

# Linking for the general diagnostic model

Xueli Xu and Matthias von Davier

*Educational Testing Service, Princeton, New Jersey, United States*

This study analyzed National Assessment of Educational Progress (NAEP) reading data using a general diagnostic model (GDM) in order to investigate and compare three strategies for linking two consecutive assessments. These strategies are compared in terms of marginal and joint expectations of skills, joint probabilities of skill patterns, and item parameter estimates. The results indicate that fixing item parameter values at their previously calibrated values is sufficient to establish a comparable scale for the subsequent year.

## INTRODUCTION

Cognitive diagnosis models (DiBello, Stout, & Roussos, 1995; Junker & Sijtsma, 2000; Maris, 1999; Tatsuoaka, 1983; von Davier 2005; von Davier & Rost, 2006) have been developed for in-depth analysis of item response data. In such models, the latent abilities or skill profiles are represented by a discrete set of real valued numbers. For example, one can specify  $\{0, 1\}$  for skill spaces with mastery/non-mastery status or  $\{-4.0, -3.8, -3.6, \dots, +3.6, +3.8, +4.0\}$  for skill spaces with more than two levels that emulate unidimensional item response theory (IRT) models. The non-continuous nature of the skill profiles makes the linking across assessments non-trivial. It is appropriate to use a linking strategy in IRT models based on linear transformations when the ability distribution is assumed to follow a standard normal distribution. However, the linear linking approach might not be appropriate for discrete latent skills. Our primary goal in this article is to compare three proposed linking strategies with respect to various aspects by using a general diagnostic model (GDM) containing discrete skill profiles. The article is organized as follows. Section 1 gives a brief introduction to GDMs in general and the model used in our study in particular. In Section 2, we introduce three proposed linking strategies. Section 3 outlines our evaluation criteria, the data we drew on, and our results, including the results of analyses we conducted relative to key subgroups from our datasets. In Section 4, we complete the article by presenting a brief discussion and conclusion.

## GENERAL DIAGNOSTIC MODELS

The general diagnostic model or GDM (von Davier, 2005) is a framework that allows researchers to integrate approaches involving confirmatory multidimensional models with discrete latent trait variables. Within the GDM framework, the flexible form of the functioning of skills (cognitive attributes) allows specification of many well-known psychometric models, such as IRT models (Lord & Novick, 1968), the fusion model (DiBello et al., 1995; Hartz, 2002), and various IRT models (for an overview, see von Davier & Rost, 2006).

The special form of the GDM that we use in our study, suitable for dichotomous and partial credit data, is represented by the following equation:

$$P(X=x | \beta_i, \alpha, q_i, \gamma_i) = \frac{\exp [\beta_{xi} + x \sum_{k=1}^K \gamma_{ik} q_{ik} \alpha_k]}{1 + \sum_{y=1}^{m_i} \exp [\beta_{yi} + y \sum_{k=1}^K \gamma_{ik} q_{ik} \alpha_k]}$$

In this equation,  $q_{ik}$  is an entry of the Q-matrix, which specifies the correspondence between item  $i$  and skill  $k$ . If skill  $k$  is required to solve item  $i$ , then  $q_{ik}=1$ ; otherwise,  $q_{ik}=0$ . The total number of skills is denoted by  $K$ . Content experts prespecify the Q-matrix, which represents a hypothesis about the relationship between students' skills and students' item responses. Thus, in the above equation,  $y$  is an index for possible scores for item  $i$ , and  $m_i$  denotes the maximum score for this item. According to the equation (1), the probability of obtaining score  $x$  on item  $i$  depends on the item parameters  $\beta_{xi}$ ,  $\beta_{yi}$ ,  $\gamma_{ik}$  and on the student skill profile  $\alpha_k$ . In this model, the values  $\alpha_k$ 's take on  $\alpha_k$  a finite set of real valued numbers that the user sets in his or her model specification.

Similar to IRT models, GDMs require that certain conditions are met to remove the indeterminacy of the scale. Different methods can be used to determine the scale. Thus, for example,  $\beta_{11} = 0$  can be fixed to a certain constant and some or all slopes set to fixed constants. For example,  $\gamma_{11}$  and  $\gamma_{1k}$  for  $K > 1$ , or the mean of the difficulties as well as the (log)-average of the slopes are set to constant values. Alternatively, in models with several ability levels, the mean and variance of the ability variables can be fixed to certain values, much like the commonly used assumption of a standard normal distribution in IRT models.

## LINKING STRATEGIES

Trend maintenance is an important concept in most large-scale assessments with multiple cycles. A considerable portion of items is common across two consecutive assessments designed to establish or continue the trend. In the remainder of this article, we denote these two consecutive assessments via  $Y1$  and  $Y2$ , set in chronological order. Because we assumed that the scale of  $Y1$  was established from a previous calibration before consideration was given to the linkage between  $Y1$  and  $Y2$ , we denote this previous calibration of  $Y1$  as  $Y1$  calibration throughout this article.

A *concurrent* calibration strategy is used in the operational linking analysis of National Assessment of Educational Progress (NAEP) data (Mislevy, 1992; Muraki & Hombo, 1999). The strategy includes three steps that endeavor to build a linkage between the *Y1 calibration* and the *Y2*. The first step involves establishing a common scale for *Y1* and *Y2* through a concurrent calibration of the data from *Y1* and *Y2*, and setting common items to have the same item parameter estimates. This step makes it possible to obtain the mean and variance of the latent ability for students in *Y1* and *Y2* in the concurrent calibration. The second step is to form a bridge between the *Y1 calibration* and the concurrent calibration by finding a linear transformation that makes the mean and the variance of the latent ability for students in *Y1* from both calibrations equal to each other. Finally, the third step involves establishing the link between the *Y1 calibration* and the *Y2* by applying this linear transformation to *Y2* from the concurrent calibration.

This concurrent calibration strategy is valid when the latent ability is assumed to follow a normal distribution. The reason for this is that, with normal distributions, any two distributions can be perfectly matched by a location and scale transformation. This is not true, however, for more general distributions, such as those that require full specification of three or more parameters. In addition, as Haberman (2005) has shown, attempts to use two-parameter-logistic (2PL) and three-parameter-logistic (3PL) models with ability distributions that are more general in nature than the standard normal distribution require careful work. Specifically, the linear transformation in Steps 2 and 3 is not appropriate for discrete latent variables. For example, if the latent skill is prespecified to have six real-valued levels  $\{-2, -1, -0.5, 0.5, 1, 2\}$ , any linear transformations other than identity (slope=1, and intercept=0) and negative identity (slope=-1, and intercept=0) are not valid. A linear transformation with slope=2 and intercept=0 leads to a set of  $\{-4, -2, -1, 1, 2, 4\}$ , which is out of the range of the original set of  $\alpha_k$ . So, in developing the linking strategy under discrete latent trait models, we have to use methods that avoid the need for linear transformations. The three strategies that we consider in this article are all based on the concurrent calibration described above. Strategy 2 is actually the first step of the concurrent calibration linking. Although Strategy 2 cannot establish a good link because Steps 2 and 3 are missing, we have included it for comparison with Strategy 1. We consider Strategy 1 to be more stringent than Strategy 2 because the parameter estimates for the common items are fixed, like those in the *Y1 calibration*. Strategy 3 also relies on a strong assumption regarding the role of the common items. Our hypothesis therefore is that the common items will be sufficient to build a link between the *Y1 calibration* and the *Y2*. Moreover, although we knew it was likely that Strategies 1 and 3 would be the same when no constraints were imposed on item parameters, we recognized that certain constraints must be imposed in many situations involving GDMs in order to make the models identifiable. While in our case these constraints would make Strategies 1 and 3 different, we suspected that the differences in most cases would be small.

The details of our three linking strategies follow:

- *Linking Strategy 1:* Under this strategy, *Y1* and *Y2* are calibrated concurrently, with the common items fixed at the values obtained from the *Y1 calibration*. This calibration does not reestimate the item parameters of the common items for *Y2* but rather assumes the parameters of these items are fixed at known values. In addition, items not common to *Y1* and *Y2* are reestimated in a joint calibration with unique sets of parameters for each of the years.
- *Linking Strategy 2:* This strategy calibrates *Y1* and *Y2* concurrently, with the common items set to be equal across (for this study) two years. This procedure involves reestimating all item parameters in a joint calibration, while assuming that the parameters of items common to *Y1* and *Y2* are equal and do not change over assessment cycles.
- *Linking Strategy 3:* This strategy establishes the link by calibrating the *Y2* assessment data separately, with common items fixed at the values obtained from the *Y1 calibration*.

## EVALUATION CRITERIA

Before presenting our analysis, we briefly outline the criteria that we used to evaluate the different strategies. Within an IRT modeling framework, a good recovery of the basic characteristics of *Y1* is often used as the criterion for a good linking. For example, in a concurrent calibration, the rationale behind Steps 2 and 3 is to make sure that the characteristics of *Y1* stay the same from the *Y1 calibration* to the concurrent calibration. If we assume a normal distribution for the latent ability in the IRT models, then the mean and the variance are sufficient to maintain the shape of the latent ability. However, the mean and variance are no longer sufficient for a discrete latent skill distribution. Thus, when we estimate multidimensional skills simultaneously, we should estimate the joint probabilities so as to describe the characteristics of the latent skill distributions. Accordingly, in this study, we also report, in addition to the joint probability distributions, the joint expectation of latent skills and the marginal probability of skills for key subgroups as the criteria by which to evaluate the three different linking strategies.

## DATA

The data that we used to compare the three linking strategies were drawn from two NAEP Grade 4 reading assessments. Our first dataset contained a subset of the 2003 assessment data, and our second dataset was a subset of the 2005 assessment. The dataset from 2003 contained 47,817 students' responses to 102 items under a partially balanced incomplete block (pBIB) design from two subscales ( $K=2$ ): reading for literary experience and reading to gain information. The 2005 dataset included 41,420 students' responses to 99 items under the pBIB design employing the same two subscales. The two assessments had 69 items in common. The data from 2003 served as the *Y1* data, while the data from 2005 served as the *Y2* data.

## ANALYSIS

In the reading framework, each item from both the *Y1* and the *Y2* data is assigned to one of the two reading subscales—reading for literary experience and reading for information. The correspondence between items and subscales defined in the framework served as our Q-matrix. This setting is equivalent to a two-dimensional IRT model that has a simple structure represented by the allocation of each item to only one of the subscales. Because model comparisons were not a focus of our linking study, we considered no other alternative Q-matrices in relation to it. Our primary goal relative to the comparison between Strategies 1 and 2 was to determine if Strategy 1 could reproduce the scale set by the *Y1* calibration. We considered that if Strategy 1 outperformed Strategy 2, then the comparison between Strategies 1 and 3 would allow us to determine if the release of concurrent calibration in Strategy 3 would allow us to recover the *Y1* characters.

## RESULTS

The results are organized as follows: comparison between Strategies 1 and 2; comparison between the three strategies; and comparisons in terms of key subgroup statistics.

### Comparison 1: Strategy 1 versus Strategy 2

The comparisons in this section are based on using fit statistics, the joint probabilities of skill patterns, and the joint and marginal expectations of skills. The fit statistics that we used in this study included the log-likelihood and the Akaike Information Criterion or AIC (Akaike, 1974) index. The AIC is defined as  $-2\ln(L) + 2p$ , where  $\ln(L)$  is the log-likelihood of the data under the model and  $p$  is the number of parameters in the model. For a given dataset, the larger the log-likelihood, the better the model fit, and the smaller the AIC value, the better the model fit.

Table 1 gives information on our model's fit statistics. Note that the number of parameters in the table is much smaller for Strategy 1 than for Strategy 2. This is because the parameters for the common items were fixed given that they were already known from the 2003 separate calibration. Therefore, we can argue that the *actual* count of parameters is unknown for this model because it involves the 2003 data, which were separately used to determine the common item parameters in this strategy. Nevertheless, a comparison solely in terms of likelihood indicates that the differences between the two strategies were not huge for these calibrations.

**Table 1: Model fit comparisons for Strategies 1 and 2**

Linking	Model parameters	Log-likelihood	AIC
Strategy 1	165	-994579.93	1989469
Strategy 2	321	-993799.31	1988200

Figure 1 compares the estimated joint probabilities of skill patterns for the 2003 data ( $Y1$ ) obtained from using Strategies 1 and 2 with those obtained from the separate calibration of 2003 (the  $Y1$  calibration). If a common scale had been maintained across calibrations, we would expect all the estimated joint probabilities within the figure to be very close to one another. Within each plot of the figure, the x-axis stands for the estimated joint probability of skill patterns for the 2003 sample from the  $Y1$  calibration, while the y-axis represents the corresponding probability from using either Strategy 1 or Strategy 2. The left-hand panel gives the contrast between the separate calibration and Strategy 1, while the right-hand panel gives the contrast between the separate calibration and Strategy 2.

Figure 2 shows the estimated joint expectation of skills for the 2003 students under Strategies 1 and 2 against those from the  $Y1$  calibration. Here, the expectation is calculated by  $E(\alpha_1, \alpha_2 | v_j)$  for each person  $v_j$ , and so we could expect the estimates for the same students from the different methods to be very close to one another if a common scale had been maintained. Again, within each plot, the x-axis stands for the estimated joint expectation for the 2003 sample from the  $Y1$  calibration, while the y-axis represents the corresponding expectation from using either Strategy 1 or Strategy 2.

Figure 3 presents the differences in marginal skill expectations for the 2003 students in the form of boxplots. Specifically, the left-hand graph shows the difference between the  $Y1$  calibration and the use of Strategy 1, while the right-hand graph represents the difference between the  $Y1$  calibration and the use of Strategy 2. The numbers 1 and 2 alongside the x-axis in each graph represent the two reading subscales, and we used  $E(\alpha_k | v_j)$  to calculate the marginal expectation for skill  $k$  for each person  $v_j$ .

Figure 1: Joint probability comparison for the 2003 data

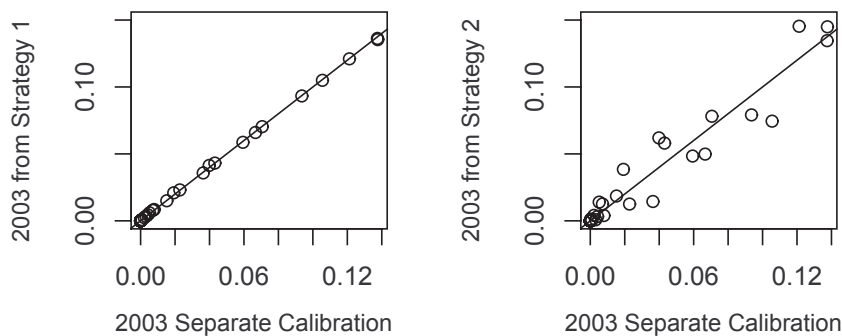


Figure 2: Joint expectation comparison for the 2003 data

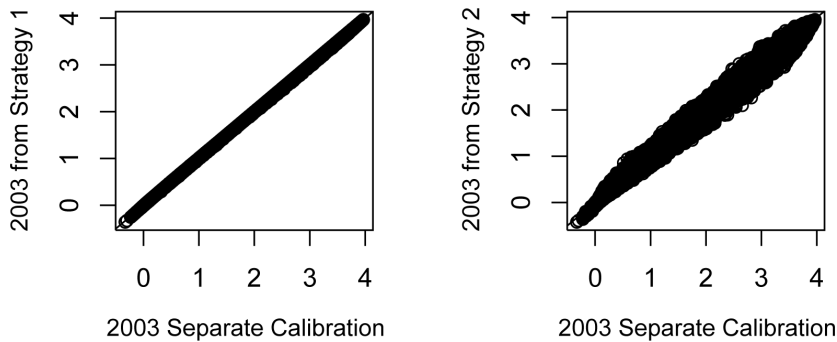
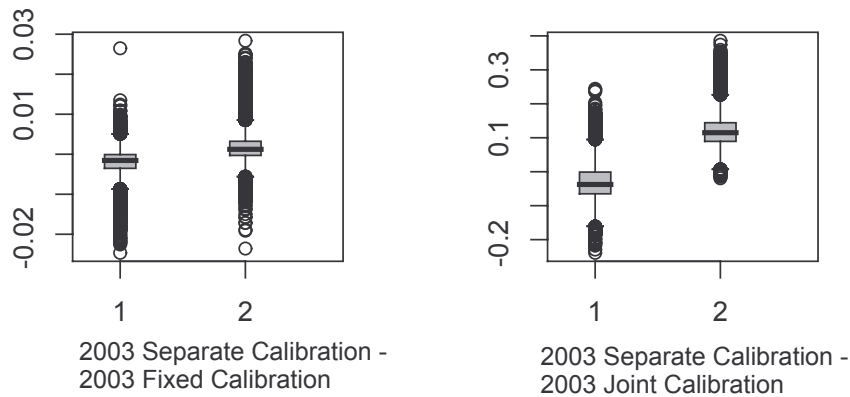


Figure 3: Marginal expectation comparison for the 2003 data



We can see from Figures 1 to 3 that the deviations from the *Y1 calibration* in terms of various statistics were smaller when we used Strategy 1 as compared to Strategy 2. Even though concurrent calibration (Strategy 2) produced a common scale for Years 2003 and 2005, it may not have produced the same scale as that established from the *Y1 calibration*. Compared to Strategy 2, Strategy 1 utilized a stronger link to connect these two consecutive assessments by using concurrent calibration coupled with fixed common-item parameter values. Therefore, Strategy 1 showed much smaller deviations from the *Y1 calibration* than did Strategy 2.

We could argue that these results would not have held if there had been fewer common items between the two tests. To answer this consideration, we investigated a case in which only 25 items were common to the two years. We randomly selected these 25 items from the original 69 common items. Table 2 gives the model-fit information for these analyses. Again, due to the fixing of item parameters in Strategy 1, the number of parameters shown in Table 2 for Strategy 1 is not accurate. Nevertheless, the difference between AIC is not large for the two strategies.

Table 2: Model fits of Strategies 1 and 2 with only 25 common items

Linking	Model parameters	Log-likelihood	AIC
Strategy 1	369	-993761.63	1988215
Strategy 2	424	-993408.44	1987612

Figures 4 to 6 show the same configuration of results as that shown in Figures 1 to 3, but this time the results relate to the 25-common-items case. The fact that the three figures present results similar to those in Figures 1 to 3 indicates that the scale established by Strategy 1 remained robust even when fewer common items were present.

Figure 4: Joint probability comparison with 25 common items from the 2003 assessment

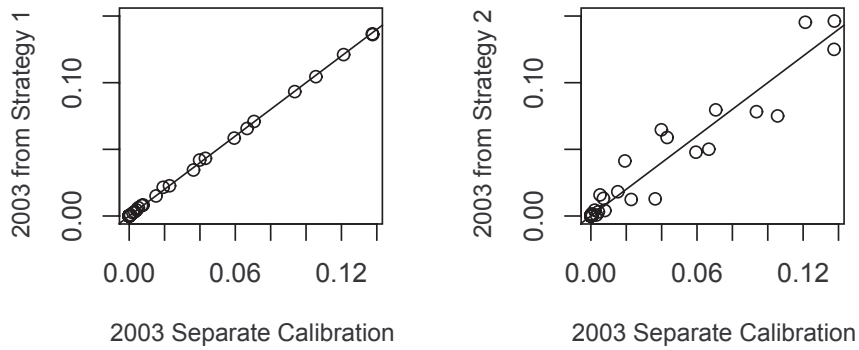


Figure 5: Joint expectation comparison with 25 common items from the 2003 assessment

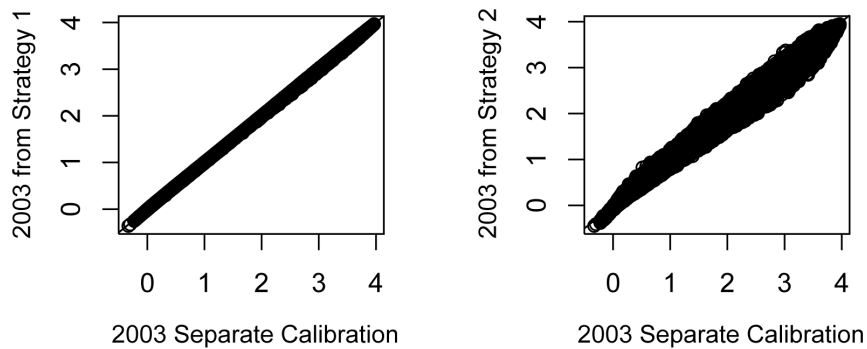
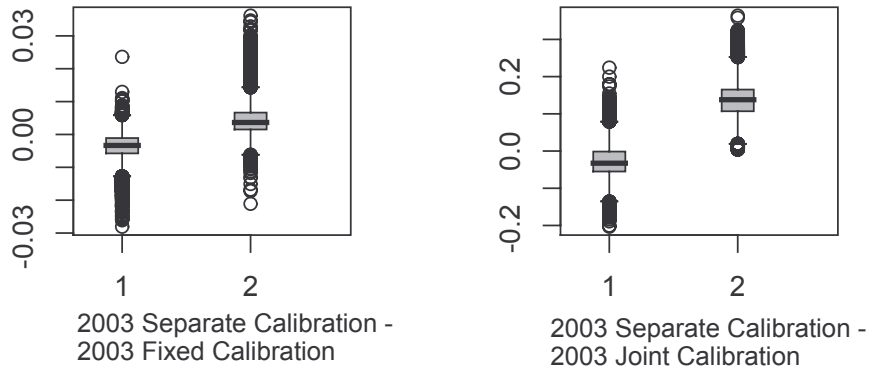


Figure 6: Marginal expectation comparison with 25 common items from the 2003 assessment



### Comparison 2: Three Strategies

Our intention with the comparisons set out in this section was to determine if we could establish the link between the two assessment years by dropping the concurrent calibration and using only fixed-item parameters for the subsequent calibrations. This required us to apply the common-item parameters obtained from the *Y1 calibration* directly to the analysis of the 2005 data. We conducted the comparison by using joint skill probabilities and marginal skill expectations. We also conducted our comparisons with either the 69 items or the 25 items held in common between the two tests.

The difference in the joint probabilities of skill patterns between the 2005 students and the 2003 students under the three different strategies is shown in Figure 7 in the form of boxplots. The numbers along the x-axis in Figure 7 stand for strategy ID (i.e., Strategies 1, 2, or 3). If a scale identical to that from the 2003 separate calibration had been set up by Strategy 3, then the boxplots for Strategy 1 and Strategy 3 depicted in the figure should be similar to each other. However, if the scale had not been identical, then the boxplots that we see in the figure for Strategies 1 and 2 should be similar to each other. Our analysis produced the former result, and we also obtained a similar result with the 25 common-items configuration, as shown in Figure 8.

Figures 9 and 10 show, again via boxplots, comparisons of the marginal skill expectations obtained from Strategies 2 and 3 with those from Strategy 1. In each graph, the numbers 1 and 2 along the x-axis stand for the two subscales measured in reading. The left-hand graph presents the difference between Strategies 3 and 1, while the right-hand panel illustrates the difference between Strategies 2 and 1. If an identical scale had been established through fixing the item parameter values, and not by concurrent calibration, then we could assume that the difference between Strategies 1 and 3 would be smaller than the difference between Strategies 1 and 2. The results shown in the figures confirm this, since the boxplots for both reading subscales shown in the left-hand graph are considerably more concentrated around 0 than are those in the right-hand graph.

Figure 7: Joint probability comparison: 2005 minus 2003

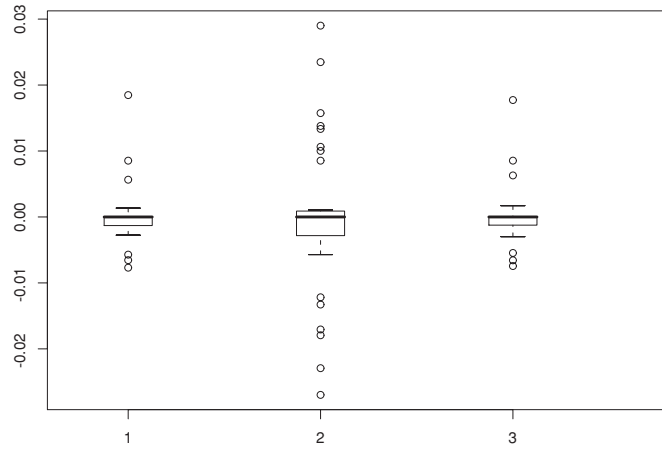


Figure 8: Joint probability comparison: 2005 minus 2003 with 25 common items

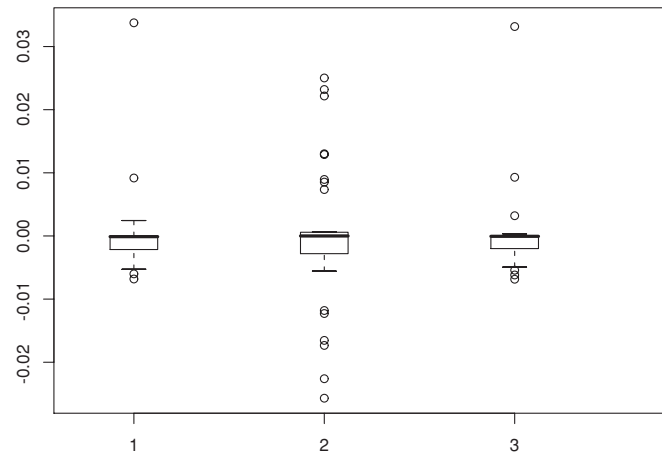


Figure 9: Marginal expectation comparison for 2005 data

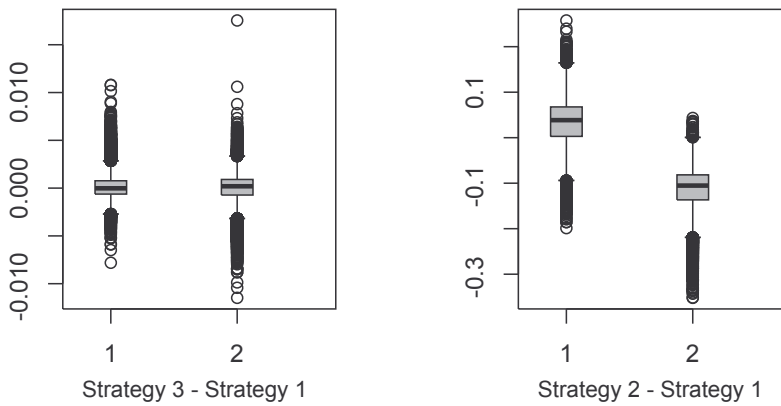
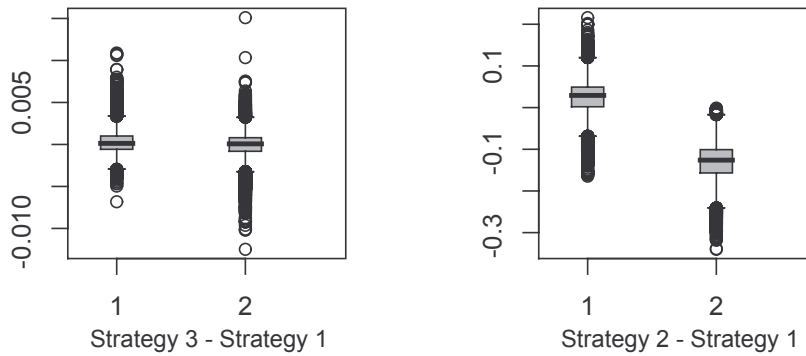


Figure 10: Marginal expectation comparison for 2005 data with the 25 common items



In summary, the comparisons in this section indicate that it *is* possible to reproduce in the 2005 separate calibration a scale identical to the 2003 separate calibration by fixing the common item parameters at the estimates obtained from the 2003 separate calibration.

### Key Subgroups Comparison

Because it is important, relative to operational NAEP reporting purposes, to consider statistics relating to key subgroups, such as the mean, standard deviation, and quantiles, we decided to consider the skills distributions for the subgroups within the cognitive diagnosis framework that had an equivalent aggregation level. More specifically, we compared the skill distributions of several key subgroups (race/ethnicity and gender) in the 2005 assessment across linking strategies. We also compared the skill profiles of key subgroups in the 2003 assessment across strategies and with those obtained from the separate calibration of 2003. Given that the results of the case with 25 common items and the case with 69 common items were again similar to each other, we report only the results of the case with 25 common items.

In the following comparisons, we derived the skill profiles for subgroups on the basis of a single-group assumption. This means that we set all subgroups to have the same prior distribution for the latent classes. We then calculated the skill profile for a subgroup by taking a weighted average of the skill profiles of students in the subgroup; here, the weights were the student weights used in the NAEP operational analysis.

We compared the skill profiles for subgroups from the 2003 assessment between the *Y1 calibration* and Strategies 1 and 2. Figures 11 and 12 show the differences in estimated marginal skill distributions. In each graph, the subgroups are represented by a capitalized initial letter. Thus, A, B, F, H, M, and W stand for Asian, Black, Female, Hispanic, Male, and White student groups. Although we might consider the differences between the *Y1 calibration* and Strategy 2 shown in Figure 12 to be small (within a range of -0.04–0.04), the differences between the *Y1 calibration* and

Strategy 1 (shown in Figure 11) are much smaller. Thus, Strategy 1 leads to an almost identical scale to the scale from the *Y1 calibration*.

Figure 11: Differences in marginal skill profile: Separate versus Strategy 1 (2003 data)

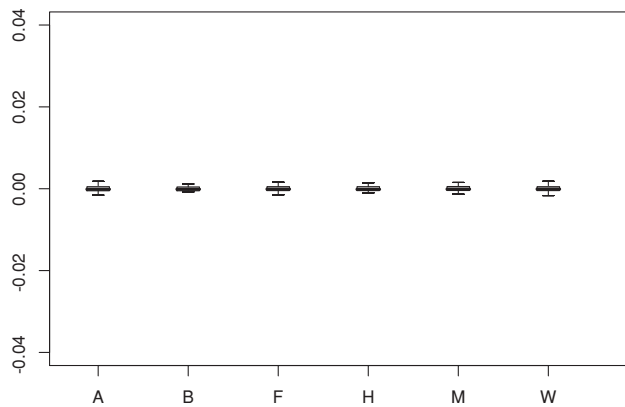
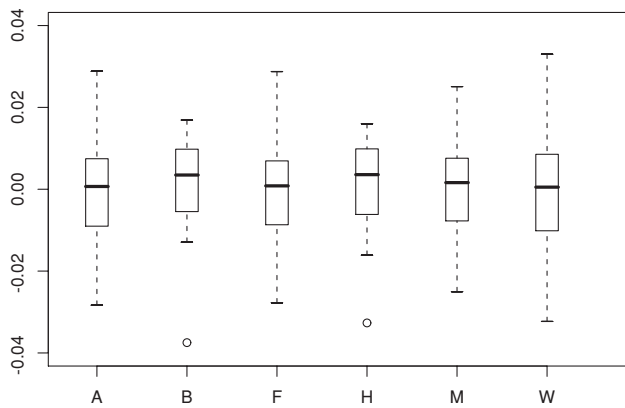


Figure 12: Differences in marginal skill profile: Separate versus Strategy 2 (2003 data)



Figures 13 and 14 show the results of our comparison (between Strategy 3 and Strategies 1 and 2) of the skill profiles for the subgroups from the 2005 assessment. From these figures, we can see that Strategy 1 is much closer to Strategy 3 than it is to Strategy 2 in terms of the estimated marginal skill profile for the key subgroups. This outcome suggests that, in the case of linking these two NAEP assessments under the GDM framework, it is possible to set and reproduce a scale by fixing the values of the common items in the two assessments, even when they hold only 25 (approximately 25% of the entire test) items in common.

Figure 13: Differences in the marginal skill profile: Strategy 3 versus Strategy 1 (2005 data)

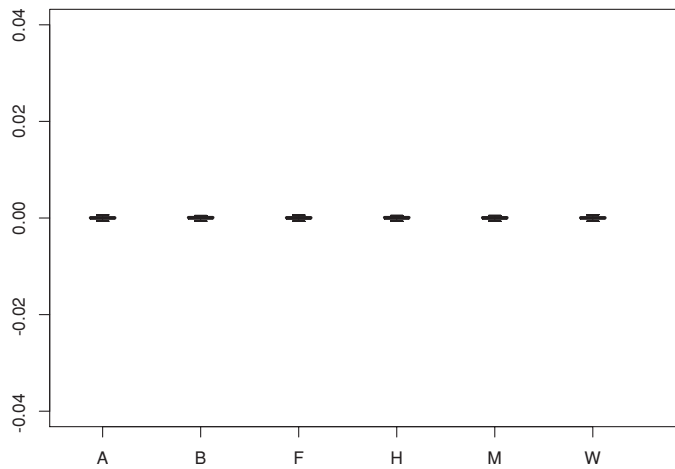
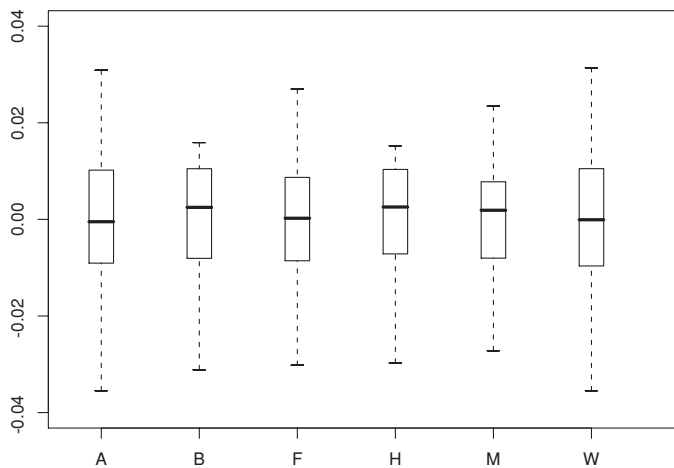


Figure 14: Differences in the marginal skill profile: Strategy 3 versus Strategy 2 (2005 data)



## DISCUSSION AND CONCLUSION

Although the non-continuous nature of the skill locations in the GDM limits searches for appropriate linking methods, the research question that needs to be answered is what leads to a scale that is maintained across assessments and links to the scale of the *Y1 calibration*. Certainly, the bridge between target (*Y2*) and baseline year (*Y1*) is built through common items, but the nature and extent of the necessary constraints are not self-evident. We therefore compared three different strategies in this study. As previously mentioned, these three strategies are variations from the concurrent calibration linking used in NAEP operations. Often, a concurrent calibration linking consists of three steps of calibration and transformation, and Strategy 2 in this present study was indeed the first step of the concurrent calibration linking, especially given the need to intentionally drop the other steps due to their inappropriateness for discrete latent skills/abilities.

Strategy 1 produced a stronger link than Strategy 2 by fixing the common items parameters at known values from the *Y1 calibration* in addition to the concurrent calibration. With Strategy 3—a simplified version of Strategy 1—we dropped the concurrent calibration and kept the common-item parameters from the *Y1 calibration* fixed at known values. In fact, Strategy 3 would have been identical to Strategy 1 if no constraint had been imposed on the item parameter estimation procedure. Even when we did place certain constraints on the estimation process, we observed only slight differences for Strategies 1 and 3, as shown in Figures 7 to 10 and Figure 13.

All the results in our study empirically demonstrated that one linking strategy—the concurrent calibration of two adjacent assessments—is not necessary when the common items are fixed at the values obtained from the *Y1 calibration*. These similar results even held up in the case where common items consisted of only 25% of the whole NAEP assessment. Generalizing this result is not advised, however, because it may not hold up in studies with different procedures for assessment development, block formation, item flagging, and selection for subsequent assessment.

To make sure the conclusion remained true for the case where only 25 items were held in common across both tests, we carried out two additional analyses based on different sets of 25 items, and obtained similar results to those shown in this current article. We also analyzed data from a Grade 8 NAEP reading assessment in 2003 and 2005, and again drew similar conclusions from these analyses. Although the purpose of our study was not focused on model comparisons, we have to mention one special model case where only two levels (mastery and non-mastery) were specified for each cognitive skill. On running such cases, we found that the concurrent calibration of 2003 and 2005 assessments was able, as in Strategy 2, to reproduce the scale established by the *Y1 calibration*.

As discussed earlier in this article, we conducted our analysis on the basis of a single-group assumption, that is, a one-skill distribution. In our future work, we intend to undertake an analysis based on a multiple-group assumption coupled with Strategy 3. Under this assumption, we will assign the subgroups unique and potentially

different prior distributions so that the skill profiles for these subgroups can be directly calculated by rerunning the software. An initial investigation of applying GDM to NAEP data (Xu & von Davier, 2006) showed that the multiple-group analysis yielded results for the racial subgroups and gender subgroups that were similar to those from the NAEP operational analyses. Our future work will also endeavor to answer additional questions, such as whether our employment of a GDM multiple-group analysis procedure will see Strategy 3 leading to a comparable scale.

## References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.
- DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan, (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Haberman, S. (2005). *Identifiability of parameters in item response models with unconstrained ability distributions* (ETS Research Report Series RR-05-24). Princeton, NJ: Educational Testing Service.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois, Champaign, IL.
- Junker, B., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65–81.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212.
- Mislevy, R. J. (1992). Scaling procedures. In E. G. Johanson & N. L. Allen (Eds.), *The NAEP 1990 technical report* (Report No. 21-TR-20, pp. 199–213). Washington DC: National Center for Education Statistics.
- Muraki, E. & Hombo, C. (1999). *Application of a multiple-group generalized partial credit model to NAEP linking procedures*. Unpublished manuscript. Princeton, NJ: Educational Testing Service.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report Series RR-05-16). Princeton, NJ: Educational Testing Service.
- von Davier, M. & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 27. Psychometrics* (pp. 643–661). Amsterdam: Elsevier.
- Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (ETS Research Report Series, RR-06-08). Princeton, NJ: Educational Testing Service.

