Research Report

Linking for the General Diagnostic Model

Xueli Xu

Matthias von Davier

February 2008

ETS RR-08-08

Linking for the General Diagnostic Model

Xueli Xu and Matthias von Davier
ETS, Princeton, NJ

February 2008

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Three strategies for linking two consecutive assessments are investigated and compared by analyzing reading data for the National Assessment of Educational Progress (NAEP) using the general diagnostic model. 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.

Key words: General diagnostic model, concurrent calibration, linking strategies, National Assessment of Educational Progress, marginal expectation, joint expectation, joint probability

1. Introduction

Cognitive diagnosis models (Tatsuoka, 1983; DiBello, Stout, & Roussos, 1995; Maris, 1999; Junker & Sijtsma, 2000; 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/nonmastery status or {-4.0, -3.8, -3.6, …, 3.6,3.8,4.0} for skill spaces with more than two levels emulating unidimensional item response theory (IRT) models. The noncontinuous nature of the skill profiles makes the linking across assessments nontrivial. 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. The primary goal of this paper is to compare three proposed linking strategies with respect to various aspects by using a general diagnostic model with discrete skill profiles. The paper is organized as follows: section 2 gives a brief introduction of the model, section 3 introduces three proposed linking strategies, section 4 introduces the data, section 5 shows the results, and section 6 includes key subgroup analysis results. A brief discussion and conclusion are included in section 7.

2. General Diagnostic Models

The general diagnostic model (GDM; von Davier, 2005) was introduced as a framework 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, 1980), the fusion model (Hartz, 2002; DiBello et al., 1995), and mixture IRT models (for an overview, see von Davier & Rost, 2006). The special form of the GDM that we used in our study, suitable for dichotomous and partial credit data, was given by

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

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*. The Q-matrix is prespecified by content area experts and represents a hypothesis about the relationship between students' skills and students' item responses.

In Equation 1, y is an index for possible scores for item i, and m_i denotes the maximum score for this item. According to Equation 1, the probability of obtaining score x on item i depends on the item parameters β_{xi}, β_{yi}, γ_{ik}, and the student skill profile α_k. In this model, the values α_k take on a finite set of real valued numbers that are set by the user in his or her model specification. Similar to IRT models, the GDM requires that certain conditions are met to remove the indeterminacy of the scale. There are different methods to determine a scale: Either one difficulty parameter (for example $\beta_{11} = 0$) as well as some or all slopes (e.g., $\gamma_{11} = 1$ and $\gamma_{1k} = 0$ for $K > 1$) are fixed to a certain constant 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.

3. 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 to establish or continue the trend. For the rest of this paper, these two consecutive assessments are denoted by *Y1* and *Y2* in chronological order. The scale of *Y1* is assumed to have been established from a previous calibration before considering the linkage between *Y1* and *Y2*. This previous calibration of *Y1* is denoted as *Y1 calibration* throughout this paper.

A concurrent calibration strategy is used in the operational linking analysis of National Assessment of Educational Progress (NAEP) data (Mislevy, 1992; Muraki & Hombo, 1999). This linking strategy includes three steps to build a linkage between *Y1 calibration* and *Y2*. In the first step, a common scale for *Y1* and *Y2* is established through a concurrent calibration of the data from *Y1* and *Y2* with common items set to have the same item parameter estimates. As a result, we obtain the mean and variance of the latent ability for students in *Y1* and *Y2* in the concurrent calibration. In the second step, a bridge between the *Y1 calibration* and the concurrent calibration is set by finding a linear transformation that will make the mean and variance of the latent ability equal for students in *Y1* from both calibrations. Finally, in the third step, the link

between *Y1 calibration* and *Y2* is established 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, since for normal distributions, any two distributions can be perfectly matched by a location and scale transformation. This is not true for more general distributions, for example, distributions that require more than three or more parameters to be specified fully. In addition, Haberman (2005) has shown that the attempts to use two-parameter-logistic (2PL) and three-parameter-logistic (3PL) models with more general ability distributions than the standard normal distribution require quite careful work. However, the linear transformation in steps 2 and 3 is not appropriate for discrete latent variables. For example, if the latent skill α_k 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 α_k. So in developing the linking strategy under discrete latent trait models, we have to use methods that avoid the need for linear transformations.

Three strategies are considered in this paper. In fact, all three strategies proposed are based on the concurrent calibration described above. Strategy 2 is indeed the first step of the concurrent calibration linking. Obviously, Strategy 2 cannot establish a good link since steps 2 and 3 are missing. But it is included for the comparison with Strategy 1. Strategy 1 is considered to be more stringent than Strategy 2 since the parameter estimates for the common items are fixed as those in *Y1 calibration*. Strategy 3 also relies on a strong assumption on the role of the common items. We hypothesize that the common items are sufficient enough to build a link between *Y1 calibration* and *Y2*. Strategies 1 and 3 will be the same when no constraints are imposed on item parameters. However, certain constraints must be imposed in many situations in order to make the models identifiable. These constraints make Strategies 1 and 3 different; although in most cases the differences are small.

The details of these three linking strategies are listed below.

- *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 will not re-estimate the item parameters of the common items for *Y2*, but rather it will assume the parameters of these items are fixed at known values. In addition, items not common to *Y1* and *Y2* will be reestimated in a joint calibration with unique sets of parameters for each of the years.

- *Linking Strategy 2:* This strategy is to calibrate *Y1* and *Y2* concurrently with the common items set to be equal across the two years. This procedure will reestimate 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:* The link is established by a strategy in which the *Y2* assessment data is calibrated separately, with common items fixed at the values obtained from the *Y1 calibration*.

4. Data, Analysis, and Results

Before the analysis, we would like to discuss the criteria used in evaluating 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 for Steps 2 and 3 is to make sure that the characteristics of *Y1* stay the same from *Y1 calibration* to the concurrent calibration. If a normal distribution is assumed for the latent ability in IRT models, the mean and 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. When we estimate multidimensional skills simultaneously, the joint probabilities should be estimated in order to describe the characteristics of the latent skill distributions. In this study, in addition to the joint probability distributions, we will also report the joint expectation of latent skills, and the marginal probability of skills for key subgroups as the criteria to evaluate the three different linking strategies.

Data

The data used to compare these three linking strategies were taken from data on two NAEP Grade 4 reading assessments. One dataset contained a subset of the 2003 assessment data, and the other 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. There were 69 items in common to these two assessments. The data from 2003 was assumed to be *Y1*, while the data from 2005 served as the *Y2* data.

Analysis and Results

Each item in both the *Y1* and *Y2* data was assigned to exactly one of the two reading subscales: reading for literary experience and reading for information. The correspondence between items and subscales serves as a Q-matrix in our analysis. This setting is equivalent to a two-dimensional IRT model with simple structure represented by the allocation of each item to only one of the subscales. Since model comparisons are not a focus of this paper, no other alternative Q-matrices were considered in this linking study. The primary goal of the comparison between Strategies 1 and 2 is to see whether Strategy 1 can reproduce the scale set by *Y1 calibration*. If Strategy 1 outperforms Strategy 2, the comparison between Strategies 1 and 3 is to see whether the release of concurrent calibration in Strategy 3 will make the recovery of *Y1* characters possible. The following result sections are organized as follows: comparison between Strategies 1 and 2, comparison among the three strategies, and comparisons in terms of key subgroup statistics.

Comparison I: Strategy 1 versus Strategy 2. The comparisons in this section are based on the use of fit statistics, the joint probabilities of skill patterns, and the joint and marginal expectations of skills. The fit statistics used in this study include the log-likelihood and the Akaike information criterion (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; the smaller the AIC value, the better the model fit. Information on model fit statistics is given in Table 1. Note that the number of parameters is much smaller for Strategy 1, which is due to the fact that the parameters for common items have been fixed as known from the 2003 separate calibration. Therefore, it can be argued that the actual count of parameters is unknown for this model, since it involves the 2003 data, which has been 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 are 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

Note. AIC = Akaike information criterion.

Figure 1 compares the estimated joint probabilities of skill patterns for data 2003 (*Y1*) obtained from using Strategies 1 and 2 with those from separate calibration of 2003 (*Y1 calibration*). All the estimated joint probabilities should be very close to each other if a common scale is maintained across calibrations. Within each plot, the *x*-axis stands for the estimated joint probability of skill patterns for 2003 sample from *Y1 calibration*, while the *y*-axis represents the corresponding probability from either using Strategy 1 or Strategy 2. The left panel gives the contrast between the separate calibration and Strategy 1, while the right panel gives the contrast between the separate calibration and Strategy 2.

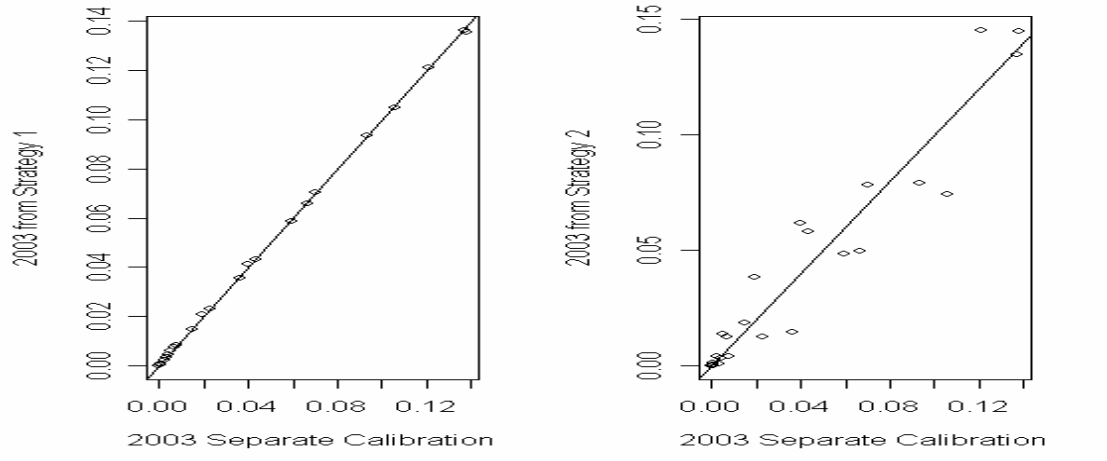

Figure 1. **Joint probability comparison for 2003.**

Figure 2 shows the estimated joint expectation of skills for the 2003 students under Strategies 1 and 2 against those from *Y1 calibration*. This expectation is calculated by $E(\alpha_1\alpha_2 \mid \nu_j)$ for each person ν_j. The estimates for the same students from different methods should be very close to each other if a common scale is 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.

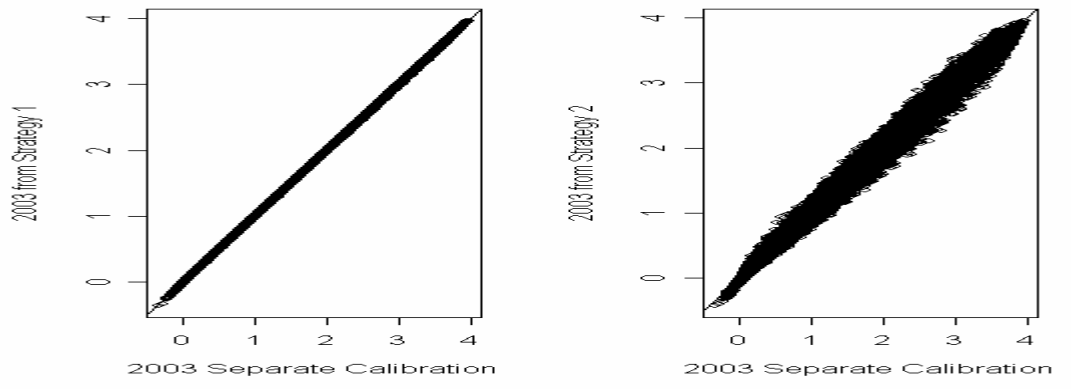

Figure 2. **Joint expectation comparison for 2003 data.**

Figure 3 presents the differences in marginal skill expectations for 2003 students in the form of boxplots. Specifically, the left graph shows the difference between *Y1 calibration* and the use of Strategy 1, while the right graph represents the difference between *Y1 calibration* and the use of Strategy 2. The numbers 1 and 2 along the *x*-axis in each graph represent the two reading subscales. The marginal expectation for skill k is calculated by $E(\alpha_k \mid \nu_j)$ for each person ν_j. It is important to note that the scale of the difference in the right graph is about 10 times greater than that in the left graph.

It can be observed from Figures 1 to 3 that the deviations from *Y1 calibration* in terms of various statistics are smaller when using Strategy 1 as compared to Strategy 2. Even though concurrent calibration (Strategy 2) produced a common scale for 2003 and 2005, it may not produce the same scale as that established from *Y1 calibration*. Compared to Strategy 2, Strategy

1 utilizes a stronger link to connect these two consecutive assessments by using concurrent calibration coupled with fixed common item parameter values. Therefore, Strategy 1 shows much smaller deviations from the *Y1 calibration* than Strategy 2 does.

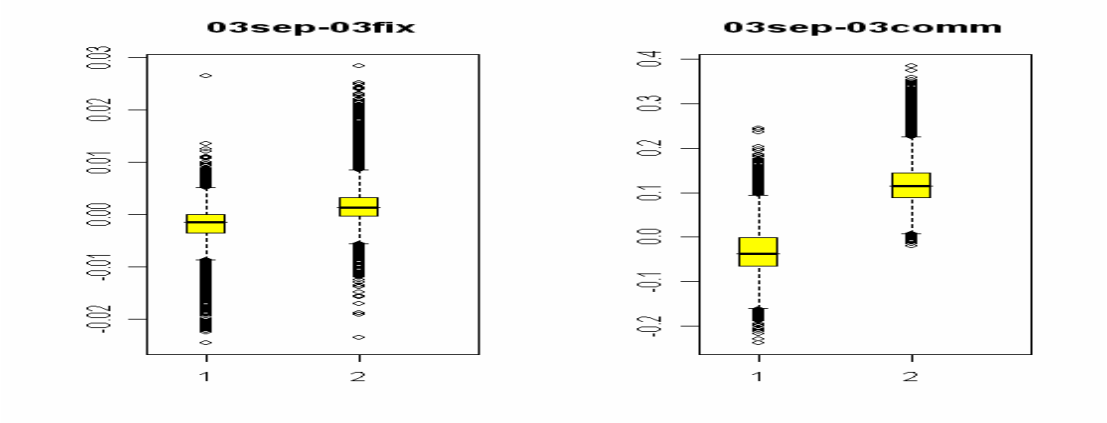

Figure 3. **Marginal expectation comparison for 2003 data.**

One might argue that the above results might not be true when fewer common items exist between the two tests. To answer this question, we investigated the case where only 25 items were common to two years. These 25 items were randomly selected 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 these two strategies. Figures 4 to 6 show the results corresponding to Figures 1 to 3 for the 25 common items.

Table 2

Model Fits of Strategies 1 and 2 With Only 25 Common Items

Model	Parameters	Log-likelihood	AIC
Strategy 1	369	-993,761.63	1988215
Strategy 2	424	-993,408.44	1987612

Note: AIC = Akaike information criterion.

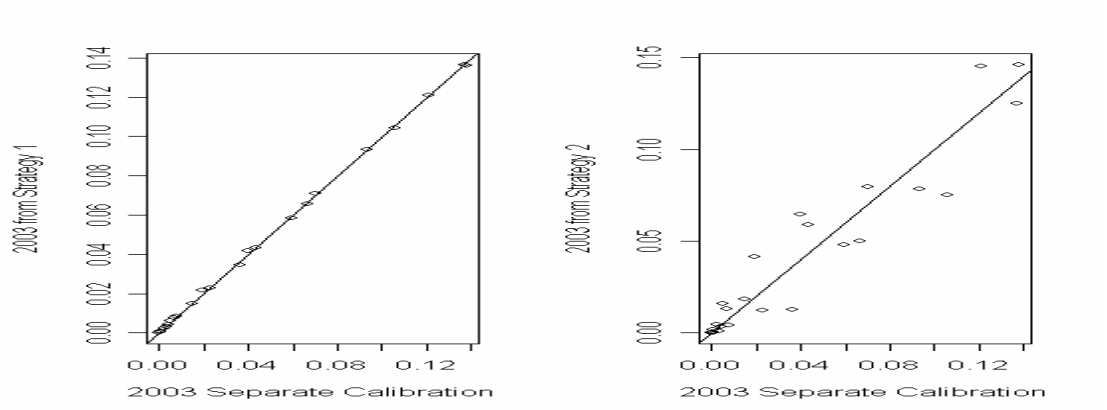

Figure 4. **Joint probability comparison for 2003 with 25 common items.**

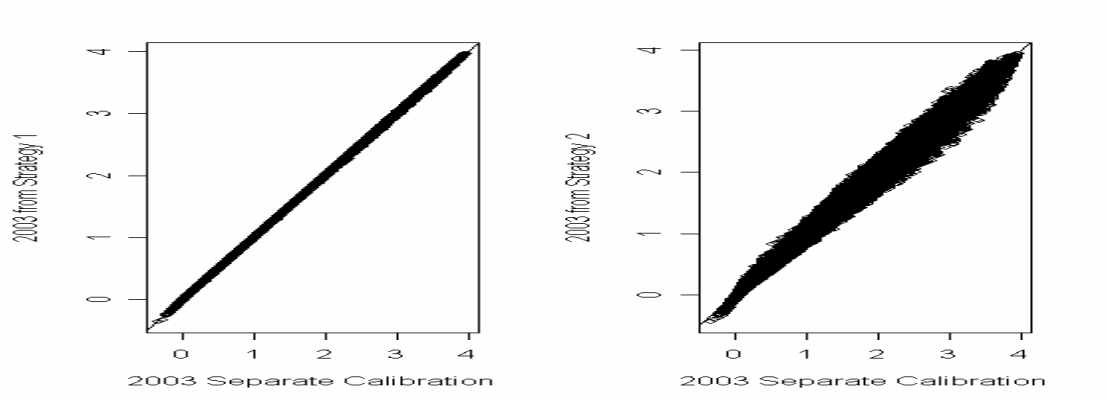

Figure 5. **Joint expectation comparison for 2003 with 25 common items.**

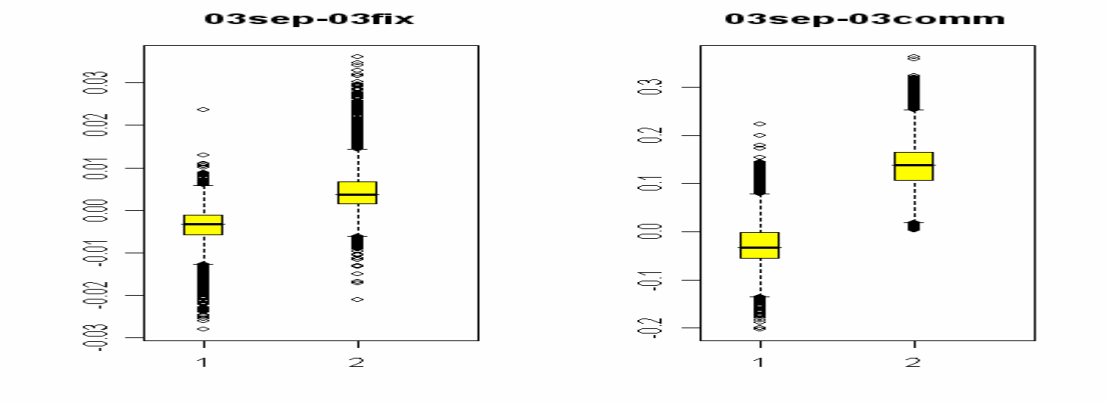

Figure 6: **Marginal expectation comparison for 2003 with 25 common items.**

Figures 4 to 6 present results that are similar to those in Figures 1 to 3. This would seem to indicate that the scale established by Strategy 1 is robust even in the case where fewer common items exist.

Comparison II: Three strategies. The comparisons in this section are meant to investigate whether we can establish the link between an assessment from two different years by dropping concurrent calibration and only using fixed item parameters for subsequent calibrations. In our case, this would mean that the common item parameters obtained from *Y1 calibration* could be applied directly to the analysis of 2005 data. In this section, the comparison was conducted by using the following: item parameter estimates, model fit, and marginal skill expectations. In addition, this comparison was investigated when either 69 common items or 25 common items existed between the two tests.

The difference in joint probabilities of skill patterns between 2005 students and 2003 students under three different strategies are shown in Figure 7 in the form of boxplots. The numbers shown along the *x*-axis in Figure 7 stand for strategy ID (i.e. Strategy 1, 2, or 3). If a scale identical to that from the 2003 separate calibration could be set up by Strategy 3, then the boxplots for Strategy 1 and Strategy 3 should be similar to each other. Otherwise, boxplots for Strategy 1 and Strategy 2 should be similar to each other. It turns out that the former is found in this study. A similar result is obtained when there are 25 common items, as shown in Figure 8.

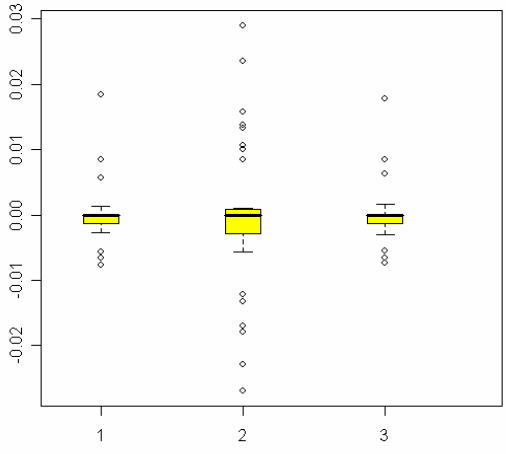

Figure 7. **Joint probability comparison: 2005 minus 2003.**

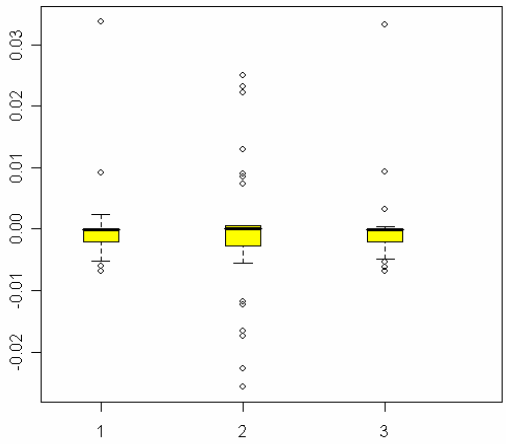

Figure 8. **Joint probability comparison: 2005 minus 2003 with 25 common items.**

The marginal skill expectations obtained from Strategies 1 and 2 are compared with those from Strategy 3 via boxplots shown in Figures 9 and 10. In each graph, numbers 1 and 2 along the *x*-axis stand for the two subscales measured in reading. The left graph presents the difference between Strategies 3 and 1, while the right panel illustrates the difference between Strategies 2 and 1. If an identical scale is established through fixing item parameter values, and not by concurrent calibration, then the difference between Strategies 1 and 3 should be smaller than the difference between Strategies 1 and 2. The results confirm this, since the boxplots for both reading subscales shown in the left graph are much more concentrated around 0 than those in the right graph.

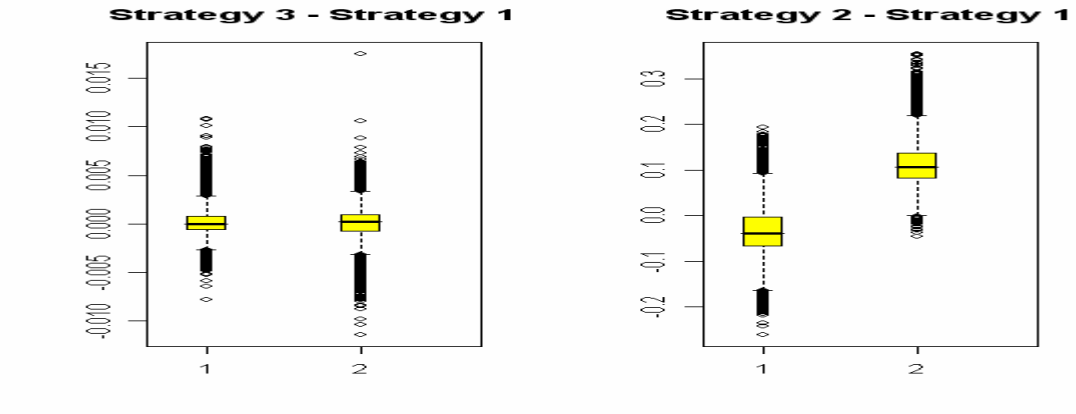

Figure 9. **Marginal expectation comparison for 2005 data.**

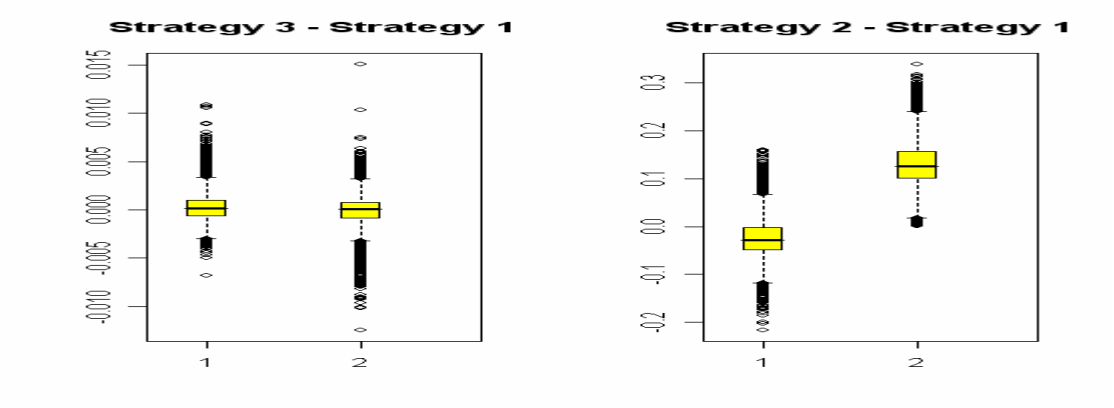

Figure 10. **Marginal expectation comparison for 2005 data with 25 common items.**

The comparisons in this section indicate that a scale identical to the 2003 separate calibration can be reproduced in the 2005 separate calibration by fixing the common item parameters at the estimates obtained from the 2003 separate calibration.

6. Key Subgroups Comparison

Statistics for key subgroups, such as the mean, standard deviation, and quantiles of subgroups, are important to consider for operational NAEP reporting purposes. In the framework of cognitive diagnosis, skill distributions for key subgroups at an equivalent aggregation level could be considered for NAEP reporting. In this paper, skill distributions of several key subgroups (race/ethnicity and gender) in the 2005 assessment are compared across linking strategies. Also, the skill profiles of key subgroups in the 2003 assessment are compared across strategies and with those obtained from the separate calibration of 2003. Since the results of the case with 25 common items and the case with 69 common items were again similar to each other, we only report the results of the case with 25 common items.

In the following comparisons, the skill profiles for subgroups were derived based on a single-group assumption. That is, all subgroups were set to have the same prior distribution for the latent classes. Then the skill profile for a subgroup was calculated by taking a weighted average of the skill profiles of students in the subgroup, where the weights were the student weights used in the NAEP operational analysis.

The skill profiles for subgroups from the 2003 assessment were compared between *Y1 calibration* and Strategies 1 and 2. The differences in estimated marginal skill distributions are shown in Figures 11 and 12. In each graph, the subgroups are represented by a capitalized initial letter.

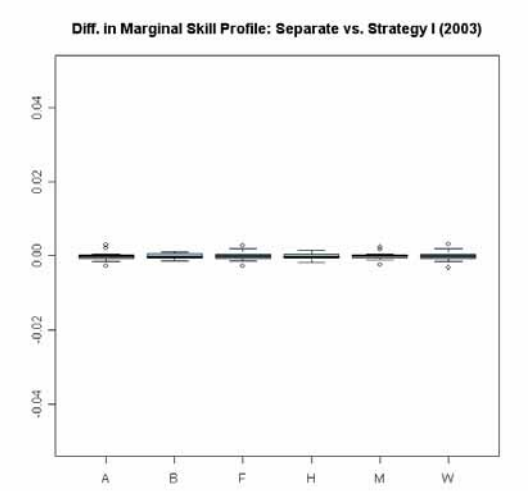

Figure 11. **Differences in marginal skill profile: separate versus Strategy 1 (2003).**

Note. A, B, F, H, M, and W stand for Asian, Black, female, Hispanic, male, and White student groups, respectively.

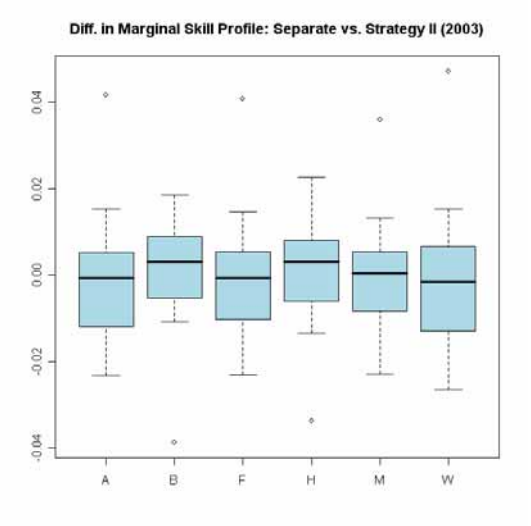

Figure 12. **Differences in marginal skill profile: separate vs. Strategy 2 (2003).**

Note. A, B, F, H, M, and W stand for Asian, Black, female, Hispanic, male, and White student groups, respectively.

Although the differences between *Y1 calibration* and Strategy 2 (shown in Figure 12) might be considered to be small (within range of -0.04 and 0.04), the difference between *Y1 calibration* and Strategy 1 (shown in Figure 11) are even smaller. It can be shown that Strategy 1 leads to an almost identical scale with the scale from *Y1 calibration*.

The skill profiles for subgroups from the 2005 assessment were compared between Strategy 3 and Strategies 1 and 2. These differences are shown in Figures 13 and 14.

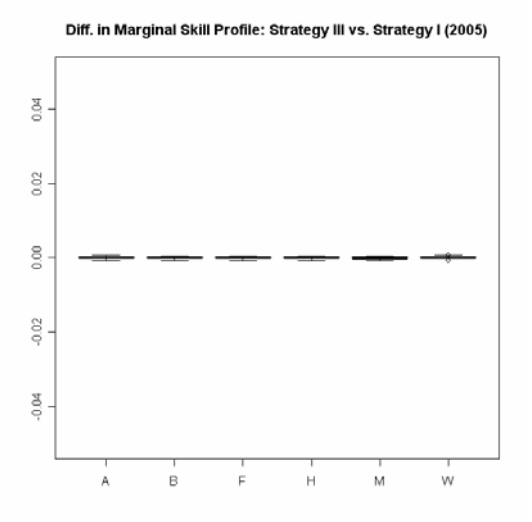

Figure 13. **Differences in marginal skill profile: Strategy 3 versus Strategy 1 (2005).**

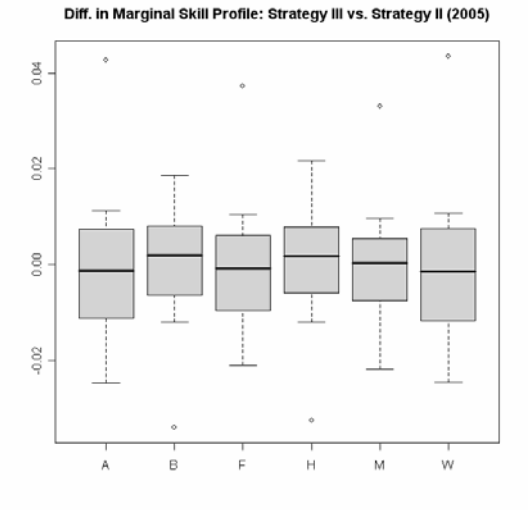

Figure 14. **Differences in marginal skill profile: Strategy 3 versus Strategy 2 (2005).**

The results presented here demonstrate that Strategy 1 is much closer to Strategy 3 than Strategy 2 in terms of the estimated marginal skill profile for key subgroups. This may imply that, in the case of linking these two NAEP assessments under the GDM framework, a scale can be set and reproduced by fixing the values of the common items in these two assessments, even when there are only 25 (approximately 25% of the entire test) items in common.

7. Discussion and Conclusion

The noncontinuous nature of the skill locations in the GDM limits the search for appropriate linking methods. The research question that needs to be answered is what linking methods lead to a scale that is maintained across assessments *Y1* and *Y2*. 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. Thus, three different strategies were compared in this study. As mentioned in Section 3, these three strategies are variations from the concurrent calibration linking used in NAEP operation. Often, a concurrent calibration linking consists of three steps of calibration and transformation. Strategy 2 in this study is indeed the first step of the concurrent calibration linking, since the other steps are dropped intentionally due to their inappropriateness for discrete latent skills/abilities. Strategy 1 produces a stronger link than Strategy 2 by fixing the common items parameters at known values from *Y1 calibration* in addition to the concurrent calibration. Strategy 3 is a simplified version of Strategy 1 by dropping the concurrent calibration and keeping the common item parameters fixed at known values from *Y1 calibration*. In fact, Strategy 3 will be identical to Strategy 1 if there is no constraint imposed on the item parameter estimation procedure. Even when certain constraints are put in the estimation process, only slight differences can be observed for Strategy 1 and Strategy 3, as shown in Figures 7 to 10 and Figure 13.

All results in this study empirically demonstrate that one strategy of linking, the concurrent calibration of two adjacent assessments, is not necessary when the common items are fixed at the values obtained from *Y1 calibration*. The similar results even hold up in the case where common items consist of only 25% of the whole NAEP assessment. This result, however, should not be generalized, since it may not hold up in studies with different procedures for assessment development, block formation, item flagging, and selection for subsequent assessment.

In order to make sure the conclusion is true for the case where only 25 items are in common to both tests, two additional analyses were carried out based on different sets of 25 items. Similar results were obtained to those shown in this paper. The authors also analyzed data from a Grade 8 NAEP reading assessment in 2003 and 2005. Similar conclusions were drawn from these analyses.

Even though the purpose of this paper was not focused on model comparisons, we have to mention one special model case where only two levels (mastery and nonmastery) are specified for each cognitive skill. The authors ran such cases and found out, in this case, the concurrent calibration of 2003 and 2005 assessments as in Strategy 2 is able to reproduce the scale established by *Y1 calibration*.

As discussed in a previous section, the analysis is conducted based on a single-group assumption, assuming only one skill distribution. A future study will focus on an analysis based on a multiple-group assumption coupled with Strategy 3. Under the multiple-group assumption, subgroups are assigned unique and potentially different prior distributions, so that the skill profiles for subgroups are 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 yields similar results for racial subgroups and gender subgroups as those from NAEP operational analyses. A future study will be able to answer further questions, such as whether a comparable scale can be established by Strategy 3 when employing a GDM multiple-group analysis procedure.

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 Rep. No. RR-05-24). Princeton, NJ: ETS.

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.

Junker, B., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24,* 65–81.

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

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187–212.

Mislevy, R. J. (1992). Scaling procedures. In E. G. Johnson & N. L. Allen (Eds.), *The NAEP 1990 technical report* (Rep. 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.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345–54.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: ETS.

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics.* Amsterdam: Elsevier.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (ETS Research Rep. No. RR-06-08). Princeton, NJ: ETS.