The Impact of Compounding Item Parameter Drift on Ability Estimation

James A. Wollack
Hyun Jung Sung
Taehoon Kang

University of Wisconsin—Madison
1025 W. Johnson St., #373
Madison, WI 53706

April 8, 2006

Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA

RUNNING HEAD: The Impact of Compounding Drift

The Impact of Compounding Item Parameter Drift on Ability Estimation

According to the invariance property of item response theory (IRT), item parameter values should be the same for all samples from a population. In practice, however, it is not always possible to satisfy the invariance property. Research has found that item parameters may be different for subgroups of examinees and across testing occasions. Change in parameter values for different subgroups is called differential item functioning (DIF; Holland & Wainer, 1993; Pine, 1977); change across time is called item parameter drift (IPD; Bock, Muraki, & Pfeiffenberger, 1988; Goldstein, 1983).

The literature on DIF is extensive. Throughout the 1990s, nearly two-thirds of the issues of *Journal of Educational Measurement* included at least one article pertaining to DIF. In contrast, the literature on IPD is quite small. One reason for this may be that the few studies that have been published have largely suggested that IPD is not as big a problem as the theory might lead one to believe. By and large, research on IPD has found that naturally occurring amounts and magnitudes of drift tend to have a very minor impact on the resulting ability scores. Wells, Subkoviak, and Serlin (2002) found that even when item discrimination (a) and item difficulty (b) parameters were increased by .5 and .4, respectively, for 20% of the items, item ability (θ) estimates were expected to deviate on the two tests by no more than 0.14 logits, for any true θ value. Similarly, Rupp and Zumbo (2003a, 2003b) found that examinees' scores were changed only slightly, unless the amount of simulated IPD was unusually large.

That IRT ability parameter estimation appears robust to even substantial amounts of IPD is of tremendous comfort to test developers who are charged with the task of equating forms from separate administrations and maintaining the test's score scale over time. Yet, the fact that

having a subset of items with considerably different item parameter estimates for two timepoints still results in reasonably similar ability estimates remains counterintuitive.

In both Wells et al. (2002) and Rupp and Zumbo (2003a, 2003b), the impact of IPD was studied across two occasions. Yet, in practice, IPD is a phenomenon that has typically been examined (and is most relevant when considered) over multiple testing occasions. Bock et al. (1988) studied IPD over a 10-year period on the College Board English and Physics Achievement Tests. Chan, Drasgow, and Sawin (1999) studied IPD over a 16-year period on the Armed Services Vocational Aptitude Battery. Veerkamp and Glas (2000) modeled the effects of item over-exposure within computerized adaptive testing by analyzing changes in item difficulty parameters across 25 simulated intervals within a testing window. Recently, DeMars (2004a, 2004b) examined patterns of IPD over four years on one test of U.S. History and political science and a second test of information literacy. And Wollack, Sung, and Kang (2005) looked for effects of IPD over six years on a college-level German placement test.

Therefore, for test developers to be comfortable employing methods that largely ignore any potential IPD, it must first be demonstrated that IRT ability estimation is robust to IPD over a multiple-year period. One concern is that the item drift problem may compound over time, as the number of drifting items and the magnitude of drift increases, particularly if drifting items are included in the linking of test forms (Kim & Cohen, 1992; Lautenschlager & Park, 1988; Shepard, Camilli, & Williams, 1984). In this study, we used both simulated data and data from a college-level test of information literacy to examine the effect of compounding IPD on a score scale over a multiple-year period. The impact of IPD on examinee ability and on item parameters was studied under eight different IRT linking designs.

Simulated Data Analysis

Data Source

The design for the simulated datasets is shown in Figure 1. In this study, we simulated a testing program that, each year, administers an alternate form of a 60-item test. In addition to the set of scored items, a set of 10 pilot items was simulated for each of the first four years. Each new form consisted of a randomly sampled 50 items from the preceding year's operation items, plus all 10 of the items that were pilot tested on the preceding form. Therefore, forms were designed so that each new form consisted entirely of items that were administered the previous year. In this way, five forms were constructed, resulting in the simulation of 100 items. Thirty-five of the 100 items appeared on all five test administrations; twenty-nine items were operational in all five years.

Insert Figure 1 About Here

Item parameters from the three-parameter logistic model (3PLM) were generated randomly for the 100 items from the following distributions:

$\alpha \sim$ Lognormal (0, 1)

$\beta \sim$ Normal (0, 1)

$\gamma \sim$ Logit normal(-1.4, 0.3).

These generating item parameters defined the base scale.

Simulation of IPD. Both compounding IPD (CIPD) and random IPD (RIPD) were simulated. The CIPD condition modeled situations where the item difficulty for some items changes systematically over time, as might be the case when curriculum is changing or when items have been over-exposed. CIPD was simulated by adding a constant δ to the item difficulty

parameter from the preceding year each time the item drifted, and 0.0 to the preceding item difficulty each time the item did not drift. For example, under CIPD, an item *i* that drifted each of the first two years before stabilizing in the fourth and fifth years was simulated with its base value, b_i, for year 1 (Y1), $b_i + \delta$ for year 2 (Y2), and $b_i + 2\delta$ for years 3 (Y3), 4 (Y4), and 5 (Y5).

RIPD was intended to model situations where certain items occasionally function differently in a particular year, though their long range difficulty estimates remain unchanged from their base values. Items selected to exhibit RIPD were simulated with difficulties equal to $b_i + \delta$ for that particular year only. In years where RIPD was not simulated for an item, the generating difficulty value was fixed at b_i. To illustrate, an item showing the same drifting pattern as above, only with RIPD—drifting in Y2 and Y3, but not in Y4 and Y5—would have been simulated with difficulty values equal to b_i for Y1, $b_i + \delta$ for Y2 and Y3, and b_i for Y4 and Y5.

Ten items exhibiting CIPD were randomly selected from the 35 items appearing on all five forms. Three of these items were simulated to drift four times (i.e., in each of years 2, 3, 4, and 5); four items drifted three times, and three items drifted twice. In addition, each year had 10 RIPD items, which were randomly selected from the 50 items not already chosen to have CIPD. Therefore, on any given form, approximately one-third of the items were simulated as drifting from the base scale.

By manipulating δ, it was possible to study the effect of different magnitudes of IPD on item and ability parameter estimation. Both moderate and large values of δ were used. In one condition, δ was fixed at 0.25, simulating a moderate amount of IPD. In the large IPD condition, δ was equal 0.40. This is the same magnitude of IPD used by Wells et al. (2002) and by

Donoghue and Isham (1998). Although all items were modeled with the 3PLM, a_i and pseudo-guessing (c_i) parameters were simulated to remain invariant across testing occasions.

All IPD in this study was simulated by adding a positive constant to the item difficulty parameters. Clearly, this is an oversimplification of how IPD actually occurs in practice, but was selected because it represents a worst-case-scenario for item and ability parameter estimation. Similar techniques have previously been used to study the effects of IPD (e.g., Wells et al., 2002).

Simulation of item responses. Item response vectors were simulated for 1,000 examinees in each of the five years. θ values for year y were randomly drawn from a Normal ($\Delta(y-1)$, 1) distribution, where $\Delta = 0$ or 0.15, and y equaled 1, 2, 3, 4, or 5. Therefore, θ was generated under two separate ability trend conditions. When $\Delta = 0$, examinee θ parameters were sampled randomly from independent Normal (0, 1) distributions for all y. When $\Delta = 0.15$, however, a linear trend in ability was simulated, with each year becoming 0.15 logits higher in ability. Under this condition, therefore, Y1 θs were generated from a Normal (0, 1) distribution, but Y5 θs were generated from a Normal (0.6, 1) distribution.

The magnitude of IPD ($\delta = 0.25$ or $\delta = 0.40$) was crossed with the ability trend ($\Delta = 0.0$ or $\Delta = 0.15$), to create four experimental conditions. Five replications of each condition were performed.

Methods

Data from Y1 were used to calibrate item parameters under the 3PLM, using the computer program MULTILOG 7.0 for Windows (Thissen, 2003). To improve the quality of item parameter recovery, the following priors were placed on the item parameter values, as per the recommendations of Orlando and Thissen (2000):

$$a \sim N(1.9,1)$$

$$b \sim N(0,1.5)$$

$$c \sim \text{logit normal}(-1.1, 0.5).$$

In addition, the maximum number of EM cycles was set at 1,000 to improve the chances of the solution converging. Default values were used for all remaining settings.

Linking Techniques. Test forms were linked using one of two methods: the fixed item parameters method and the test characteristic curve (TCC; Stocking & Lord, 1983) method.

When linking using fixed item parameters, parameters for all operational items are constrained to equal their scale values, based on a previous administration. Parameters for any items that have not been previously administered, i.e., pilot items, are freely estimated.

When linking using the TCC method, parameters for all years are independently estimated. Estimates for each new year are placed onto the Y1 metric by transforming the parameter estimates from the new year, such that $a^* = a/A$, $b^* = Ab + K$, and $\hat{\theta}^* = A\hat{\theta} + K$, so as to minimize the difference between the Y1 TCC and that based on the newly transformed parameter estimates. Note that the estimation of the A and K coefficients is done using a subset of the items referred to as the anchor items. At a minimum, these items must be common to the two forms being linked, but may also possess other desirable properties, such as being representative of the test blueprint or being free from DIF or IPD. Items not included in the linking set are not used to estimate A and K, but are still transformed to the base metric after A and K have been determined. The computer program EQUATE (Baker, 1990) was used to estimate the A and K coefficients for the TCC method.

Linking Methods. When applying the two linking techniques above, new forms may be linked directly to the original Y1 metric through the common items, or indirectly to the Y1

metric by linking to the immediately preceding form (which was linked to its predecessor, and so on back to the Y1 metric). An illustration of the distinction between the two types of linking is provided in Figure 2, for a hypothetical three-year period.

Insert Figure 2 About Here

Direct linking is advantageous because it minimizes linking error between forms and links directly to the scale of interest. However, there are potential problems with directly linking over multiple years, as is illustrated in Table 1. Table 1 shows a hypothetical testing program which administers a different form of the test each year over a five year period. In this example, each form contains 10 operational items (numbered 1-10) and four pilot items (numbered 11-14). Given the design in Table 1, indirect linking will always involve 10 items because all operational items were selected from the previous form. With direct linking, however, the number of items available (shown in boldface type in Table 1) decreases with each new form. For example, a direct linking between Forms 1 and 5 would involve only three items. Therefore, a testing program that uses direct linking will need to make a concerted effort to retain an adequate number of items from the base year in each subsequent form, but doing so may lead to IPD problems due to over-exposure. Therefore, while direct linking minimizes linking error between forms by linking directly to the scale of interest, indirect linking will result in more items on which to link. Indirect linking may also be attractive because adjacent forms are expected to be more similar with regard to the behavior of individual items.

Insert Table 1 About Here

Treatment of Drift Items. Including into the linking process items that exhibit IPD may be problematic (Kim & Cohen, 1992; Lautenschlager & Park, 1988; Shepard, Camilli, & Williams, 1984). Therefore, linking methods were examined both under conditions where items were not tested for IPD, and conditions where items were first tested for IPD using the likelihood ratio (LR) test for differential item functioning (DIF; Thissen, Steinberg, and Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993) with concurrent calibration. Although only item difficulty coefficients were manipulated to simulate IPD, it is conceivable that simulated difficulty drift manifests itself in changes in item discrimination or pseudo-guessing estimates. Therefore, augmented models allowed all three parameters for the tested item to be freely estimated. Items with LR χ^2 values that were greater than or equal to $\chi^2_{3,.95} = 7.82$ were treated as drifting. Any items found to have drifted were not included among the linking set, but instead had their parameter values re-estimated. In conditions where items were not tested first for IPD, linking was based on all common items between the two years.

Models. Eight different models were used to link subsequent forms to the Y1 metric. The different models varied with respect to linking technique (Fixed vs. TCC), linking method (Direct vs. Indirect), and the way in which drifting items were treated (IPD Testing vs. No IPD Testing). A summary of these models is provided in Table 2.

Insert Table 2 About Here

Outcome Measures. To investigate the impact of IPD and the linking model on ability estimation, Y5 scale means, standard deviations, biases, and root mean squared errors (RMSE) for each of the eight models were compared. In addition, the same outcome variables were assessed for Y5 item parameters.

Results

Ability estimation. Sample $\hat{\theta}$ means under the eight models for Y5, after being linked to the Y1 scale, are given in Table 3, averaged over the five replications. The first column presents the generating mean θ value for Y5, averaged across replications. Table 4 presents Y5 sample pooled standard deviations for the eight models, pooled over the five replications. Again, the pooled standard deviation among generating values is presented in the first column. Means and standard deviations are reported separately for the four ability trend ($\Delta = 0.0$ or $\Delta = 0.15$) x magnitude of drift ($\delta = 0.25$ or $\delta = 0.40$) conditions.

Insert Tables 3-4 About Here

Several patterns are clear from Tables 3 and 4. For all models, the true Y5 mean and standard deviation were underestimated. For the means, the amount of underestimation tended to increase as Δ and δ increased. The standard deviation estimates, however, were quite stable across the four different conditions. Finally, by examining the final column in Table 3, which presents the standard deviation among the average $\hat{\theta}$ for the eight linking models, it is clear that the between-model differences in average $\hat{\theta}$ increased as Δ and δ increased. The standard deviation among the models in Condition 4 was more than twice as large as the standard deviation in Condition 1. There was no noticeable difference between conditions with respect to between model standard deviations of $\hat{\theta}$.

Average correlations between θ and $\hat{\theta}$ were either 0.93 or 0.94 for all models and all conditions. Furthermore, inter-model $\hat{\theta}$ correlations were 1.0 between all pairs of models. That

all models were a linear transformation away from being virtually identical suggests that all differences between the models were attributable to differences in the linking procedures.

Average biases and RMSEs are presented Tables 5 and 6, respectively. Data in part (a) of the table report on the four different generating conditions, whereas data in parts (b) and (c) show the two main effects, magnitude of drift and ability trend, respectively.

Insert Tables 5-6 About Here

The data from Table 5a mirror the results of Table 3: all models showed some negative bias and the amount of bias increased as Δ and δ increased. What is also evident from Table 5a is that the amount of bias varied considerably as a function of model choice and condition. Bias ranged from as little as -0.04 for Model 5 in Conditions 1 and 2 to as high as -0.52 for Model 2 in Condition 4. Overall, the average bias was nearly three times greater in Condition 4 (-0.31) than in the Condition 1 (-0.11).

Table 6a also revealed differences in terms of RMSE, though the differences were generally less pronounced. Average RMSEs were higher for Condition 4 than for Conditions 1-3. However, 26 of the 32 RMSE values were between 0.38 and 0.46, indicating that the order of magnitude was similar for most values.

From Tables 5b and 6b, one can see that the magnitude of IPD had a moderate effect on bias, but very little effect on RMSE. In general, the amount of negative bias increased as the magnitude of IPD increased. Bias and RMSE were lower in Models 5-8 than in Models 1-4. This result is not surprising, since IPD testing was incorporated into Models 5-8, but not in Models 1-4. Bias and RMSE were smallest for Model 7, regardless of the amount of IPD.

Table 5c shows a large main effect for ability trend in terms of bias. A moderate main effect is also shown for RMSEs in Table 6c. When the mean of the ability distribution remained stable over time (i.e., Conditions 1 and 2), Models 5 and 6, both of which involved IPD testing and fixed item parameters, produced the smallest biases and RMSEs. When $\Delta = 0.0$, Model 7, which involved direct TCC linking with IPD testing, produced biases and RMSEs that were almost identical to those of Model 6. However, when the mean of the ability distribution increased over time (Conditions 3 and 4), Models 5 and 6 performed much less well than Model 7. In the $\Delta = 0.15$ conditions, Model 7 was clearly the best at recovering underlying parameters.

The effects of linking technique, IPD testing, and linking method are shown in Table 7 for bias and Table 8 for RMSE. From Tables 7a and 8a, one can see that, over all conditions, there was less bias and smaller RMSEs in θ estimates when IPD testing was utilized prior to linking. The main effects for linking technique and linking method were quite small. However, it was often the case that differences in bias or RMSE between levels of linking technique, IPD testing, or linking method changed as Δ and δ varied. The extent of these changes are illustrated by examining the interaction effects shown in Tables 7b-d for bias and 8b-d for RMSE. As an example, from Table 7b, we see that the difference in bias between TCC and fixed linking was -0.14 − -0.09 = -0.05 for Condition 1, but was -0.25 − -0.37 = 0.12 for Condition 4.

Insert Tables 7-8 About Here

Tables 7b and 8b show that linking technique was relatively unaffected by the magnitude of IPD, but was affected noticeably by an upward trend in ability. Furthermore, fixed linking was more influenced by changes in the ability distribution than was TCC linking. The effect of IPD testing on bias (Table 7c), on the other hand, was increasingly important as the magnitude of

IPD increased, but was not especially sensitive to whether there was a trend in ability. A corresponding pattern for RMSE (Table 8c) was not apparent. Finally, there was no evidence of significant interaction effects involving direct versus indirect linking (see Tables 7d and 8d).

Item parameter estimation. Sample item difficulty means for Y5 under the eight models are given in Table 9, averaged over the five replications. One will note that the true mean item difficulty value, presented in the first column, is common for fixed values of δ, and is larger for Conditions 2 and 4 (where $\delta = 0.40$). Similarly, Table 10 presents the item difficulty pooled standard deviations for the eight models.

Insert Tables 9-10 About Here

Item difficulty values were, on average, underestimated. The variance among models in item difficulty estimates was considerably larger when $\delta = 0.40$ than when $\delta = 0.25$. In addition, the estimates were less variable than the parameter values.

Tables 11 and 12 present the average biases and RMSEs, respectively, for item difficulty. Examining these tables, it is clear that the models are not equally adept at recovering underlying item parameter values. Models 1-3 all showed considerable bias, each producing average biases of at least -0.14. Model 7, on the other hand, was unbiased, and also consistently produced the smallest RMSE values. Furthermore, as Δ and δ increased, bias increased for all models except Model 7 (which remained essentially unbiased in all conditions) and Model 8 (which showed approximately the same magnitude of bias in all conditions). Patterns of RMSEs showed that the variability of estimates of item difficulty were particularly affected by values of Δ and δ in Models 1 and 2. Linking method appeared to have only a very small effect, although it is

important to realize that biases in the direct linking condition were consistently less than or equal to those for indirect linking.

Insert Tables 11-12 About Here

The effects of magnitude of IPD and ability trend on item difficulty estimation in model are shown in panels (b) and (c) of Tables 11 and 12. As was found with ability estimation, Models 5-8 were all affected very little by increasing magnitudes of simulated IPD. In contrast, Models 1-4 all showed increased bias, and Models 1-2 showed increased RMSEs, in Conditions 2 and 4.

Models using TCC linking (i.e., Models 3, 4, 7, and 8) were unaffected by an ability trend, whereas models using fixed linking (i.e., Models 1, 2, 5, and 6) all showed increased bias when the mean of the ability distribution increased over time. Interestingly, this same pattern was not observed for ability estimation, although the magnitude of the bias was less in the TCC conditions than in the fixed linking conditions. Ability trend appeared to have very little effect on the RMSEs among difficulty estimates for any of the models.

The item difficulty effects of linking technique, IPD testing, and linking method are shown in Table 13 for bias and Table 14 for RMSE. Panel (a) shows that both linking technique and IPD testing had important effects on the bias and RMSE of item difficulty estimates. The impact of Δ and δ on these effects are shown more fully in panels (b) - (d). Table 13b shows little difference between the fixed and TCC linking procedures when $\Delta = 0.0$, but a rather large difference (in favor of the TCC method) when $\Delta = 0.15$. Similarly, Tables 13c and 14c show that IPD testing becomes more important as the magnitude of IPD increases. The main and interaction effects for linking method (Tables 13d and 14d) were all quite small.

Insert Tables 13-14 About Here

The same outcome measures reported for item difficulty were also estimated for item discrimination and item pseudo-guessing parameters. Biases and RMSEs of a_i and c_i were virtually identical in all four Δ x δ conditions, and were not noticeably or systematically different across any of the eight linking models. Therefore, tabled results for a_i and c_i are not shown here. Within the context of this study, this result is not surprising because IPD was simulated by manipulating the b_i values only. Had the a_i or c_i also been allowed to exhibit drift, it is possible that some differences across models or simulating conditions would have emerged.

Summary of Simulation Study

When item parameters drift over multiple years, the corresponding effects may compound in a way that has important implications for ability estimation. This simulation study examined the effectiveness of different item linking strategies at recovering underlying model parameters in the face of varying magnitudes of IPD and the presence or absence of a linear trend in the mean of the ability distribution. Unlike previous studies (e.g., Wells et al., 2002) which showed that, between two years, bias in ability estimates did not exceed 0.14, even when a considerable amount of IPD was present, in this study, we studied the compounding effects of IPD over a multiple-year period and found that ability biases that were substantially larger. In fact, for some models in certain conditions, biases were as large as -0.52 (see Model 2 in Condition 4). Also, differences between models within a condition were substantial. The ranges (i.e., absolute value of the maximum minus the minimum) in biases among the eight models were .15, .24, .23, and .39 for Conditions 1-4, respectively. Similar patterns were found with item difficulty estimates.

There were substantial differences between models using fixed and TCC linking, and between those that tested for IPD before linking and those that did not. Models that tested first for IPD were uniformly better, but the difference became more pronounced as δ increased from 0.25 to 0.40. As an example, when $\delta = 0.40$, bias increased an average of just .02 for Models 5-8 over its value when $\delta = 0.25$, compared to an increase of .09 for Models 1-4.

There were similar differences observed when the mean of the ability distribution increased over time. Fixed linking appeared to be a reasonable strategy provided the mean of the ability distribution was the same for different test forms. When the mean increased across time, however, fixing item parameters resulted in much more biased estimates than were found using the TCC method. In fact, average biases for the fixed linking models were higher by 0.22 when an ability trend existed. By comparison, average biases increased, also, in the TCC models, but only by 0.05.

The effect of linking method was small. However, what patterns were discernible suggested that direct linking is to be preferred over indirect linking. Of course, unless the base test consists of many items, one ramification of direct linking is that the exposure rates of linking items will be higher than might otherwise be desired. This in turn may cause unusually large magnitudes of IPD or large numbers of drifting items among the anchor set. This study did not examine any possible effects caused by anchor items that are more likely to be affected by IPD.

Overall, Model 7 appeared to best reproduce the underlying ability parameters. In the $\Delta = 0.0$ conditions, Models 5 and 6 actually produced less bias than Model 7; however, the differences between Models 5 and 6 and Model 7 were slight, and the overall amount of bias in Model 7 was still fairly small. In the $\Delta = 0.15$ conditions, on the other hand, Model 7 was much preferred to Models 5 and 6. Because it can never be known for certain how much items will

drift and whether the ability distribution over time is changing, the results of this study suggest that it is safest to use direct TCC linking with IPD testing.

The analysis of item difficulty was very similar to that for ability. Model 7, again, recovered parameters better than competing models. In fact, Model 7 recovered difficulty parameters without (or with only the most minimal amounts of) bias for all conditions. Given that item parameters were recovered so well, it is a bit surprising that ability parameters were not recovered better. Though this remains an area for future study, one possible explanation is that the true pattern of IPD was not well recovered, leading to incorrect sets of items being included in the anchor set. Although Models 5-8 included testing for IPD, it was often the case that items simulated to drift were not correctly identified (Type II errors), whereas items that were not simulated to drift were identified (Type I errors). For example, in Model 7, the Type I error rates were 0.19, 0.38, 0.30, and 0.36 for Conditions 1-4, respectively. Consequently, many items that should have been used as anchors (i.e., their parameter values had not changed) were excluded because of false positive results on the LR test. At the same time, power in Model 7 was 0.30, 0.43, 0.38, and 0.50 in Conditions 1-4, respectively. This suggests that, in all conditions, at least half of the drifting items were not detected as drifting and were included among the anchor set. These items, though they demonstrated empirical effects that were too small to be detected, may collectively have caused small amounts of bias in ability estimation, particularly when one considers that all IPD was simulated in the same direction. If it is true that misidentifying drifting items may cause ability estimates to be biased, it could further explain why direct linking outperformed indirect linking. It is possible that choosing a different IPD detection method, such as Lord's χ^2 with a common c_i, might improve the pattern of correct and incorrect detections (Donoghue & Isham, 1998).

A Real-Data Example

Data Source

The Information Seeking Skills Test (ISST) is a 53-item computer-delivered exam developed by reference librarians to assess information literacy among college students. Over a four year period (ranging from the 1998-99 to 2001-02 academic years), 8,721 students at James Madison University completed the assessment (353 students the first year, 2,671 the second year, 2,741 the third year, and 2,956 the fourth year). The same form of the ISST was used for all four years. One item was dropped because of poor statistics.

Beginning in the Fall of 1999, it became required of students at James Madison University that they take and pass the ISST during their first year. Due to the changes in availability and use of technology on campus, in addition to the change in stakes between the first and second year, it was assumed that the behavior of some of the items on this test might change over this four-year period. In fact, Demars (2004) found several of the items on this test included systematic linear IPD.

Methods

The same eight linking models were used to link the forms across all four year. Year 4 means and standard deviations were computed for all model parameters under each of the models, and correlations and root mean square differences were computed for all parameters between all models. To evaluate the impact of any differences on examinees, pass rates were also examined for a variety of hypothetical cut-scores.

Note that because all items were common across the four years and Models 1 and 2 fix the parameter estimates for all common items, results for Models 1 and 2 are identical in this real data analysis.

Results

Means and standard deviations for all model parameters are shown in Table 15 for the eight linking models. For the most part, the models performed fairly similarly with regards to ability estimation. Model 3 produced the highest mean $\hat{\theta}$ and the lowest variability, whereas Model 6 produced the lowest mean $\hat{\theta}$ and the most variability.

Insert Table 15 About Here

The eight models varied more in terms of item parameter estimation. Average item a_i values ranged from 0.74 (Model 6) to 1.01 (Model 4). Standard deviations among a_i values were very similar, with the exception of that for Model 6 which was slightly higher. Average b_i values ranged from -0.96 (Model 6) to -0.49 (Models 1 and 2) and standard deviations ranged from 0.89 (Model 4) to 2.01 (Models 1 and 2). Although these differences appear large, these numbers are somewhat spurious. One item in the Year 1 calibration sample was estimated to have a difficulty of 12.18. In Year 4, however, the item difficulty estimates for this item from Models 3-8 ranged from 2.61 (Model 4) to 4.70 (Model 5). Because Models 1 and 2 use fixed item parameters and no IPD testing, the item difficulty was 12.18 for both Models 1 and 2. Therefore, although there was considerable variance among the item characteristic curves for this item, all looked reasonably similar for $-2 \leq \theta \leq 2$. The estimation of pseudo-guessing parameters was very similar across models.

Table 16 shows the average differences and RMSDs between linking models for $\hat{\theta}$, $a_{i,}$ and b_i. Average differences and RMSDs for c_i were uniformly very small, so are not given here. Most models were quite similar, on average, with respect to ability estimation (Table 16a). The largest mean difference was 0.16, between Models 3 and 6. Model 7, which worked best in the

simulation study, differed by no more than 0.11 with any model (Model 6). RMSDs were similarly modest, never exceeding 0.23 (between Models 3 and 6). The largest RMSD between Model 7 and another model was 0.17 (Models 1 and 2).

Insert Table 16 About Here

Item discrimination values (Table 16b) were estimated similarly across models, with the possible exception of Model 4, which consistently estimated items with higher discrimination. All average differences among other models were no greater than 0.14, though RMSDs did get as large as 0.35.

Item difficulty estimates were different for sets of models. In particular, Models 5 and 6 appeared similar to each other, but rather different from the remaining models. Models 1, 2, 3, 4, 7, and 8 appeared, on average, to recover similar estimates of item difficulty. However, RMSDs for Models 1 and 2 were appreciably higher than those for other models. This is likely due to the one extreme item mentioned earlier, which had its difficulty estimated at 12.18 for Models 1 and 2, but no higher than 4.70 for any other model.

Correlations between $\hat{\theta}$ values under the eight models were extremely high, ranging from 0.97 – 1.00. This suggests, in addition to the $\hat{\theta}$ values being of similar magnitude (as was shown in Tables 15 and 16), the rank-ordering of examinees is also very similar.

Correlations among item parameters were lower, and are given in Table 17 for the a_i (upper triangle) and b_i (lower triangle) values. Item parameter estimates for the four TCC models (Models 3, 4, 7 and 8) all correlated perfectly with each other. Though the data are not shown, $\hat{\theta}$ values for these four models also correlated perfectly. This result is expected because the four models apply a linear transformation on $\hat{\theta}$, a_i, and b_i. The transformation coefficients will differ

for each of the models, but the transformation will always be linear and will be applied to all items and people, and correlations are invariant to linear transformation. Correlations between Models 5 and 6 and Models 3, 4, 7, and 8 were strong. Correlations between Models 1 and 2 and other models—particularly the four TCC models—were fairly modest. Correlations for the c_i values are not provided because (a) they were estimated well by all models, and (b) there was too little variance in their estimated values.

Insert Table 17 About Here

The impact of the differences between models can be illustrated by examining pass rates for a variety of different hypothetical cut scores. Table 18 shows the percentage of students who would have passed the ISST for six different cut scores, ranging from $\theta = -0.25$ to 1.00, in increments of 0.25. This particular θ range was selected because for values outside this range, there was very little difference among the models. Within this range, however, there are some noteworthy differences. Over this range of potential cut scores, Model 3 uniformly produced the highest passing rates. The passing rate for Model 6 was lowest for $\theta \leq 0.25$. Models 1 and 2 were lowest for $\theta \geq 0.75$. And Models 1, 2, and 6 were tied for lowest at $\theta = 0.50$.

Insert Table 18 About Here

Although differences in $\hat{\theta}$ values were small, the ramifications of these differences could be rather large, depending on the location of the cut score. For $0.00 \leq \theta \leq 0.75$, the maximum difference between passing rates was at least 0.10, and in some cases was as high as 0.15. For example, if the cut score during Year 4 had been set at $\theta = 0.50$, approximately 1,508 students

would have passed under Models 1, 2, or 6, approximately 1,744 students would have passed under Model 7, and approximately 1,951 students would have passed under Model 3. The difference (between Models 3 and 6) represents a total of 443 students; these are the students who would pass under some models, but not under all models.

Summary of Real-Data Study

This study used four years' worth of data from the ISST to demonstrate the effects of IPD and the linking model on ability and item parameter estimation. Relatively small differences were observed between $\hat{\theta}$ values from the eight models, but these differences were neither trivial nor inconsequential. Depending on the location of the cut score, the passing rates for the models differed by as much as 0.15. Important differences between models in passing rates existed over about 1.25 logits on the ability scale.

Differences in item parameter estimates were generally small, though there were some notable exceptions. Though, on average, Models 1 and 2 produced item parameter estimates that were fairly similar to those from other models, there was also quite a bit more variability in those estimates than with other models. Also, the RMSDs between Models 1 and 2 and other models were quite large, indicating that there were some items for which there were rather large differences. In addition, Models 5 and 6 produced difficulty estimates that were noticeably easier than those from other models.

It may be comforting to know that, among the eight models, parameter estimates for Model 7, which was best at recovering parameters in the simulation study, tended to be in the middle, rather than consistently higher or lower than the others. Therefore, although the differences between the most and least extreme models was substantial, the differences between each model and the best model were more reasonable.

Conclusions

The presence of unaccounted for IPD holds potential to negatively affect the linking process, possibly resulting in spurious estimates of examinee ability. Although previous research has shown that IRT ability estimation is robust to the presence of normally occurring amounts and magnitudes of IPD, studies have not fully investigated this longitudinally, where IPD and linking errors may compound over time.

In this study, eight different models for linking and accounting for IPD were considered and applied to both simulated data and a test of information literacy over multiple years. The results of this study showed that choice of linking/IPD model can have a large effect on the resulting $\hat{\theta}$, as well as on passing rates. Models that tested items for IPD and excluding drifting items from the linking process tended to recover the underlying parameters better, particularly as the magnitude of IPD increased. Models that used TCC linking also performed better than models that fixed the parameter values for anchor items, particularly for conditions where the mean of the ability distribution increased over time. Finally, directly linking each new form to the base form was slightly preferable to indirectly linking it; however the ramifications of potentially increasing the amount and/or magnitude of IPD among the linking items due to over-exposure remains to be addressed. Model 7, which directly linked to the base metric after first removing drifting items, performed fairly similarly across simulated conditions, and was consistently the best or among the best of the models.

Although the simulation study demonstrated that the treatment of drifting items and the linking technique can affect ability estimation, the real data analysis showed that, in practice, the differences may be less pronounced. There are several reasons that this may be the case. About one-third of the items each year were simulated as drifting by increasing the item difficulty by

0.4 logits. In practice, it is unlikely that all drifting items will drift in a common direction. Some may become easier, while others become harder, thereby essentially canceling each other out. Furthermore, not all drifting items will change by as much as 0.4 logits. However, some items may have even larger magnitudes of IPD, particularly if they are administered in two very different locations in the test (Oshima, 1994), as might be the case if end-of-test items are reserved for pilot testing. Finally, having as many as one-third of the items drift between two years may be higher than is encountered in many testing programs.

References

Baker, F. B., Al-Karni, A., & Al-Dosary, I. M. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement, 50,* 529-549.

Bock, R., Muraki, e., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.

Chan, K,-Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84*, 610-619.

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

DeMars, C. E. (2004b, April). *Item parameter drift: The impact of the curricular area.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20,* 369-377.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*, 51-66.

Lautenschlager, G. J., & Park, D.-G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement, 12*, 365-376.

Orlando, M. & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models, *Applied Psychological Measurement, 24*, 50-64.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200-219.

Pine, S. M. (1977). Application of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Application of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association* (Research Rep. No. 77-1, pp. 37-43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Rupp, A. A., & Zumbo, B. D. (2003a, April). *Bias coefficients for lack of invariance in unidimensional IRT models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Rupp, A. A., & Zumbo, B. D. (2003b). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *The Alberta Journal of Educational Research, XLIX*, 264-276.

Shepard, L., Camilli, G., & Williams, D M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*, 93-128.

Stocking, M., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 207-210.

Thissen, D. (2003). *MULTILOG 7.0: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago, IL: Scientific Software, Inc.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Veerkamp, W. J. J., & Glas, C. A. W (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics, 25*, 373-390.

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. (2005, April). *Longitudinal effects of item parameter drift*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Table 1

Illustration of Direct Versus Indirect Linking

Form 1	Form 2	Form 3	Form 4	Form 5
1	**1**			
2	**2**	**1**	**1**	
3				
4	**3**	**2**		
5	**4**			
6				
7	**5**	**3**	**2**	**1**
8	**6**	**4**	**3**	**2**
9				
10	**7**			
11	**8**	**5**	**4**	**3**
12	**9**	**6**		
13	**10**	**7**	**5**	
14				
	11	8	6	4
	12	9		
	13	10	7	5
	14			
		11	8	6
		12		
		13	9	7
		14	10	
			11	8
			12	9
			13	
			14	10

Table 2

Description of Linking and Drift Models Studied

	Common-Item Linking Technique		Linking Method		Drift Testing	
	Fixed Item Parameters	TCC Method	Directly with Y1	Indirectly with Y1	NO	YES LR Test
Model 1	X		X		X	
Model 2	X			X	X	
Model 3		X	X		X	
Model 4		X		X	X	
Model 5	X		X			X
Model 6	X			X		X
Model 7		X	X			X
Model 8		X		X		X

Table 3

Y5 Mean $\hat{\theta}$ for Linking Models

	Linking Models									
	True	1	2	3	4	5	6	7	8	SD
1. $\Delta = 0.0, \delta = 0.25$	-0.01	-0.12	-0.15	-0.20	-0.16	-0.05	-0.09	-0.09	-0.14	0.05
2. $\Delta = 0.0, \delta = 0.40$	0.00	-0.15	-0.22	-0.28	-0.23	-0.04	-0.08	-0.10	-0.17	0.08
3. $\Delta = 0.15, \delta = 0.25$	0.58	0.30	0.22	0.35	0.38	0.34	0.36	0.44	0.42	0.07
4. $\Delta = 0.15, \delta = 0.40$	0.61	0.21	0.09	0.26	0.32	0.32	0.35	0.48	0.39	0.12

Table 4

Y5 Pooled Standard Deviations of $\hat{\theta}$ for Linking Models

	Linking Models									
	True	1	2	3	4	5	6	7	8	SD
1. $\Delta = 0.0, \delta = 0.25$	1.00	0.82	0.82	0.83	0.86	0.85	0.85	0.84	0.85	0.02
2. $\Delta = 0.0, \delta = 0.40$	1.01	0.83	0.81	0.87	0.88	0.86	0.86	0.87	0.91	0.03
3. $\Delta = 0.15, \delta = 0.25$	1.00	0.84	0.83	0.84	0.83	0.87	0.84	0.85	0.83	0.02
4. $\Delta = 0.15, \delta = 0.40$	1.01	0.81	0.82	0.83	0.84	0.87	0.85	0.84	0.85	0.02

Table 5

Bias for Linking Models

5a. Bias for ability trend x magnitude of drift conditions

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
1. $\Delta = 0.0, \delta = 0.25$	-0.11	-0.14	-0.19	-0.15	-0.04	-0.08	-0.08	-0.13	-0.11	0.05
2. $\Delta = 0.0, \delta = 0.40$	-0.15	-0.21	-0.28	-0.22	-0.04	-0.08	-0.10	-0.17	-0.16	0.08
3. $\Delta = 0.15, \delta = 0.25$	-0.28	-0.37	-0.23	-0.20	-0.25	-0.22	-0.14	-0.17	-0.23	0.07
4. $\Delta = 0.15, \delta = 0.40$	-0.40	-0.52	-0.35	-0.29	-0.29	-0.26	-0.13	-0.22	-0.31	0.12
	-0.24	-0.31	-0.26	-0.22	-0.15	-0.16	-0.11	-0.17	-0.20	0.11

5b. Bias for levels of magnitude of drift

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
$\delta = 0.25$	-0.20	-0.25	-0.21	-0.18	-0.14	-0.15	-0.11	-0.15	-0.17	0.05
$\delta = 0.40$	-0.27	-0.36	-0.31	-0.26	-0.17	-0.17	-0.11	-0.19	-0.23	0.09
Average	-0.24	-0.31	-0.26	-0.22	-0.15	-0.16	-0.11	-0.17	-0.20	0.07

5c. Bias for levels of ability trend

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
$\Delta = 0.0$	-0.13	-0.18	-0.23	-0.19	-0.04	-0.08	-0.09	-0.15	-0.14	0.06
$\Delta = 0 .15$	-0.34	-0.44	-0.29	-0.24	-0.27	-0.24	-0.13	-0.19	-0.27	0.09
Average	-0.24	-0.31	-0.26	-0.22	-0.15	-0.16	-0.11	-0.17	-0.20	0.10

Table 6

RMSE for Linking Models

6a. RMSE for ability trend x magnitude of drift conditions

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
1. $\Delta = 0.0, \delta = 0.25$	0.40	0.41	0.59	0.40	0.38	0.39	0.39	0.40	0.42	0.07
2. $\Delta = 0.0, \delta = 0.40$	0.41	0.44	0.46	0.43	0.37	0.38	0.38	0.40	0.41	0.03
3. $\Delta = 0.15, \delta = 0.25$	0.46	0.52	0.43	0.42	0.43	0.42	0.39	0.41	0.43	0.04
4. $\Delta = 0.15, \delta = 0.40$	0.55	0.64	0.51	0.46	0.46	0.45	0.38	0.42	0.48	0.08
	0.45	0.50	0.50	0.43	0.41	0.41	0.39	0.41	0.44	0.06

6b. RMSE for levels of magnitude of drift

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
$\delta = 0.25$	0.43	0.46	0.51	0.41	0.41	0.40	0.39	0.40	0.43	0.04
$\delta = 0.40$	0.48	0.54	0.48	0.45	0.42	0.41	0.38	0.41	0.45	0.05
Average	0.45	0.50	0.50	0.43	0.41	0.41	0.39	0.41	0.44	0.05

6c. RMSE for levels of ability trend

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
$\Delta = 0.0$	0.40	0.42	0.53	0.42	0.37	0.38	0.39	0.40	0.41	0.05
$\Delta = 0 .15$	0.50	0.58	0.47	0.44	0.45	0.44	0.39	0.42	0.46	0.06
Average	0.45	0.50	0.50	0.43	0.41	0.41	0.39	0.41	0.44	0.05

Table 7

Effects on $\hat{\theta}$ of Linking Technique, IPD Testing, and Linking Method on Bias

Table 7a. Bias: Main Effects of Linking Technique, IPD Testing, and Linking Method

	Linking Technique		IPD Testing		Linking Method	
	Fixed	TCC	No	Yes	Indirect	Direct
1. $\Delta = 0.0, \delta = 0.25$	-0.09	-0.14	-0.15	-0.08	-0.12	-0.10
2. $\Delta = 0.0, \delta = 0.40$	-0.12	-0.19	-0.22	-0.10	-0.17	-0.14
3. $\Delta = 0.15, \delta = 0.25$	-0.28	-0.19	-0.27	-0.20	-0.24	-0.23
4. $\Delta = 0.15, \delta = 0.40$	-0.37	-0.25	-0.39	-0.22	-0.32	-0.29
	-0.22	-0.19	-0.26	-0.15	-0.21	-0.19

Table 7b. Differences in Bias (TCC − Fixed): Interaction Effect for Linking Technique

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	-0.05	0.10	0.02
$\delta = 0.40$	-0.07	0.12	0.02
Average	-0.06	0.11	0.02

Table 7c. Differences in Bias (Yes − No): Interaction Effect for IPD Testing

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	0.07	0.08	0.07
$\delta = 0.40$	0.12	0.16	0.14
Average	0.09	0.12	0.11

Table 7d. Differences in Bias (Direct − Indirect): Interaction Effect for Linking Method

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	0.02	0.01	0.02
$\delta = 0.40$	0.03	0.03	0.03
Average	0.03	0.02	0.02

Table 8

Effects on $\hat{\theta}$ of Linking Technique, IPD Testing, and Linking Method on RMSE

Table 8a. RMSE: Main Effects of Linking Technique, IPD Testing, and Linking Method

	Linking Technique		IPD Testing		Linking Method	
	Fixed	TCC	No	Yes	Indirect	Direct
1. $\Delta = 0.0, \delta = 0.25$	0.39	0.45	0.45	0.39	0.40	0.44
2. $\Delta = 0.0, \delta = 0.40$	0.40	0.42	0.43	0.38	0.41	0.41
3. $\Delta = 0.15, \delta = 0.25$	0.46	0.41	0.46	0.41	0.44	0.43
4. $\Delta = 0.15, \delta = 0.40$	0.52	0.45	0.54	0.43	0.49	0.48
	0.44	0.43	0.47	0.40	0.44	0.44

Table 8b. Differences in RMSE (TCC − Fixed): Interaction Effect for Linking Technique

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	0.05	-0.05	0.00
$\delta = 0.40$	0.02	-0.08	-0.03
Average	0.04	-0.06	-0.01

Table 8c. Differences in RMSE (Yes − No): Interaction Effect for IPD Testing

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	-0.06	-0.04	-0.05
$\delta = 0.40$	-0.05	-0.11	-0.08
Average	-0.06	-0.08	-0.07

Table 8d. Differences in RMSE (Direct − Indirect): Interaction Effect for Linking Method

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	0.04	-0.01	0.01
$\delta = 0.40$	-0.01	-0.02	-0.01
Average	0.02	-0.02	0.00

Table 9

Y5 Mean Item Difficulty for Linking Models

	Linking Models									
	True	1	2	3	4	5	6	7	8	Mean
1. $\Delta = 0.0, \delta = 0.25$	0.21	0.16	0.12	0.11	0.14	0.24	0.18	0.20	0.16	0.04
2. $\Delta = 0.0, \delta = 0.40$	0.31	0.14	0.06	0.12	0.18	0.32	0.25	0.30	0.24	0.09
3. $\Delta = 0.15, \delta = 0.25$	0.21	0.09	0.01	0.12	0.16	0.11	0.15	0.21	0.19	0.06
4. $\Delta = 0.15, \delta = 0.40$	0.31	0.07	-0.05	0.11	0.18	0.21	0.19	0.34	0.24	0.12

Table 10

Y5 Pooled Standard Deviations of Item Difficulty for Linking Models

	Linking Models									
	True	1	2	3	4	5	6	7	8	Mean
1. $\Delta = 0.0, \delta = 0.25$	1.09	0.98	1.02	0.97	1.00	0.95	1.06	0.98	1.00	0.03
2. $\Delta = 0.0, \delta = 0.40$	1.16	0.92	0.94	1.03	1.04	0.98	1.01	1.03	1.07	0.05
3. $\Delta = 0.15, \delta = 0.25$	1.09	1.03	0.99	1.02	1.01	1.08	1.02	1.04	1.01	0.03
4. $\Delta = 0.15, \delta = 0.40$	1.16	0.95	1.00	1.04	1.05	1.05	1.03	1.05	1.07	0.04

Table 11

Bias for Linking Models

11a. Bias for ability trend x magnitude of drift conditions

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
1. $\Delta = 0.0, \delta = 0.25$	-0.05	-0.09	-0.10	-0.07	0.03	-0.03	0.00	-0.05	-0.05	0.04
2. $\Delta = 0.0, \delta = 0.40$	-0.17	-0.25	-0.19	-0.13	0.01	-0.06	-0.01	-0.06	-0.11	0.09
3. $\Delta = 0.15, \delta = 0.25$	-0.12	-0.19	-0.08	-0.05	-0.10	-0.06	0.00	-0.02	-0.08	0.06
4. $\Delta = 0.15, \delta = 0.40$	-0.24	-0.35	-0.19	-0.13	-0.10	-0.12	0.03	-0.06	-0.15	0.12
	-0.15	-0.22	-0.14	-0.09	-0.04	-0.07	0.00	-0.05	-0.09	0.09

11b. Bias for levels of magnitude of drift

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
$\delta = 0.25$	-0.09	-0.14	-0.09	-0.06	-0.03	-0.05	0.00	-0.03	-0.06	0.04
$\delta = 0.40$	-0.20	-0.30	-0.19	-0.13	-0.04	-0.09	0.01	-0.06	-0.13	0.09
Average	-0.15	-0.22	-0.14	-0.09	-0.04	-0.07	0.00	-0.05	-0.09	0.08

11c. Bias for levels of ability trend

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
$\Delta = 0.0$	-0.11	-0.17	-0.14	-0.10	0.02	-0.04	-0.01	-0.06	-0.08	0.06
$\Delta = 0 .15$	-0.18	-0.27	-0.14	-0.09	-0.10	-0.09	0.02	-0.04	-0.11	0.08
Average	-0.15	-0.22	-0.14	-0.09	-0.04	-0.07	0.00	-0.05	-0.09	0.08

Table 12

RMSE for Linking Models

12a. RMSE for ability trend x magnitude of drift conditions

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
1. $\Delta = 0.0, \delta = 0.25$	0.49	0.49	0.33	0.31	0.36	0.43	0.31	0.31	0.38	0.08
2. $\Delta = 0.0, \delta = 0.40$	0.57	0.58	0.38	0.35	0.38	0.46	0.33	0.33	0.42	0.10
3. $\Delta = 0.15, \delta = 0.25$	0.47	0.46	0.35	0.34	0.41	0.39	0.34	0.34	0.39	0.05
4. $\Delta = 0.15, \delta = 0.40$	0.59	0.64	0.37	0.33	0.36	0.42	0.31	0.32	0.42	0.13
	0.53	0.54	0.36	0.33	0.37	0.42	0.32	0.33	0.40	0.09

12b. RMSE for levels of magnitude of drift

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
$\delta = 0.25$	0.48	0.48	0.34	0.33	0.38	0.41	0.32	0.33	0.38	0.07
$\delta = 0.40$	0.58	0.61	0.38	0.34	0.37	0.44	0.32	0.32	0.42	0.12
Average	0.53	0.54	0.36	0.33	0.37	0.42	0.32	0.33	0.40	0.09

12c. RMSE for levels of ability trend

	Linking Models									
	1	2	3	4	5	6	7	8	Mean	SD
$\Delta = 0.0$	0.53	0.53	0.36	0.33	0.37	0.44	0.32	0.32	0.40	0.09
$\Delta = 0 .15$	0.53	0.55	0.36	0.34	0.38	0.41	0.32	0.33	0.40	0.09
Average	0.53	0.54	0.36	0.33	0.37	0.42	0.32	0.33	0.40	0.09

Table 13

Effects on Item Difficulty of Linking Technique, IPD Testing, and Linking Method on Bias

Table 13a. Bias: Main Effects of Linking Technique, IPD Testing, and Linking Method

	Linking Technique		IPD Testing		Linking Method	
	Fixed	TCC	No	Yes	Indirect	Direct
1. $\Delta = 0.0, \delta = 0.25$	-0.04	-0.06	-0.08	-0.01	-0.06	-0.03
2. $\Delta = 0.0, \delta = 0.40$	-0.12	-0.10	-0.18	-0.03	-0.13	-0.09
3. $\Delta = 0.15, \delta = 0.25$	-0.12	-0.04	-0.11	-0.04	-0.08	-0.08
4. $\Delta = 0.15, \delta = 0.40$	0.20	-0.09	-0.23	-0.06	-0.17	-0.13
	-0.12	-0.07	-0.15	-0.04	-0.11	-0.08

Table 13b. Differences in Bias (TCC − Fixed): Interaction Effect for Linking Technique

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	-0.02	0.08	0.03
$\delta = 0.40$	0.02	0.11	0.07
Average	0.00	0.10	0.05

Table 13c. Differences in Bias (Yes − No): Interaction Effect for IPD Testing

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	0.07	0.07	0.07
$\delta = 0.40$	0.15	0.17	0.16
Average	0.11	0.12	0.11

Table 13d. Differences in Bias (Direct − Indirect): Interaction Effect for Linking Method

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	0.03	0.00	0.02
$\delta = 0.40$	0.04	0.04	0.04
Average	0.03	0.02	0.03

Table 14

Effects on Item Difficulty of Linking Technique, IPD Testing, and Linking Method on RMSE

Table 14a. RMSE: Main Effects of Linking Technique, IPD Testing, and Linking Method

	Linking Technique		IPD Testing		Linking Method	
	Fixed	TCC	No	Yes	Indirect	Direct
1. $\Delta = 0.0, \delta = 0.25$	0.44	0.32	0.40	0.35	0.38	0.37
2. $\Delta = 0.0, \delta = 0.40$	0.50	0.35	0.47	0.38	0.43	0.42
3. $\Delta = 0.15, \delta = 0.25$	0.43	0.34	0.40	0.37	0.38	0.36
4. $\Delta = 0.15, \delta = 0.40$	0.50	0.33	0.48	0.35	0.43	0.41
	0.47	0.34	0.44	0.36	0.41	0.40

Table 14b. Differences in RMSE (TCC − Fixed): Interaction Effect for Linking Technique

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	-0.12	-0.09	-0.11
$\delta = 0.40$	-0.15	-0.17	-0.16
Average	-0.13	-0.13	-0.13

Table 14c. Differences in RMSE (Yes − No): Interaction Effect for IPD Testing

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	-0.05	-0.03	-0.04
$\delta = 0.40$	-0.10	-0.13	-0.11
Average	-0.07	-0.08	-0.08

Table 14d. Differences in RMSE (Direct − Indirect): Interaction Effect for Linking Method

	$\Delta = 0.0$	$\Delta = 0.15$	Average
$\delta = 0.25$	-0.01	0.00	0.00
$\delta = 0.40$	-0.01	-0.02	-0.02
Average	-0.01	-0.01	-0.01

Table 15

Year 4 Parameter Means and Standard Deviations for Linking Models

	Linking Models							
	1	2	3	4	5	6	7	8
Mean $\hat{\theta}$	0.53	0.53	0.67	0.57	0.53	0.51	0.62	0.65
St. Dev. $\hat{\theta}$	0.57	0.57	0.47	0.52	0.61	0.62	0.58	0.58
Mean *a*	0.82	0.82	0.88	1.01	0.75	0.74	0.87	0.85
St. Dev. *a*	0.38	0.38	0.37	0.43	0.35	0.35	0.37	0.36
Mean *b*	-0.49	-0.49	-0.59	-0.50	-0.82	-0.96	-0.61	-0.60
St. Dev. *b*	2.01	2.01	1.03	0.89	1.23	1.17	1.04	1.06
Mean *c*	0.24	0.24	0.27	0.27	0.25	0.24	0.27	0.27
St. Dev. *c*	0.04	0.04	0.06	0.06	0.04	0.05	0.06	0.06

Table 16

Average Differences and RMSDs Between Linking Models

16a. Ability Statistics

	Linking Models							
	1	2	3	4	5	6	7	8
Model 1		0.00	-0.14	-0.05	-0.01	0.01	-0.09	-0.12
Model 2	0.00		-0.14	-0.05	-0.01	0.01	-0.09	-0.12
Model 3	0.22	0.22		0.09	0.13	0.16	0.05	0.02
Model 4	0.15	0.15	0.12		0.04	0.06	-0.04	-0.07
Model 5	0.12	0.12	0.21	0.14		0.02	-0.09	-0.11
Model 6	0.12	0.12	0.23	0.14	0.09		-0.11	-0.13
Model 7	0.17	0.17	0.12	0.07	0.13	0.14		-0.03
Model 8	0.18	0.18	0.12	0.10	0.15	0.16	0.03	

16b. Item Discrimination Statistics

	Linking Models							
	1	2	3	4	5	6	7	8
Model 1		0.00	-0.06	-0.19	0.07	0.08	-0.05	-0.03
Model 2	0.00		-0.06	-0.19	0.07	0.08	-0.05	-0.03
Model 3	0.35	0.35		-0.13	0.13	0.14	0.01	0.03
Model 4	0.42	0.42	0.15		0.26	0.28	0.15	0.17
Model 5	0.24	0.24	0.28	0.38		0.01	-0.12	-0.10
Model 6	0.26	0.26	0.26	0.37	0.22		-0.13	-0.11
Model 7	0.34	0.34	0.01	0.16	0.28	0.25		0.02
Model 8	0.34	0.34	0.03	0.18	0.26	0.24	0.02	

16c. Item Difficulty Statistics

	Linking Models							
	1	2	3	4	5	6	7	8
Model 1		0.00	0.10	0.00	0.32	0.47	0.11	0.11
Model 2	0.00		0.10	0.00	0.32	0.47	0.11	0.11
Model 3	1.53	1.53		-0.09	0.23	0.37	0.02	0.01
Model 4	1.55	1.55	0.16		0.32	0.46	0.11	0.10
Model 5	1.36	1.36	0.56	0.64		0.15	-0.21	-0.21
Model 6	1.55	1.55	0.61	0.69	0.64		-0.36	-0.36
Model 7	1.52	1.52	0.02	0.19	0.55	0.60		-0.01
Model 8	1.52	1.52	0.04	0.20	0.55	0.61	0.02	

Note: Upper diagonal contains average pairwise differences. Lower diagonal contains root mean squared differences between model pairs.

Table 17

Correlations Between Linking Models for Item Discrimination and Difficulty

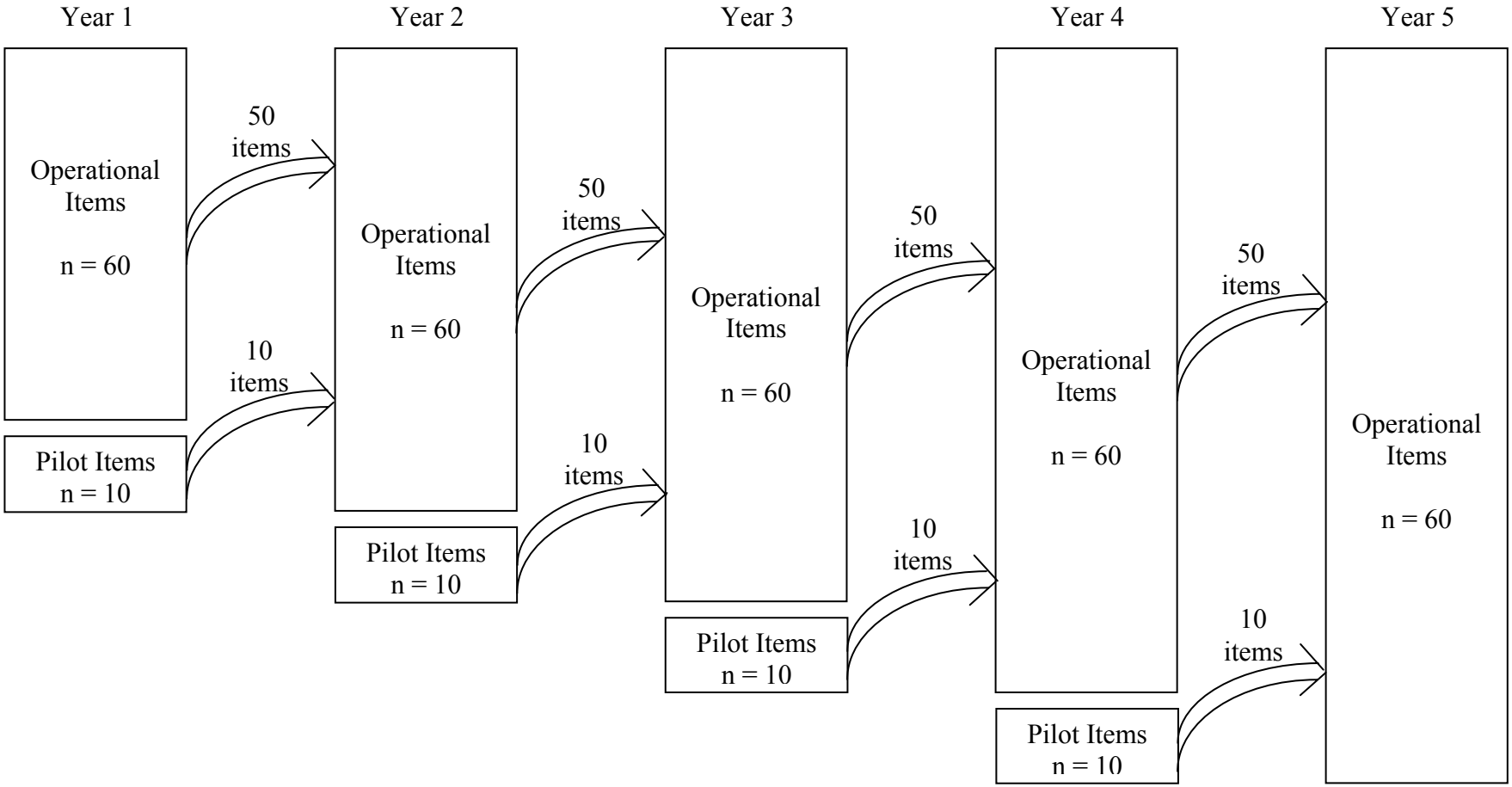
	Linking Models							
	1	2	3	4	5	6	7	8
Model 1		1.00	0.57	0.57	0.80	0.76	0.57	0.57
Model 2	1.00		0.57	0.57	0.80	0.76	0.57	0.57
Model 3	0.66	0.66		1.00	0.75	0.81	1.00	1.00
Model 4	0.66	0.66	1.00		0.75	0.81	1.00	1.00
Model 5	0.76	0.76	0.91	0.91		0.80	0.75	0.75
Model 6	0.67	0.67	0.91	0.91	0.86		0.81	0.81
Model 7	0.66	0.66	1.00	1.00	0.91	0.91		1.00
Model 8	0.66	0.66	1.00	1.00	0.91	0.91	1.00	

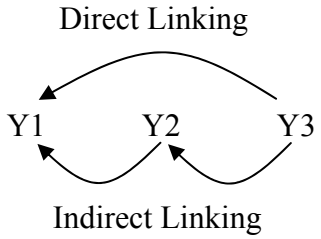
Note: Upper diagonal contains correlations between item discrimination values. Lower diagonal contains correlations between item difficulty values.

Table 18

Linking Models' Pass Rates for Different Cut Scores

	Linking Models							
Cut Score	1	2	3	4	5	6	7	8
$\theta = -0.25$	0.92	0.92	0.97	0.95	0.91	0.90	0.94	0.94
$\theta = 0.00$	0.84	0.84	0.92	0.88	0.81	0.80	0.87	0.87
$\theta = 0.25$	0.69	0.69	0.82	0.74	0.68	0.67	0.74	0.76
$\theta = 0.50$	0.51	0.51	0.66	0.56	0.53	0.51	0.59	0.61
$\theta = 0.75$	0.34	0.34	0.46	0.37	0.36	0.35	0.42	0.43
$\theta = 1.00$	0.20	0.20	0.25	0.20	0.22	0.21	0.24	0.26


Year 1
Year 2
Year 3
Year 4
Year 5
Operational Items
n = 60
Pilot Items
n = 10
50 items
10 items
Operational Items
n = 60
Pilot Items
n = 10
50 items
10 items
Operational Items
n = 60
Pilot Items
n = 10
50 items
10 items
Operational Items
n = 60
Pilot Items
n = 10
50 items
10 items
Operational Items
n = 60


Direct Linking
Y1
Y2
Y3
Indirect Linking