Application of Multilevel IRT to Multiple-Form Linking When Common Items Are Drifted

Chanho Park[1]
Taehoon Kang[2]
James A. Wollack[1]

[1] University of Wisconsin-Madison
[2] ACT, Inc.

April 11, 2007

Paper presented at the 2007 annual meeting of the National Council on Measurement in Education, April 10 – April 12, Chicago, IL.

Application of Multilevel IRT to Multiple-Form Linking When Common Items Are Drifted

When scores obtained from different forms of a test are compared, it is necessary to link the forms onto a common scale. When the forms are linked by common items and the common items have differential item functioning or item parameter drift (IPD), however, the linking process can produce inaccurate results. Common practice, thus, has been to detect and eliminate IPD items before linking the forms. This study, by building in the testing occasion as a third level in the model, develops a three-level item response theory methodology for linking in the presence of IPD across multiple testing occasions. This model may reduce linking error by unifying linking and IPD detection. The developed model has two different statistical identification methods depending on the assumptions on the model. Using the two model identification methods, two different simulation studies were conducted. In both studies, the model was fit using a Markov chain Monte Carlo estimation procedure. The results show that linking results were encouraging in both simulation studies in spite of the relatively poor performance in detecting IPD. Cautions when interpreting results, implications of simulation conditions, and limitations or suggestions for future studies are also discussed.

Application of Multilevel IRT to Multiple-Form Linking When Common Items Are Drifted

In educational testing, in order to compare scores obtained from different forms of a test it is necessary to link the forms onto a common scale. When linking is based on item response theory (IRT) models, a certain number of common items are needed to ensure the equivalence of linked scales due to the invariance principle of IRT. However, the linking process by common items can produce inaccurate results when some of the anchor items have problems such as differential item functioning (DIF; Holland & Thayer, 1993) or item parameter drift (IPD; Bock, Muraki, & Pfeiffenberger, 1988). When multiple forms from multiple waves are linked, IPD is a bigger concern to test practitioners, and it is a common practice to eliminate the IPD items before linking. Reducing the number of common items, however, may increase linking error.

The problem of linking when DIF or IPD exists in common items can be addressed by building in a higher level to the traditional IRT models. Chu & Kamata (2005) formed a three-level hierarchical generalized linear model (HGLM) and demonstrated how the DIF analysis could be integrated into the linking process in an HGLM framework. By adding a DIF grouping variable in the third level, they found that the information contained in DIF items could improve the precision of linking. In addition, the magnitude of DIF effects was estimable.

Although Chu & Kamata (2005) made an important contribution to improve linking, it was not without some limitations. First, their model and notation were highly specific to the HLM software package. It will appeal more to IRT users if the model is

expressed in IRT notation. Second, the model was presented only in terms of the Rasch model, thus restricting the number of testing programs to which it will apply. Finally, they studied only the DIF cases between two groups in equating, and it was not clearly stated how the method could be used for IPD over more than two occasions.

This study aims to extend the work of Chu & Kamata (2005) by developing a multilevel IRT methodology for linking in the presence of IPD across multiple testing occasions. As a general model, a three-level three-parameter logistic (3PL) model will be used. Since the Rasch, one-parameter logistic (1PL), and two-parameter logistic (2PL) models are special cases of the 3PL model, the method can easily incorporate the other IRT models. By doing so, a framework that unifies IPD and linking in dichotomous IRT models can be obtained.

The primary purpose of this study is to present the general framework and illustrate the model in simulation studies. Two simulation studies will be conducted by using different model identification methods in order to demonstrate that this new methodology functions as expected. Since myriads of IPD detection techniques and linking methods are available, it is not intended in this study to compare the results of this method with those of other methods.

Development of a 3-level IRT model

This study aims to develop a general method to solve problems in linking due to IPD. When different forms are administered to different groups of examinees, the items are common across the forms, and the 3PL model is fit to the data, the three-level IRT model can be formulated as follows:

At level 1, the probability of correctly answering an item is given by the 3PL. More precisely,

$$P\left(U_{ijk}=1 \mid a_{ik}, b_{ik}, c_{ik}; \theta_{jk}\right) = p^{*} + c_{ik}\left(1-p^{*}\right), \text{ where } p^{*} = \frac{\exp\left[1.7a_{ik}\left(\theta_{jk}-b_{ik}\right)\right]}{1+\exp\left[1.7a_{ik}\left(\theta_{jk}-b_{ik}\right)\right]};$$

$U_{ijk}=1$ is the correct response to the i^{th} item by the j^{th} person during the k^{th} occasion (1,…, *i*,…, I; 1,…, *j*,…, J; 1,…, *k*,…, K); θ is the person ability, and *a*, *b*, and *c* are the item discrimination, difficulty, and pseudo-guessing parameters, respectively. Note that the *k* subscript is added to an ordinary 3PL model to accommodate the third level (occasion).

At level 2,

$$\theta_{jk} = \mu_{k} + \varepsilon_{jk};$$

where $\varepsilon_{jk} \sim N\left(0,\ \sigma_{k}^{2}\right)$, and μ_{k} is the ability mean for occasion *k*. The parameters of item *i* are assumed to be the same for all examinees in occasion *k*, and thus remain the same as in level 1.

Finally at level 3,

$$\mu_{k} = \nu_{0} + \tau_{k};$$

$$a_{ik} = \alpha_{i},\ \ c_{ik} = \gamma_{i},\ b_{ik} = \begin{cases} \beta_{i}, & k=1 \\ b_{i,(k-1)} + \xi_{i,(k-1)}x_{i,(k-1)}, & k=2,\ldots,K \end{cases}$$

where ν_{0} is the overall mean and $\tau_{k} \sim N\left(0,\ \kappa^{2}\right)$. $X\left(=\left\{x_{i,(k-1)}\right\}\right)$ is a dummy variable matrix for the IPD with I rows and (K-1) columns, where $x_{i,(k-1)}=1$ denotes that the i^{th}

item is an IPD candidate in the k^{th} occasion. More about X will be discussed in the next section.

Since IPD is defined as parameter shifts from the previous occasion, IPD effect parameters ($\xi_{i,(k-1)}$) are necessary only from the second occasion while item parameters for the first occasion have only fixed effects. From the second occasion, $\xi_{i,(k-1)}$ denotes (possibly random) IPD effects in the k^{th} occasion. Although only IPD in the *b*-parameter (*b*-IPD) will be considered in this study, IPD for all three parameters is possible, and its implementation is straightforward. Here, α_i, β_i, and γ_i are item *i*'s discrimination, difficulty, and pseudo-guessing parameters, respectively, for the base group. Item *i* has *b*-IPD for the k^{th} occasion if $\xi_{i,(k-1)}$ is significantly different from 0. Also, the magnitude of the IPD parameter, $\xi_{i,(k-1)}$, can be taken as an estimate of the IPD effect size. IPD for discrimination and pseudo-guessing parameters could be estimated for occasions two through *K* by adding $\zeta_{i,(k-1)}x_{i,(k-1)}$ and $\eta_{i,(k-1)}x_{i,(k-1)}$ to the $a_{i,(k-1)}$ and $c_{i,(k-1)}$ values, respectively.

Model Identification

Because latent variables do not have inherent scales, IRT models are over-parameterized models and the estimated model parameters are thus exact only up to a linear transformation (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000). Because item parameters are considered structural parameters and ability parameters incidental parameters (Hambleton & Swaminathan, 1985), it is common practice in IRT to fix the metric of ability parameter (θ) to resolve the indeterminacy. It is usually assumed that θ follows a normal distribution, and that the mean of θ is 0. Fixing the

mean of θ to a constant is sufficient to identify the 1PL model or the Rasch model, although 2PL or 3PL models need an additional constraint such as fixing the standard deviation of ability.

In multilevel IRT models, the existence of multiple levels makes identifying the model more complicated. The three-level model formulated above contains K underidentified models when the groups are assumed to be independent. Thus, the parameters are indeterminate within groups and between groups. Hence, double indeterminacy exists in the suggested model. To resolve this issue, first, σ_k^2 are arbitrarily fixed to equal 1.0, following a common practice in IRT. ν_0 can also be set to 0 without any loss of generality. Further, assuming all groups are independent, the model can be fully identified if κ^2 is 0, i.e., all group means are equal to ν_0. This identification condition is the same as assuming a standard multivariate normal distribution on the ability parameters. That is,

$$\Theta_K \sim \text{MVN}(\mathbf{0}_\text{K}, \mathbf{I}_{\text{K}\times\text{K}}),$$

where Θ_K is the person ability parameters in K occasions,

$\mathbf{0}_\text{K}$ is a zero-vector with K elements,

and $\mathbf{I}_{\text{K}\times\text{K}}$ is a $K \times K$ identity matrix.

This condition may be reasonable when all the occasions are considered to be sampled from the same population. When this condition holds, all elements of X become 1.0, and

all items are IPD candidates. The items having IPD parameters significantly different from 0 will be adjudged to have IPD.

When studying IPD, however, fixing the group means may not be a reasonable assumption when the testing occasions constitute relatively long periods of time (e.g., years). When a test form is administered each year for multiple years, the construct being measured by the test may change over time, as may characteristics of the examinee population or item pool. Changes in both person and item characteristics can be confounded, which complicates detecting IPD items and linking the scales.

If ability means (μ) shift across groups and thus cannot be fixed, the model can be identified by placing constraints on item parameters instead of on ability parameters. The items used to link the scales should be common items containing no IPD. These non-IPD items can be found by using IPD or DIF detection techniques such as the Mantel-Haenszel test (Holland & Thayer, 1988), the SIBTEST (Shealy & Stout, 1993), Lord's χ^2 test (Lord, 1980), or the likelihood ratio test (LR-test; Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993). It may seem counterintuitive to use IPD detection methods before applying the suggested model because the advantage of the suggested model is that IPD detection and linking are unified under one general framework. However, the purpose of applying IPD detection methods is to find non-IPD items (rather than IPD items) to serve as anchor items. Under this framework, items not included in the anchor will become IPD candidates in the three-level IRT model and will be tested for possible IPD.

This two-stage process—finding IPD candidates first and then finding IPD items—is analogous to the two steps of applying the LR-test with iterative linking. When using the LR-test, first, one item at a time becomes an IPD candidate, and all the other

items function as anchor items. The results from this first run are sub-optimal because the anchor items are not "pure," that is, not IPD-free. From the results of the first test, anchor items and IPD candidates are separated, and the LR-test is conducted again to finally detect IPD items from the candidates. Therefore, the suggested method can be thought of as replacing the second run of the LR-test. One of the strengths of the suggested method is that IPD candidate items need not be eliminated from the linking. Instead, they remain in the model and information contained in those items will contribute to the linking results.

MCMC Estimation

A three-level IRT model was fit to simulated data using a Markov chain Monte Carlo (MCMC) procedure implemented in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). This approach involved initial specification of the model and a prior for all model parameters. Using a Metropolis Hastings algorithm, WinBUGS was used to simulate parameter observations from the joint posterior distribution of the model parameters. The success of the algorithm was evaluated by whether the chain converged to a stationary distribution, in which case characteristics of that posterior distribution (e.g., the sample mean for each parameter) could be taken as point estimates of the model parameters. In the current application, the following priors were chosen for the model parameters:

$$\mu_k \sim N(0,\ 1),$$

$$\zeta_i \sim LogNormal(0,\ 1),$$

$$\delta_i \sim N(0,\ 1),$$

$$\eta_i \sim Beta(5,\ 17),$$

$$\xi_{i(k-1)} \sim N(0,\ .5^2).$$

In MCMC estimation, several additional issues require consideration in monitoring the sampling history of the chain. WinBUGS, by default, will use an initial 4,000 iterations to "learn" how to generate values from proposal distributions to optimize sampling under Metropolis Hastings. Therefore, the initial 5,000 iterations were discarded as "burn-in," and an additional 10,000 iterations were simulated and inspected for convergence using visual inspection, as well as convergence statistics available in CODA (Best, Cowles, & Vines, 1996).

Study 1

The purpose of Study 1 is to demonstrate how the model works in a straightforward, two-group, common-item setting, where linking is conducted in the presence of IPD. Also note that this is the case directly applicable to the study of DIF.

Administration of a 60-item test to 1,000 examinees on two testing occasions was simulated using the 3PL. Out of the 60 items, a randomly selected 15 items were manipulated to contain IPD. Base scale item parameters for these 60 items were generated according to the following distributions: $\alpha \sim$ Lognormal $(0, .5^2)$, $\beta \sim N(0, 1)$, $\gamma \sim$ Logit normal$(-1.4, .3^2)$. IPD was simulated by adding a constant δ to the item difficulty. Two δ values of .25 and .4 were used for simulating moderate and large magnitudes of IPD, respectively. Similar values of δ have been used in other studies of IPD (Donoghue

& Isham, 1998; Wells, Subkoviak, & Serlin, 2002; Wollack, Sung, & Kang, 2006). In all cases of IPD, δ was added to the base β_i value so that the item difficulty during the second testing occasion was always larger (i.e., the item was harder) than it was in the first occasion.

To simulate the data, it was assumed that the two groups were sampled from the same normally distributed population. Therefore, the model was identified by fixing the θ metric for the two groups; that is, it was assumed $\Theta_2 \sim \text{MVN}(\mathbf{0}_2, \mathbf{I}_{2\times 2})$ as seen in the Model Identification section. Due to the computational demands of MCMC estimation, each of the two conditions (δ = .25 or .4) was replicated only five times.

Results of Study 1

Visual inspection of the chains as well as convergence statistics available in CODA (Best et al., 1996) confirmed convergence of the chains. Since the suggested 3-level IRT model combines IPD testing and linking, it was examined how the model performed in both IPD testing and linking.

The criterion to detect IPD was such that an item was concluded to have IPD if the ξ parameter for the item did not contain 0 in the central 95% of the posterior distribution. That is, an item was concluded to be free from IPD if the product of the 2.5th percentile and the 97.5th percentile of the simulated values of the posterior distribution of ξ was negative. If an IPD-flagged item (i.e., $\xi \neq 0$) was truly simulated to have IPD, it was considered a true positive (TP). On the other hand, if an IPD-flagged item was not simulated to have IPD, it was classified as a false positive (FP). Note that FP is

compatible to empirical Type I error, and TP may be equivalent to the power when Type I error is properly controlled.

TP rate is the proportion of statistically significant ξ values among the 15 true IPD items, while FP rate is the proportion of statistically significant ξ values among the 45 non-IPD items. FP and TP rates for Study 1 are shown in Table 1. When the IPD effect was moderate ($\delta = .25$), the TP rate of .60 and FP rate of .04 demonstrate that the suggested model was functioning relatively well considering that detection of IPD is rather difficult with only moderate effects for 3PL models. When the IPD effect was large ($\delta = .4$), detection rate of TP increased, but this increase in the TP rate was at the expense of an inflated FP rate. As will be seen later, the increase of effect size did not necessarily improve test results.

Insert Table 1 About Here

It was also of interest how well ξ parameters were recovered. Table 2 shows the mean bias (ξ) and RMSE (ξ) across the 5 replications. When $\delta = .25$, ξ estimates were unbiased. There was a slightly negative bias when $\delta = .4$. Mean RMSE values were the same for the two conditions, with a reasonable value of .12.

Insert Table 2 About Here

Since only *b*-IPD was considered in this study, it was examined how well the *b*-parameters were recovered in the second testing occasion. Table 3 shows the mean bias and RMSE of b_2 estimates, as well as the mean correlation coefficients between $\hat{b}_2$ and b_2. Negligible biases, small RMSE, and high correlation coefficients (.99 for both) demonstrate successful recovery of the b_2 parameters for both values of δ.

Insert Table 3 About Here

Conclusion of Study 1

The results of Study 1 were generally as expected and provide support for the proposed 3-level IRT model for linking and detecting IPD. The results of Study 1 were particularly encouraging under conditions of moderate IPD. Although the model was able to identify 84% of IPD items in the δ = .4 condition, the larger than expected FP rate is cause for some concern and warrants further study.

Study 2

Study 1 demonstrated successful application of the suggested model in a straightforward linking application. Here, application of the three-level IRT model is extended to a more complex situation in which IPD is allowed to compound over the course of a five-year period.

The simulation design for this study was modeled after that used by Wollack et al. (2006) for examining the longitudinal effects of IPD on a score scale. A 60-item test was simulated over five years, with an additional 10 pilot items for each of the first four years. Rather than using the same items over the entire five-year period, this study simulated a testing program in which the operational form of the test changed somewhat from year to year. Each year, the 60 operational items consisted of a random 50 of the 60 operational items and all 10 pilots that were administered during the preceding year. In this manner, 100 items were simulated in total, of which 35 items appeared in all five forms. Base scale item parameters for these 100 items were simulated from the same distributions as those used in Study 1: $\alpha \sim$ Lognormal $(0, .5^2)$, $\beta \sim N(0, 1)$, $\gamma \sim$ Logit normal$(-1.4, .3^2)$.

IPD was simulated as either compounding IPD (CIPD) or random IPD (RIPD). CIPD was simulated by adding a constant δ to the item difficulty from the preceding year each time the item drifted, and 0.0 to the preceding item difficulty each time the item did not drift. RIPD was simulated by adding δ to the item's base (Year 1) value. Again, the two δ values of .25 and .4 were used for simulating moderate and large amount of IPD, respectively.

CIPD was simulated on 10 randomly selected items from among the 35 that were common to all five forms. Of these 10 CIPD items, three items were simulated to drift four times (drifting by δ every year), four items drifted three times, and three items drifted two times. RIPD was simulated on 10 items each year, randomly selected from among the 50 non-CIPD items.

The number of examinees was fixed to 1,000 for each year. Examinee θ parameters were simulated in two different ways. In one condition, all θ for all years were generated from independent $N(0, 1)$ distributions. In the second condition, the mean of the ability

distribution was 0.0 for year 1, but increased by 0.15 for each of years 2 through 5. The variance was fixed at 1.0 for all years.

Since ability means (μ) may vary in one of the simulation conditions, a different identification method than in Study 1 was employed here. For the base occasion, θ scales were fixed to $N(0, 1)$. From the second to the fifth occasions, anchor items were linked to the previous year. There is one caveat to identifying the model in this way as was discussed in the Model Identification section: doing so requires a pre-analysis to find anchor items. Because there are many IPD detection methods available and these methods vary in terms of effectiveness, for purposes of this study, rather than selecting one technique and conducting such an analysis, we chose to simulate two situations where different IPD-screening devices with known detection properties were administered. In the first condition, we simulated a situation where the IPD test identified (correctly) 70% of the true IPD items and (incorrectly) 5% of the non-IPD items. In the second condition, 95% of the true IPD and 10% of the non-IPD items were detected. It is worthwhile to mention that the purpose of this pre-analysis is not to find IPD items, rather to find anchor items that do not contain IPD. Therefore, it might be reasonable to sacrifice some control of the FP rate in exchange for a very high TP rate, provided a sufficient number of non-IPD items remain for the anchor. Based on the conditions simulated, the former condition (70% / 5%) will identify a larger anchor that will contain a moderate number of IPD items, while the latter condition (95% / 10%) will identify a smaller anchor that will be almost entirely free of IPD items. Specifically, in the 70% / 5% condition, approximately 48 among the 60 common items were identified as anchor items each year. Of those 48 anchor items, approximately five were true IPD items that were falsely classified as anchor items. In the 95% / 10% condition, on the

other hand, approximately 39 of the 60 common items were selected as anchor items each year, with approximately one of those items being incorrectly included in the anchor.

The different identification rule led to changes in the *X* matrix in the specification of the Level 3 model. Here, the dummy $x_{i,(k-1)}$ is assigned 0 if item *i* is an anchor item in the k^{th} testing occasion; otherwise, it becomes 1. In other words, if pre-analysis found that the parameter of item *i* drifted for occasion *k*, the item becomes an IPD-candidate by setting $x_{i,(k-1)} = 1$. Since these IPD-candidates include 5% or 10% of FPs as well as 70% or 95% of TPs, the suggested method will further screen out FPs by conducting IPD detection tests and link the scales over the five occasions at the same time. As with Study 1, five replications were conducted for each of the eight simulation conditions.

Results of Study 2

Chain convergence was also confirmed by visual inspection and convergence statistics. To investigate the functionality of the suggested method, mean FP and TP rates were first examined. Then it was examined how well the μ parameters were recovered for each year using the criteria of bias and RMSE. Finally, the bias and RMSE of ξ estimates were examined.

Tables 4 and 5 show the mean FP and TP rates over 5 years. In all cases, FP rates were low. As will be discussed later, these FP rates are bounded by those of pre-analyses—5% and 10%. Considering the existence of upper bounds, however, FP rates dropped considerably from those of pre-analyses. Table 4 shows that FP rates are lower than .05 even when those of pre-analyses were .10 (95% / 10%).

Insert Table 4 About Here

TP rates, again, are bounded by those in pre-analyses—70% and 95%. Still, the obtained TP rates are considerably lower than their upper bounds, as seen in Table 5, signifying that true IPD items are rather hard to detect in this complicated design. TP rates were ever so slightly higher in the condition where ability means increased. Between the two anchoring conditions, the 95% / 10% condition had higher TP rates on average by about .085. Therefore, a smaller, but purer anchor resulted in higher TP rates than a larger, more contaminated anchor.

Comparing the detection rates of CIPD and RIPD items, TP rates of RIPD items were much higher than those of CIPD items. Note that the *b*-parameters of CIPD items were drifted two to four times while those of RIPD items were drifted only once. It was more difficult to detect CIPD each time the item was drifted than it was to detect IPD happening once at random. When a CIPD item parameter was drifted four times, it rarely happened that all four drifts were detected. As expected, increasing the effect size (from $\delta = .25$ to $\delta = .4$) corresponded to an increase of approximately 0.09 in the TP rates.

Insert Table 5 About Here

Tables 6 to 9 show the biases and RMSEs of the ability means (μ) averaged over the five replications. Tables 6 and 7, which report the mean bias for fixed and increasing μ, respectively, show that there were negative biases in Year 2 and positive biases in Year 5 for all conditions. The magnitudes of biases were, in general, fairly small, although some of the biases in Year 5 were substantial. Positive biases in later years (particularly in Year 5) may be due to the undetected IPD items. Since we simulated IPD by adding some positive constant (δ) to item difficulties, the undetected ξ corresponding to the IPD items in the anchor set will be absorbed by the θ estimates, thereby increasing estimates of μ. Indeed, biases were usually larger (more positive) in the large IPD (δ = .4) condition than in the moderate IPD (δ = .25) condition, and they cumulated over the years, culminating in Year 5. Somewhat larger absolute biases and RMSEs in Year 5 for the 95% / 10% condition than for the 70% / 5% condition may be due to the smaller anchor size, but further study is necessary before conclusions are made. No noticeable differences were observed between the increased μ condition and the fixed μ condition. The small differences are likely due to estimation errors.

Insert Tables 6 and 7 About Here

The average RMSEs of μ (in Tables 8 and 9 for fixed and increasing μ, respectively) were also very small, possibly except for in Year 5. Differences were not noticeable between the 70% / 5% and 95% / 10% conditions, and were quite small between the fixed μ and the increased μ conditions. Increasing δ made estimating μ

somewhat less precise in Year 5. However, we could not draw any general conclusions, possibly due to the small number of replications.

Insert Tables 8 and 9 About Here

Finally, it was also examined how the IPD effects (ξ) were recovered for all IPD candidates. Tables 10 and 11 show the biases and RMSEs of ξ for all IPD candidates averaged over five replications. Overall, negative biases are evident for all conditions. This may be due either to the nature of the priors ($\xi_{i(k-1)} \sim N\left(0,\ \ .5^2\right)$) or to nature of the IPD simulation, with all δ's being positive. Further exploration is needed.

Table 10 shows that there is no noticeable difference between the fixed μ and the increased μ conditions. The 95% / 10% condition had smaller absolute biases than the 70% / 5% condition, and δ = .4 had larger absolute biases than did δ = .25. These results were all consistent with earlier findings. The RMSEs of ξ (shown in Table 11) exhibited the same pattern as did the biases. Similar results were shown between the fixed μ and the increased μ conditions; the 95% / 10% condition had smaller RMSEs than the 70% / 5% condition did, and the δ = .4 condition had larger RMSEs than the δ = .25 condition.

Insert Tables 10-11 About Here

Conclusion of Study 2

In Study 2, the proposed three-level IRT model was evaluated using mean FP and TP rates, bias and RMSE of μ and ξ parameters. Since the identification condition of Study 2 requires a pre-analysis, FP and TP rates of the suggested method are upper-bounded by those of the pre-analysis. The suggested method lowered the FP rates from .05 to .01 and from .10 to .03 or .04. These decreases are, however, partly due to the decreases in TP rates: from .7 to .29-.62 and from .95 to .36-.76. It was also found out that RIPD items were easier to detect than CIPD items. Person ability means (μ) were, overall, recovered reasonably well, although biases and RMSEs of μ were rather high in later testing occasions. Estimation of IPD effects (ξ) for all IPD candidates had negative biases and relatively high RMSEs for all conditions.

The manipulated simulation conditions of Study 2 revealed noticeable findings as follows. First, large IPD effect size ($\delta = .4$) resulted in higher TP detection rates but made it harder to recover ability means and IPD effects. Making the anchor purer (95% / 10%) at the expense of fewer anchor items generally led to favorable results. Fixing or increasing μ over testing occasions resulted in only very minor changes.

Overall, the results of Study 2 were generally as expected and provide support for the proposed three-level IRT model when common anchor items detected in a pre-analysis were used to link the scales at the third level. The model produced acceptable results in the various conditions studied, although the TP rate is a bit low. However, investigations into IPD often are not about studying the characteristics of those items, but are about finding a set of suitable anchor items for the item linking process. In this capacity, in spite of the low TP detection rates, the method proposed here performed well at recovering underlying item difficulty and examinee ability parameters.

Discussion and Conclusion

Studies 1 and 2 demonstrated that the proposed 3-level IRT model functioned generally satisfactorily under various conditions from a simple and straightforward two-occasion situation to a rather complicated five-occasion situation. It was also demonstrated that the model could be identified differently by having different assumptions. If level-3 groups were assumed to be sampled from independent and identical normal distributions, the model could be identified by allowing independent standard normal distributions on person ability parameters. If the ability means were assumed to vary or if it was in doubt whether they were fixed or changed, a pre-analysis would be required to find anchor items to link the scales. It is important to note that the purpose of this pre-analysis was to find non-IPD anchor items rather than to screen out IPD items.

Although one of the strengths of the proposed method is that IPD detection and linking are unified under one general multilevel IRT framework, poor performance of IPD testing in the two simulation studies may seem discouraging. FP rates were sometimes higher than expected in the $\delta = .4$ condition in Study 1, and TP rates in both studies were generally low. However, unlike other IPD or DIF detection methods such as the LR-test, which is designed to control Type I error at a nominal level, it is unknown whether the IPD detection method employed in this study controls Type I error rates at a certain nominal rate or how powerful the method is in various conditions. Since an item was concluded to have IPD if simulated posterior distributions of its IPD parameter (ξ) did not include 0 in the central 95% regions, the results may well be affected by the choice of priors. We applied Bayesian analysis by implementing MCMC estimation

techniques, and significance testing in Bayesian analysis is rather under-explored territory.

In spite of the unsatisfactory results in IPD detection, general linking results were encouraging. Instead of discarding IPD-suspected items, the suggested method fully utilized the information of all the items without regard to their parameter drifts. Therefore, loss of information is minimized in the current method. Also, it may reduce the steps in conducting linking. Linking is often a laborious procedure including detection of IPD, deletion of the IPD items, calculation of linking coefficients, and scaling of item and person parameters. If some assumptions are satisfied, all these steps can be completed in one step using the proposed three-level IRT model, thereby reducing the risk of mistakes at each stage.

Some of the findings of this study need to be cautiously interpreted. First, RIPD items showed better detection rates than CIPD items. However, the parameters of CIPD items drifted two to four times, and with the shown detection rates, the CIPD items were detected at least once in almost all cases. Low detection rate of CIPD items suggests that it is difficult to detect each and every drift when it accumulates over time. It is quite likely that this result stems from the way in which IPD was modeled. In this study, ξ measured the amount of IPD since the previous testing occasion. Therefore, the maximum $E(\xi)$ value was 0.40. Had ξ been modeled as the amount of IPD since the base year, the cumulative $E(\xi)$ values would have been much higher and, presumably, easier to detect. It is unclear, however, that linking to the base year is desirable given that (as will be discussed shortly) a large IPD effect size may worsen linking results.

Another point of caution is that FP rates in Study 2 are bounded by those of the pre-analyses—5% and 10%, and it was not assumed that they are controlled at a certain

rate. The purpose of this study is not powerful detection of IPD with reasonable Type I error control, but improvement of linking results by incorporating IPD detection into linking procedures.

Generally, as expected, having a smaller, but relatively pure anchor (95% / 10%) appeared to lead to better results than having a larger, less pure anchor. It is hard to generalize this finding, however, because we did not test this condition more extensively. It is true that the purer anchor improved linking results in almost all conditions, but, as shown in Tables 6 to 9, μ estimates in Year 5 in Study 2 had larger biases and larger RMSEs in the 95% / 10% condition than in the 70% / 5% condition. This may be due to the effects of a smaller (although purer) anchor size. The tradeoff between anchor purity and anchor size requires further exploration.

Larger IPD effect size ($\delta = .4$) only helped with detection of TPs and generally made it worse to recover μ and ξ parameters. It is impossible for researchers to choose the effect size of IPD in practice, and larger effect sizes were usually studied in the context of IPD detection. However, this study showed the relationship of effect sizes with both IPD detection and linking. If precise linking is a bigger concern, linking results need to be more cautiously interpreted when large IPD effects are found.

Although this study successfully demonstrated how the suggested model functions in various conditions, there are limitations or future research suggestions of this study. First, we did not present whether this method performs better or worse than other methods. It is expected that this method will produce better results since errors are reduced in this unified framework; however, we postponed comparison with other methods until this method is more fully established. Second, as aforementioned, relatively poor performance of IPD detection warrants further study. Also, it is

worthwhile to examine if better detection of IPD through an iterative procedure to purify the anchor items can improve linking. Third, it will be of interest to observe how this method functions in other various conditions, such as studying what changes will be observed if both negative and positive IPD items are simulated or less favorable TP / FP ratios are used in pre-analyses. Fourth, since an MCMC approach was applied here, the effects of priors are of interest too. Finally, larger number of replications will help show a clearer pattern for all conditions.

In educational measurement it is often necessary to compare ability estimates obtained from different forms of a test. For correct inference from linked test scores, it is important to reduce linking errors. Although there have been endeavors to improve linking methods, they have been vulnerable to problems such as IPD, and these problems were considered unrelated to linking. However, this study provided a framework for simultaneously considering linking and IPD, thereby minimizing linking error due to IPD and improving the quality of linking over time.

References

Best, N., Cowles, M.K., & Vines, K. (1996). CODA*: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.30. Cambridge, UK: MRC Biostatistics Unit.

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

Chu, K. & Kamata, A. (2005). Test equating in the presence of DIF items. *Journal of Applied Measurement*, *6*, 342-354.

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

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

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum.

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

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

Shealy, R., & Stout, W.F. (1993). An item response theory model for test bias. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Spiegelhalter D. J., Thomas A., Best N.G., & Lunn D. (2003). WinBUGS Version 1.4 User Manual. MRC Biostatistics Unit, Cambridge.

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

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*. 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*. Hillsdale, NJ: Erlbaum.

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). *The impact of compounding item parameter drift on ability estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Table 1. Mean true positive (TP) and false positive (FP) rates

	$\delta = .25$	$\delta = .4$
TP	.60	.84
FP	.04	.19

Table 2. Mean bias and RMSE of ξ estimates

	$\delta = .25$	$\delta = .4$
Bias (ξ)	.00	-.06
RMSE (ξ)	.12	.12

Table 3. Bias, RMSE, and correlation coefficients of difficulty parameters for Year 2

	$\delta = .25$	$\delta = .4$
Bias (b_2)	.01	-.01
RMSE (b_2)	.17	.13
$\mathrm{Cor}\left(\hat{b}_2,\ b_2\right)$	.99	.99

Table 4. Mean FP rates of IPD-detected items

	μ fixed		μ increased	
	70% / 5%	95% / 10%	70% / 5%	95% / 10%
δ = .25	.01	.03	.01	.03
δ = .4	.01	.03	.01	.04

Table 5. Mean TP rates of IPD-detected items

		μ fixed		μ increased	
		70% / 5%	95% / 10%	70% / 5%	95% / 10%
CIPD	δ = .25	.29	.36	.35	.36
	δ = .4	.35	.42	.41	.47
RIPD	δ = .25	.49	.60	.55	.64
	δ = .4	.60	.73	.62	.76

Table 6. Mean bias (μ) when μ fixed

Year		1	2	3	4	5
True μ		0	0	0	0	0
70% / 5%	δ = .25	0*	-.05	-.01	-.01	.06
	δ = .4	0*	-.04	.01	.03	.05
95% / 10%	δ = .25	0*	-.03	.00	.02	.09
	δ = .4	0*	-.03	.05	.08	.12

* Fixed for model identification

Table 7. Mean bias (μ) when μ increased

Year		1	2	3	4	5
True μ		0	.15	.30	.45	.60
70% / 5%	δ = .25	0*	-.07	-.04	-.01	.01
	δ = .4	0*	-.04	.03	.06	.13
95% / 10%	δ = .25	0*	-.05	-.02	.01	.05
	δ = .4	0*	-.00	.05	.08	.18

* Fixed for model identification

Table 8. Mean RMSE (μ) when μ fixed

Year		1	2	3	4	5
True μ		0	0	0	0	0
70% / 5%	δ = .25	0*	.05	.05	.02	.07
	δ = .4	0*	.07	.03	.04	.07
95% / 10%	δ = .25	0*	.04	.04	.03	.10
	δ = .4	0*	.05	.06	.09	.12

* Fixed for model identification

Table 9. Mean RMSE (μ) when μ increased

Year		1	2	3	4	5
True μ		0	.15	.30	.45	.60
70% / 5%	δ = .25	0*	.08	.05	.05	.04
	δ = .4	0*	.06	.11	.09	.16
95% / 10%	δ = .25	0*	.07	.03	.04	.06
	δ = .4	0*	.05	.10	.09	.19

* Fixed for model identification

Table 10. Mean bias (ξ)

	μ fixed		μ increased	
	70% / 5%	95% / 10%	70% / 5%	95% / 10%
δ = .25	-.11	-.08	-.10	-.09
δ = .4	-.17	-.14	-.17	-.13

Table 11. Mean RMSE (ξ)

	μ fixed		μ increased	
	70% / 5%	95% / 10%	70% / 5%	95% / 10%
δ = .25	.20	.18	.21	.19
δ = .4	.30	.26	.29	.26