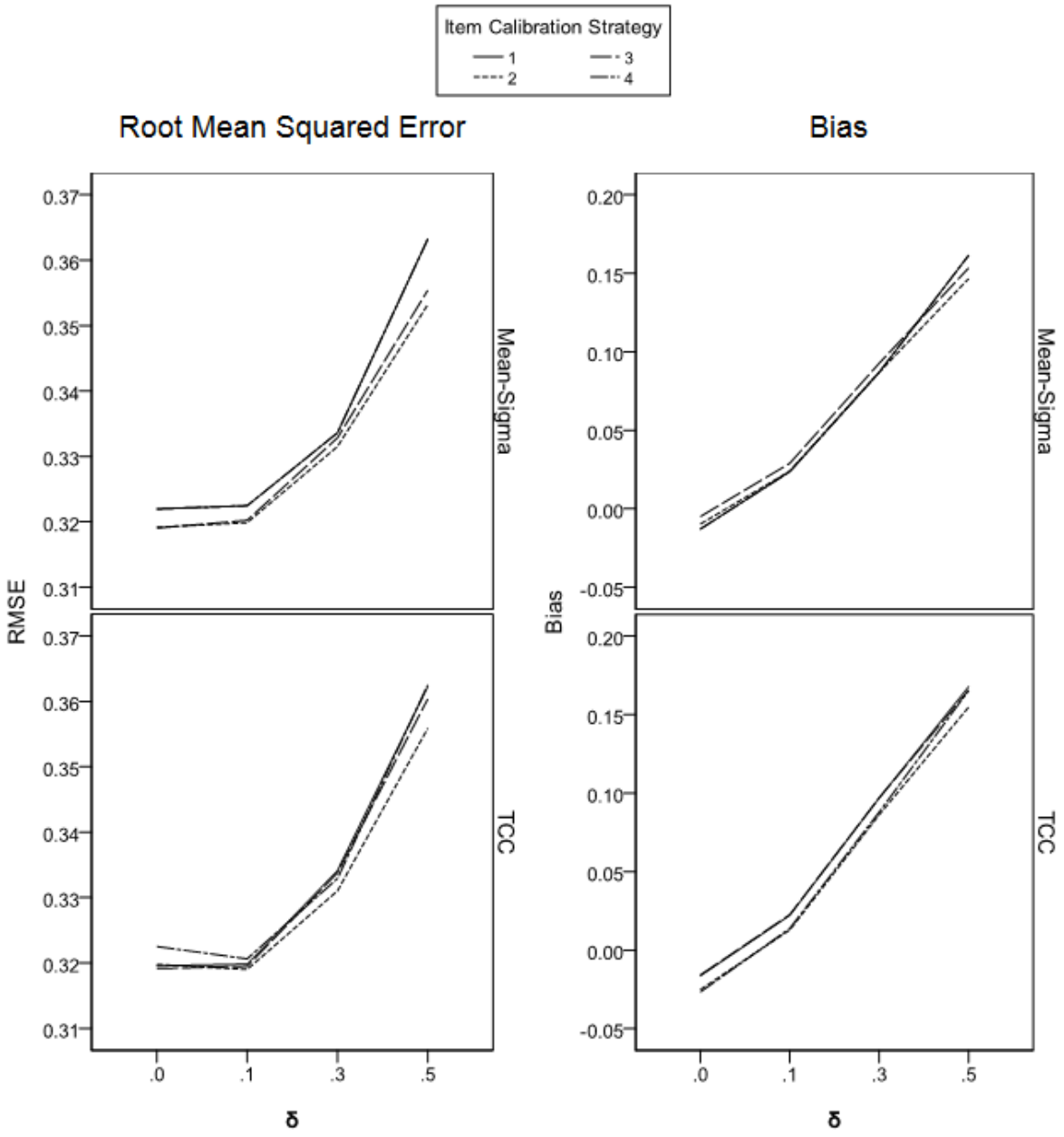

Figure 5. Root mean squared error (left) and bias (right) in proficiency estimation

We also investigated the distribution of the proficiency estimates. The mean proficiency estimate was sensitive to IPD and increased dramatically as IPD increased. A choice of scaling method and calibration strategy did not seem to influence the mean of the proficiency distribution, which again was what could be expected from the results shown in Figure 4 (middle). On the other hand, the influence of IPD on the standard deviation (SD) of the proficiency distribution varied depending on the choice of scaling method. The SD of the proficiency estimates was less influenced by IPD with the TCC method than with the mean-sigma method.

Finally, we classified the examinees into the four achievement levels (Below Basic, Basic, Proficient, and Advanced) based on the proficiency estimates to evaluate decision accuracy. The classification accuracy for each proficiency level is reported in Table 3. Overall, the classification accuracy was about 79% in the absence of IPD among the linking items. In practice, it is not unusual to see this level of classification accuracy with four proficiency categories

(i.e., three cut scores). As the magnitude of IPD increased, the overall classification accuracy decreased. Table 3 also shows the classification accuracy broken down for each proficiency level. Generally, the lower the examinee's proficiency level, the heavier the impact of IPD on the classification accuracy. For instance, for those whose true proficiency level was 'Below Basic,' the classification accuracy dropped to around 65% when δ = 0.5 regardless of the scaling method used. Among the four calibration strategies, Strategies 2 and 3 tended to yield slightly higher classification accuracy, but the difference was not very meaningful. For those whose true proficiency level was 'Advanced,' the classification accuracy increased as the IPD increased. Largely because the unidirectional IPD of this study resulted in positively biased proficiency estimates (Figure 5), the number of cases where examinees were falsely classified into the categories below 'Advanced' was reduced. It is interesting that the level of IPD had a major impact on the classification rates. The impact of IPD on the classification accuracy that we observed in this study, however, should not be imprudently generalized to other test programs that have different locations of cut scores, different examinee distributions, and/or different psychometric properties (for example, the test information function).

Summary and Conclusions

This study addressed three research questions, the first of which examined the effect of b-parameter drift on the *c*-parameter estimates. With the traditional item calibration strategy (Strategy 1), the *c*-parameter estimates showed a small change due to IPD. In the most problematic condition (33.3% of linking items with an IPD of 0.5 on the *b*-parameters), the mean change of *c*-parameter estimates due to IPD was about 0.02 (positive) regardless of scaling method and was

Table 3. Classification Accuracy for Each Level of Proficiency

Scaling Method	δ	Calibration Strategy	Below Basic (N = 683)	Basic (N = 1,444)	Proficient (N = 1,828)	Advanced (N = 1,045)	Overall (N = 5,000)
Mean-Sigma	0	1	81.0	74.4	80.6	81.8	79.1
		2	80.9	74.5	80.9	82.1	79.3
		3	80.6	74.4	80.8	82.6	79.3
		4	81.0	74.4	80.6	81.8	79.1
	0.1	1	77.9	73.0	81.0	83.8	78.8
		2	78.1	73.2	81.2	83.9	79.0
		3	77.9	73.0	80.9	84.4	78.9
		4	77.9	73.0	81.0	83.8	78.8
	0.3	1	72.1	69.7	80.6	87.6	77.7
		2	71.8	69.9	80.9	87.5	77.9
		3	71.7	69.6	80.4	88.1	77.7
		4	71.9	69.7	80.6	87.6	77.7
	0.5	1	63.4	63.7	79.0	90.5	74.9
		2	65.0	64.8	80.2	89.9	75.7
		3	65.1	64.5	79.4	90.6	75.5
		4	63.4	63.7	79.0	90.5	74.9
TCC	0	1	80.6	75.0	81.2	81.1	79.3
		2	81.2	75.3	81.2	80.4	79.3
		3	80.7	75.0	81.1	81.3	79.3
		4	81.5	74.8	80.7	80.5	79.1
	0.1	1	77.8	73.4	81.4	83.6	79.0
		2	78.4	73.8	81.6	82.7	79.1
		3	77.8	73.4	81.3	83.7	79.0
		4	78.7	73.5	81.1	83.2	79.0
	0.3	1	71.1	69.2	80.5	88.3	77.6
		2	71.8	69.9	81.0	87.5	77.9
		3	71.3	69.3	80.4	88.4	77.6
		4	72.2	69.6	80.4	87.9	77.7
	0.5	1	64.3	63.6	78.4	91.5	75.0
		2	65.1	64.2	79.5	90.7	75.5
		3	64.7	63.6	78.6	91.5	75.0
		4	64.2	63.7	78.4	91.2	74.9

smaller for the non-IPD items. It should be noted though that with a mean shift of .02 and the SD of the c-parameter estimates about .043, the effect size of the mean shift approached .50. The impact was significant.

For the second research question, we examined the effect of adjusting the *c*-parameter estimates on the scale stability. With calibration Strategies 2 and 3 where the *c*-parameters were either fixed to 0.20 or fixed to the estimates from the previous year, the *c*-parameters were not affected by the level of IPD. In addition to robustness against the uniform IPD, Strategies 2 and 3 offered other advantages over Strategy 1. Fixing *c*-parameters for the linking items to the estimates of the previous year (Strategy 3) seemed to be the most appropriate procedure because it automatically resulted in *c*-parameters that were identical across test forms. Since the *a*- and *b*-parameters were estimated with the *c*-parameters fixed, in practice, potential IPD effect would not be compounded in the *c*-parameter estimates. Calibration Strategy 2 nearly always showed results similar to calibration Strategy 3, but this approach, generally would be less suitable because model fit by fixing *c*-parameters to a value of .20 (or any other suitable constant) would be reduced in comparison with using estimates of the *c*-parameters (except in the case of smaller sample sizes and poor estimates of the *c*-parameter in the model).

With no adjustment to the *c*-parameters (Strategy 1), two different *c*-parameter estimates for each linking item become available. To handle such a situation, for example, Hambleton et al. (1991) suggested using the average of the two estimates of the *c*-parameter (Strategy 4). In principle, this solution also applies when concurrent item calibration is being used to link two test forms. In fact, our study results showed that Strategy 4 achieved more stable *c*-parameter estimation and was more robust against IPD than Strategy 1. The main problem with Strategy 4, however, is that this option is not always available. For example, it is hard to justify averaging the *c*-parameter estimates for each linking item across years. The averaged *c*-parameter estimates could be used in Year 2 but not with Year 1 results because scores already would have been reported.

Another practical advantage of calibration Strategies 2 and 3 over the other strategies is that their *a*- and *b*-parameter estimates became substantially more reliable (smaller *RMSE*) than those produced with the other strategies. As a result, Strategies 2 and 3 produced a moderate reduction in the standard error of the scaling efficient estimation. With extremely large sample sizes ($N > 5{,}000$), however, this advantage may be less meaningful because the standard errors of item parameter estimates and scaling coefficient estimates would already be very small. With a moderate sample size, however, calibration Strategies 2 or 3 offer the benefit of making more stable item parameter estimates available for the equating process.

The last research question addressed in this study compared the scaling methods. While the scaling methods were inconsequential for the most part, it appeared that Strategies 2 and 3 reduced classification errors by as much as 4%, an improvement of substantial practical significance.

When the 3PL model is used, it is vital to remember that the DIF/IPD on the item difficulty (i.e., uniform DIF/IPD) often affects the c-parameter estimation as well as the a-parameter estimation. Although c-parameters are often thought of as 'guessing' parameters, item difficulty estimates actually are influenced by guessing because examinees often provide answers based on their partial knowledge of item contents and distracters. Thus, when (uniform) DIF/IPD occurs on an item, not all of the DIF/IPD effect is reflected in the *b*-parameter estimates—some of the effect is absorbed by the *c*-parameter estimates.

Unfortunately, some IRT-based DIF/IPD detection methods cannot simultaneously evaluate the changes in multiple parameters due to DIF/IPD and are less powerful for detecting DIF/IPD. Moreover, when tests are equated, the *c*-parameter estimates, influenced by DIF/IPD, can also affect the equating results. Thus, situations where DIF/IPD is expected require solutions to control the *c*-parameter estimates across administrations. This study concentrated on the three different strategies (item calibration Strategies 2, 3, and 4) that have been observed in practice to make identical *c*-parameter estimates across multiple test occasions. Focusing on practical consequences rather than theoretical implications, we found all three of these strategies were effective in practice, with Strategies 2 and 3, but especially Strategy 3, deemed to be the best of the three choices.

This was a complex study, involving the manipulation of numerous variables and the complicated process of sorting through the findings about the roles of choice of equating method, strategies

for handling *c*-parameter estimates, and the level of IPD. Also, we knew that IPD's impact on the *c*-parameters was always going to be small to modest so when we carried out the study we expected to see, at most, a small but potentially practical impact. In fact, fixing the *c*-parameter estimates for linking items in the second test administration to the same values in the first administration offers a definite advantage for the quality of test score equating and model fit. The main finding from this study would suggest that, for protection against IPD, a change in current practices for handling *c*-parameter estimates is very much in order.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (chaps. 17-20). Reading, MA: Addison-Wesley.

Bock, R., Muraki, E., & Pfeiffenberger, W. (1998). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275–285.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education, 17*(3), 265–300.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*(1), 33–51.

Haebara, T. (1980). Equating logistic ability scales by weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J., (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457–459.

Han, K. T. (2012). Fixing the *c* parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation, 17*(1). Available online: http://pareonline.net/getvn.asp?v=17&n=1.

Han, K. T., & Guo, F. (2011). Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing. *GMAC Research Report RR-11-02*, January 1, 2011. Reston, VA: Graduate Management Education Council.

Han, K. T., & Wells, C. S. (2007, April). *Impact of differential item functioning on test equating and proficiency estimates*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale NJ: Erlbaum.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247–264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179–193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4.1: IRT item analysis and test scoring for rating-scale data* [Computer program]. Chicago, IL: Scientific Software, Inc.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned area between two item response functions. *Applied Psychological Measurement, 14,* 197–207.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63–84.

Stocking, M. L., & Lord, F. M. (1983). *Developing a common metric in item response theory. Applied Psychological Measurement, 7,* 201–210.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26,* 77–87.

Wollack, J. A., Sung, H. J., & Kang, T. (2006, April). *The impact of compounding item parameter drift on ability estimation.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Note:

The views and opinions expressed in this article are those of the authors and do not necessarily reflect those of the Graduate Management Admission Council® or the University of Massachusetts.

Acknowledgement:

The authors wish to thank Paula Bruggeman of GMAC® for review and valuable comments.

Citation:

Han, Kyung T., Wells, Craig S., & Hambleton, Ronald K. (2015). Effect of Adjusting Pseudo-Guessing Parameter Estimates on Test Scaling When Item Parameter Drift Is Present. *Practical Assessment, Research & Evaluation*, 20(16). Available online: http://pareonline.net/getvn.asp?v=20&n=16

Corresponding Author:

Kyung T. Han
Graduate Management Admission Council
11921 Freedom Dr. Suite 300
Reston, VA 20190

Khan [at] gmac.com