Longitudinal Effects of Item Parameter Drift

James A. Wollack
Hyun Jung Sung
Taehoon Kang

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

April 12, 2005

Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada

RUNNING HEAD: Longitudinal Effects of IPD

Longitudinal Effects of Item Parameter Drift

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

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

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

Yet, the fact that having a subset of items with different item parameter estimates for two timepoints still results in reasonably similar ability estimates remains counterintuitive.

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

Therefore, for test developers to be comfortable employing methods that largely ignore any potential IPD, it must first be demonstrated that IRT ability estimation is robust to IPD over a multiple-year period. One concern is that the item drift problem may cumulate over time, as the number of drifting items and the magnitude of drift increases, particularly if drifting items are included in the linking of test forms (Kim & Cohen, 1992; Lautenschlager & Park, 1988; Shepard, Camilli, & Williams, 1984). In this study, we used data from a college-level German placement test to examine the effect of compounding IPD on a score scale over a 7-year period. The impact of drift on examinee ability was studied under ten different IRT linking designs.

Methods

Data Source

This study analyzed seven years' worth of data from a German placement test (Forms 90X through 96X) used at a large, midwestern university. Students taking the German Placement Test during those years were administered approximately 55 items each year, though the number of operational (i.e., scored) items ranged from 30 to 53 (with an average of 42.4). The remaining items were pilot tested for purposes of studying their properties and estimating item parameters. The particular seven year period studied here is well suited for research on IPD because the items on the test remained virtually unchanged from year to year. In fact, on any given form (excluding the initial form), every operational item was administered—either operationally or as a pilot—in the previous year. Over half of the items on Form 96X (i.e., 17 of 32 items) were administered in all seven years. The test was administered annually to 750 – 1,500 students.

Models for Linking and Controlling IPD

Data from Form 90X was used to calibrate the item parameters under the three parameter logistic model, using the computer program MULTILOG 7.0 for Windows (Thissen, 2003). For simplicity, lower asymptote parameters were fixed at .2 for all items. MULTILOG default settings were used, except the maximum number of cycles of the EM algorithm was increased from 25 (the program default value) to 1000, to increase the likelihood of convergence.

Ten different methods were used to link the remaining forms to the 90X metric. The different methods varied with respect to both the method used for linking and the way in which drifting items were treated. A summary of these models is provided in Table 1. The models are described in detail below.

Insert Table 1 About Here

Models 1 and 2: Common item linking with fixed item parameters. In Models 1 and 2, item parameters were estimated for Forms 91X through 96X by fixing operational items at their scale values, based on a previous administration. Any items that had not been previously administered, i.e., pilot items, were freely estimated. Because Models 1 and 2 did not involve testing for IPD, once an item's parameters were estimated, the item retained those values throughout all future forms of the test.

Models 1 and 2 differed with regard to whether the link to the 90X metric was done directly or indirectly. In the case of item linking over multiple years, new forms may be linked directly to the original metric through the items common between the original and new form. However, it is also possible to indirectly link to the original metric by linking to the immediately preceding form (which was linked to its predecessor, and so on back to the original metric). From a theoretical perspective, each model offers its advantages. Direct linking has the advantage of minimizing linking errors, which could compound over multiple years. However, because forms administered over two years are likely to contain more items in common than forms administered several years apart, indirect linking has the advantage of producing a larger set of linking items. The problem of linking forms over multiple years is illustrated in Table 2 for five hypothetical forms, each containing 10 operational items (numbered 1-10) and four pilot items (numbered 11-14). Given the design in Table 2, indirect linking will always involve 10 items because all operational items were selected from the previous form. With direct linking, however, the number of items available (shown in boldface type in Table 2) decreases with each

new form. For example, a direct linking between Forms 1 and 5 would involve only three items. In this study, Model 1 used a direct link to 90X, whereas Model 2 used an indirect link to 90X.

Insert Table 2 About Here

Models 3 and 4: Common item equating using the test characteristic curve (TCC) method (Stocking & Lord, 1983). In Models 3 and 4, item parameters were estimated for Forms 91X through 96X using MULTLOG, and linked to the 90X metric using the TCC method, as implemented in the EQUATE computer program (Baker, 1990). In Model 3, Forms 91X through 96X were linked directly to 90X. In Model 4, forms were indirectly linked to 90X.

Models 5 and 6: Common item equating with fixed item parameters and IPD testing. Models 5 and 6 were identical to Models 1 and 2, except that prior to linking, items were tested individually for IPD using the likelihood ratio (LR) test for DIF (Thissen, Steinberg, and Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993). The LR tests were conducted by comparing the values of −2 times the log likelihood (−2log(L)) from a compact model, in which parameters for all items common to the two forms were constrained to be equal, and an augmented model, in which item difficulty and discrimination parameters for one common item were estimated separately for the two groups. A separate augmented model was calibrated for each item common to the two forms. Drift among the lower asymptote parameters was not considered; c parameters were fixed at .2 for all items in both the compact and augmented models. Items with LR test statistics (i.e., $-2\log(L)_{compact} + 2\log(L)_{augmented,(i)}$) greater than $\chi^2_2(.95) = 5.99$ were identified as drifting and were not included among the linking set. Instead, drifting items were, for purposes of the calibration/linking process, treated as pilot items and had their parameters re-estimated.

Item parameter estimates for all non-drifting items were fixed at their scale values. Again, the linking to the 90X metric may be done either directly and indirectly. Some advantages of both direct and indirect linking were discussed earlier. However, an additional theoretical advantage of indirect linking applies within the context of testing for IPD prior to linking. If a change in the environment (e.g., curricular changes) make an item easier, it is likely that the original scale value will, from that point forward, be inappropriate. In many cases, though, a drifting item will eventually stabilize and begin to behave similarly from year to year. A model using indirect linking, by virtue of using the most recent estimate of an item's difficulty, should lead to a smaller magnitude of scale drift and fewer items exhibiting IPD (hence a larger anchor set) than a model based on direct linking.

In Model 5, items were directly assessed for IPD from the 90X scale by testing items for IPD between the current form and Form 90X. Non-drifting items had their parameter values fixed at the 90X scale values. Indirect drift was assessed in Model 6 by testing for IPD between the current year and the preceding year. Parameter values for non-drifting were fixed to equal their values from the preceding year.

Models 7 and 8: Common item TCC linking with IPD testing. These conditions were exactly like Models 5 and 6, except that instead of linking through fixed item parameters, the TCC method was used. Model 7 linked directly to Form 90X, and Model 8 linked indirectly to 90X.

Model 9: Concurrent calibration with equality constraints on common items. Using MULTILOG, all seven test forms were calibrated simultaneously. The calibration was performed as a single group design. Equality constraints on common items were imposed by having all items common across years appearing in the same column. Between the operational

and pilot items for all seven forms, a total of 121 different items and 7,093 examinees were included in this run.

Direct comparison of parameter estimates from a concurrent calibration with those from separate calibrations are not possible without equating because the underlying metrics may be different. Therefore, following the concurrent calibration, Model 9 item parameter estimates were linked back to the original 90X metric (i.e., based on the initial calibration of 90X) using the TCC method. All 55 items from Form 90X were used in this linking.

Model 10: Concurrent equating with equality constraints and IPD testing. This model was similar to Model 9 in that the seven forms were calibrated simultaneously. To allow for comparisons with other models, Model 10 item parameter estimates were also linked back to the original 90X metric using the procedure described above. Model 10 differed from Model 9 in that, instead of constraining parameters to be equal for all common items across forms, the results of Model 8 (Common item TCC linking with IPD testing) were used to determine where equality constraints were appropriate. Common items across different forms appeared in the same column provided the item did not exhibit IPD. Items that did show significant drift appeared in separate columns and were estimated separately. It was possible for common items to be held constant across certain years and estimated separately for others. As an example, one item showed no IPD across the first 4 forms (i.e., 90X – 93X), but drifted between 93X and 94X. On Form 95X, that same item did not drift (from 94X), but did drift again on Form 96X. For purposes of Model 10, there were three separate estimates of that item's parameters: one for Forms 90X – 93X, one for Forms 94X and 95X, and one for Form 96X.

Outcome measures

To investigate the impact of IPD and the linking model on ability estimation, θ estimates from Form 96X under the ten different models were compared via several methods. Descriptive statistics included scale means and standard deviations for each of the ten models, and correlations, mean differences, and root mean square differences (RMSD) between θ estimates for all pairs of models. In addition, the impact of the particular linking/IPD model was studied by examining true score functions and passing rate functions. To better understand the amount and magnitude of naturally occurring drift over time, summary tables are provided describing the nature of the compounding drift in this study.

Results

Form 96X had 53 items and was completed by 1,188 examinees. Of the 53 total items, 32 items appeared on a previous form (in fact, all 32 were common to the previous Form 95X) and 17 were common to all seven forms, though item locations did change slightly. A total of 21 items were new to Form 96X and did not appear on previous forms.

Sample means and standard deviations for Form 96X under the ten models are given in Table 3. For all but Models 4 and 7, the means were very close to 0.00 and the standard deviations were between 0.84 and 0.90. Under both Models 4 and 7, examinees' θ estimates were higher and less variable than under the remaining eight models. This was particularly true under Model 4, where the mean and standard deviation of $\hat{\theta}$ were 0.46 and 0.76, respectively.

Insert Table 3 About Here

Inter-model correlations were 1.0 between all pairs of models. That all models are a linear transformation away from being identical suggests, as would be expected, that all differences between the models are attributable to differences in the linking procedures.

Average pairwise differences and RMSDs are presented in the upper and lower triangles of Table 4, respectively. Negative average differences mean that the higher numbered model had a higher θ estimate than the lower numbered model. As an example, the average difference between Models 1 and 5 was -0.02. Therefore, the average θ estimate for Models 5 was 0.02 logits higher than that for Model 5.

Insert Table 4 About Here

The data from Table 4 are largely consistent with the pattern from Table 3: θ estimates from Models 4 and 7 are different than those from the other eight models. Average differences and RMSDs among the other eight models were all quite small, and suggest that there was little difference between them in this study.

An important consideration in determining the impact of the different linking/IPD models is to evaluate differences in cutscores or percentages of examinees scoring above particular scores, based on the different models. Figure 1 shows the relationship between true θ level and observed true scores under the ten models. For a fixed θ level, Figure 1 shows the range in true scores under the different models. Under most circumstances, examinees would receive dramatically lower true score estimates under Models 4 and 7, than under the other eight models. As an example, an examinee with a true θ level of 0.5 has a true score estimate of 30.16 under Model 4 and 32.61 under Model 7. Under the remaining eight models, the true score estimates range from 33.42 to 34.20. The largest difference between the minimum and maximum true

score estimates occurred at $\theta = 0.1$, where the Model 4 true score estimate was 24.88 and the Model 8 estimate was 29.29, for a difference of 4.42 raw score points. Over the range $-2.5 \leq \theta \leq 2.5$, the average difference between true score estimates for Model 4 and the average of the eight more similar models was 2.13 raw score points. By any standard, this is a very large difference.

Insert Figure 1 About Here

Differences between Model 7 and the remaining eight models were noteworthy, but considerably smaller. The average difference between true score estimates for Model 7 and the average of the eight more similar models was 0.74 raw score points. Whereas Model 4 consistently led to the lowest true score estimate, Model 7 produced the second lowest true score estimate only for $\theta < 1.41$. Above $\theta = 1.41$, there was very little difference between Model 7 and the eight more similar models.

Although Models 1, 2, 3, 5, 6, 8, 9, and 10 appeared from Figure 1 to be generally quite similar, there were still some differences between them that could have important implications for examinees. When $\theta < 0$, the range of these eight models (i.e., the maximum true score minus the minimum true score) was never more than 0.55 points and the standard deviation among the models was always below 0.22. However, for $\theta > 0$, the eight models were less similar. For θ values between 0.4 and 1.5, the range of true score estimates was above 0.75 and the standard deviation was at least 0.28. The maximum range and standard deviation, occurring at $\theta = 1$, were 0.85 and 0.30, respectively. Depending on the purposes of a test, differences of this magnitude between different linking models may be nontrivial.

Average differences (upper diagonal) and RMSDs (lower diagonal) on estimated true scores are shown in Table 5 for the ten models. Negative average differences mean that the

higher numbered model had a higher true score estimate than the lower numbered model. From Table 5, one can again see that Model 4 was the most dissimilar, followed by Model 7. After that, however, it becomes more difficult to discern a clear pattern. Models 1, 2, 5, and 6—all of which involved fixing item parameters—appeared quite similar to each other. Model 3 produced the smallest RMSDs with Models 1, 2, and 9, all of which did not incorporate IPD testing, but the smallest average difference was with Model 10. Model 9 produced low average differences when compared to Models 3, 8, and 10, and low RMSDs when compared to Models 3 and 8. Model 8 appeared most similar to Model 9, and Model 10 was most similar to Model 3.

Insert Table 5 About Here

As indicated earlier, the differences between models became even greater for higher ability examinees. Table 6 shows the average differences and RMSDs in estimated true scores for examinees with $\theta \geq 0$. Here, the average differences and RMSDs suggest that Models 4, 7, 8, and 9 were all producing results different from those of other models. In many cases, the statistics for Models 8 and 9 were larger than for Model 7. For the most part, these differences for higher ability examinees are still small-to-moderate.

Insert Table 6 About Here

A second way to assess the impact of linking/IPD models is to examine the passing rates for each model for different passing scores on the $\hat{\theta}$ metric. These functions, shown in Figure 2, identify the percentage of examinees scoring at or above a particular $\hat{\theta}$. That is, they indicate the percentage of examinees who would pass the test, were the passing score set at $\hat{\theta}$. The data

in Figure 2 resemble patterns shown in other figures and tables. Passing rates for Models 1, 2, 3, 5, 6, 8, 9, and 10 were similar for all values of $\hat{\theta}$. Similarly, for a given passing rate, there was very little variance among the $\hat{\theta}$s for these eight models. Except for the most extreme passing rates (i.e., passing or failing only 5% of the examinees), the range in $\hat{\theta}$s among these eight models was always less than 0.10. Models 4 and 7 are visibly shifted to the right of the other eight models. Far fewer people score at or above a particular $\hat{\theta}$ under Models 4 and 7, and the $\hat{\theta}$ corresponding to a fixed passing rate is much higher.

Insert Figure 2 About Here

Discussion

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

In this study, ten different models for linking and accounting for IPD were considered and applied to a German Placement Test dataset over a seven year period. The results of this study showed that choice of linking/IPD model can have a large effect on the resulting $\hat{\theta}$ and estimated true score distributions, as well as on passing rates. Models 4 and 7 consistently produced different patterns of results than the remaining models. Under most conditions, Models 1, 2, 3, 5, 6, 8, 9, and 10 produced similar results. However, non-negligible differences in expected true scores between the models existed for certain θ levels, most notably for $\theta \geq 0$.

From a theoretical perspective, the differences between the ten models were considerable. Therefore, it is a bit surprising that more or larger differences were not found. Clearly, one possible reason for this is that, even in the context of an ongoing testing program where different items may be drifting at different times and IPD may compound over time, IRT is sufficiently robust to estimate examinee ability well. However, the theory underlying IRT suggests that failing to account for IPD may well result in linking error which, in turn, will have an effect on ability estimation. Furthermore, it is reasonable to assume that the amount of linking error is influenced by the number and magnitude of drifting items, as well as the number of links containing error.

One alternative explanation for why more differences were not found is that the nature of the naturally occurring IPD in this study was not conducive to finding large differences in ability estimation. To explore this, we examined the properties of the 17 items that were common to all seven test forms. Table 7 shows the item discrimination and item difficulty parameter estimates from Model 10. Items that did not drift across years are shown in a single box. As an example, item 1 exhibited IPD between Forms 90X and 91X, and again between Forms 95X and 96X, but not between Forms 91X and 95X. Therefore, within the concurrent calibration for Model 10, this item had its parameters estimated three times: once using only the examinees completing Form 90X, once using all the examinees completing Forms 91X, 92X, 93X, 94X, and 95X, and once using only the examinees completing Form 96X. In addition, Table 7 shows the average amount of drift and the average absolute drift (AAD) in each parameter.

Insert Table 7 About Here

There are several things of note in Table 7. First, the total number of drifting items was rather small. Among these 17 items, an average of three drifted between any two years. Between Forms 91X and 92X, only one of these items drifted, and none drifted between Forms 94X and 95X. Overall, seven of these items did not drift at all, and only one item (item 12) drifted more than twice.

This same pattern is observed when looking at all the items on the test forms (not just these 17). Across the six time points at which IPD was evaluated, the number of drifting items ranged from 4 to 8, with an average number of just 5.8. Given that adjacent forms shared an average of 41 items in common, 5.8 drifting items constituted just 14% of the items.

In addition to having relatively few drifting items, the magnitude of drifting items in this dataset was rather small. Studied magnitudes of a and b drift have generally been at least 0.4 for difficulty and 0.3 for discrimination (e.g., see Donoghue & Isham, 1998; Wells et al, 2002). Yet, across the 17 common items, the AAD was just 0.21 for the a parameters and 0.24 for the b parameters. This may be because the test in this study was designed to measure basic language mechanics and reading comprehension, fundamental skills which were an essential part of all German curricula. Also, the items were administered in very nearly the same order each year, thereby eliminating location effects—a major source of IPD (Oshima, 1994).

Finally, the pattern of IPD in this study appeared to be that of random drift rather than systematic drift. From Table 7, the average amount of drift between any two drifting items (among the 17 common items) was just 0.01. Consequently, it is likely that positively and negatively drifting items largely canceled each other out. Furthermore, only one of the individual items showed a consistent pattern of drift. Of the six items that drifted more than once (items 1, 3, 8, 11, 12, and 13), only item 11 got progressively easier and less discriminating over

time. The remaining items either showed an alternating pattern of becoming easier and harder, or more discriminating and less discriminating, or both. Therefore, in this study, the effects of IPD did not appear to cumulate as anticipated.

One thing that is impossible to conclude from this study is which of the ten studied models is robust to IPD. As the next phase in this study, we are planning a simulation study in which we will examine the performance of these ten models under different types, amounts, and magnitudes of IPD. We are hopeful that a thorough simulation study will help to quantify the extent of the cumulating drift problem and identify linking strategies to help account for IPD so as to best maintain the integrity of the score scale.

References

Baker, F. B. (1990). EQUATE: Computer program for equating two metrics in item response theory [Computer program]. Madison, WI: University of Wisconsin Department of Educational Psychology, Laboratory of Experimental Design.

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

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

DeMars, C. E. (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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 1

Description of Linking and Drift Models Studied

	Linking Technique			Linking Method		Drift Testing	
	Common Item						YES
	Fixed Item Parameters	TCC Method	Concurrent Calibration	Directly with 90X	Indirectly with 90X	NO	LR Test
Model 1	X			X		X	
Model 2	X				X	X	
Model 3		X		X		X	
Model 4		X			X	X	
Model 5	X			X			X
Model 6	X				X		X
Model 7		X		X			X
Model 8		X			X		X
Model 9			X	X		X	
Model 10			X	X			X

Table 2

Illustration of Direct Versus Indirect Linking

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

Table 3

Means and Standard Deviations for Linking and IPD Models

	Linking and IPD Models									
	1	2	3	4	5	6	7	8	9	10
Mean	0.01	0.02	0.00	0.46	0.03	0.03	0.13	-0.02	-0.01	-0.02
S.D.	0.88	0.87	0.87	0.76	0.87	0.88	0.81	0.84	0.86	0.90

Table 4

Average Differences and RMSDs on θ Between Linking and IPD Models

	Linking and IPD Models									
	1	2	3	4	5	6	7	8	9	10
Model 1		-0.01	0.01	-0.45	-0.02	-0.01	-0.11	0.03	0.02	0.03
Model 2	0.02		0.02	-0.44	-0.01	0.00	-0.10	0.04	0.03	0.04
Model 3	0.04	0.04		-0.46	-0.03	-0.02	-0.12	0.02	0.01	0.02
Model 4	0.47	0.46	0.47		0.43	0.01	-0.09	0.05	0.04	0.05
Model 5	0.03	0.03	0.03	0.45		0.01	-0.09	0.05	0.04	0.05
Model 6	0.03	0.03	0.03	0.45	0.02		-0.10	0.04	0.03	0.04
Model 7	0.14	0.13	0.14	0.34	0.12	0.12		0.14	0.13	0.14
Model 8	0.06	0.06	0.04	0.49	0.06	0.06	0.15		-0.01	0.00
Model 9	0.04	0.04	0.02	0.48	0.04	0.04	0.14	0.03		0.01
Model 10	0.04	0.05	0.04	0.50	0.06	0.05	0.17	0.06	0.04	

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

Table 5

Average Differences and RMSDs on True Scores Between Linking and IPD Models

	Linking and IPD Models									
	1	2	3	4	5	6	7	8	9	10
Model 1		0.03	-0.08	2.07	0.11	0.05	0.68	-0.28	-0.18	-0.15
Model 2	0.11		-0.11	2.04	0.08	0.03	0.65	-0.31	-0.21	-0.18
Model 3	0.14	0.15		2.15	0.19	0.14	0.76	-0.19	-0.09	-0.06
Model 4	2.45	2.39	2.54		-1.96	-2.01	-1.39	-2.34	-2.24	-2.21
Model 5	0.11	0.11	0.22	2.35		-0.05	0.57	-0.38	-0.28	-0.25
Model 6	0.06	0.11	0.18	2.40	0.06		0.62	-0.33	-0.23	-0.20
Model 7	0.86	0.79	0.93	1.63	0.76	0.82		-0.95	-0.85	-0.82
Model 8	0.47	0.42	0.33	2.70	0.52	0.50	1.08		0.10	0.13
Model 9	0.30	0.27	0.16	2.62	0.36	0.33	1.00	0.18		0.03
Model 10	0.22	0.31	0.21	2.66	0.32	0.26	1.08	0.47	0.31	

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

Table 6

Average Differences and RMSDs on True Scores
Between Linking and IPD Models for $\theta \geq 0$

	Linking and IPD Models									
	1	2	3	4	5	6	7	8	9	10
Model 1		-0.06	-0.17	1.69	0.08	0.05	0.36	-0.60	-0.37	-0.07
Model 2	0.08		-0.11	1.75	0.14	0.11	0.42	-0.54	-0.31	0.00
Model 3	0.18	0.15		1.86	0.24	0.22	0.53	-0.44	-0.20	0.10
Model 4	2.25	2.26	2.42		-1.61	-1.64	-1.33	-2.29	-2.06	-1.75
Model 5	0.08	0.14	0.26	2.18		-0.03	0.28	-0.68	-0.45	-0.14
Model 6	0.05	0.13	0.23	2.22	0.04		0.31	-0.65	-0.42	-0.11
Model 7	0.63	0.62	0.78	1.65	0.58	0.61		-0.96	-0.73	-0.42
Model 8	0.62	0.56	0.45	2.76	0.69	0.67	1.11		0.23	0.54
Model 9	0.39	0.34	0.21	2.59	0.46	0.44	0.94	0.24		0.31
Model 10	0.18	0.23	0.16	2.41	0.23	0.20	0.81	0.57	0.33	

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

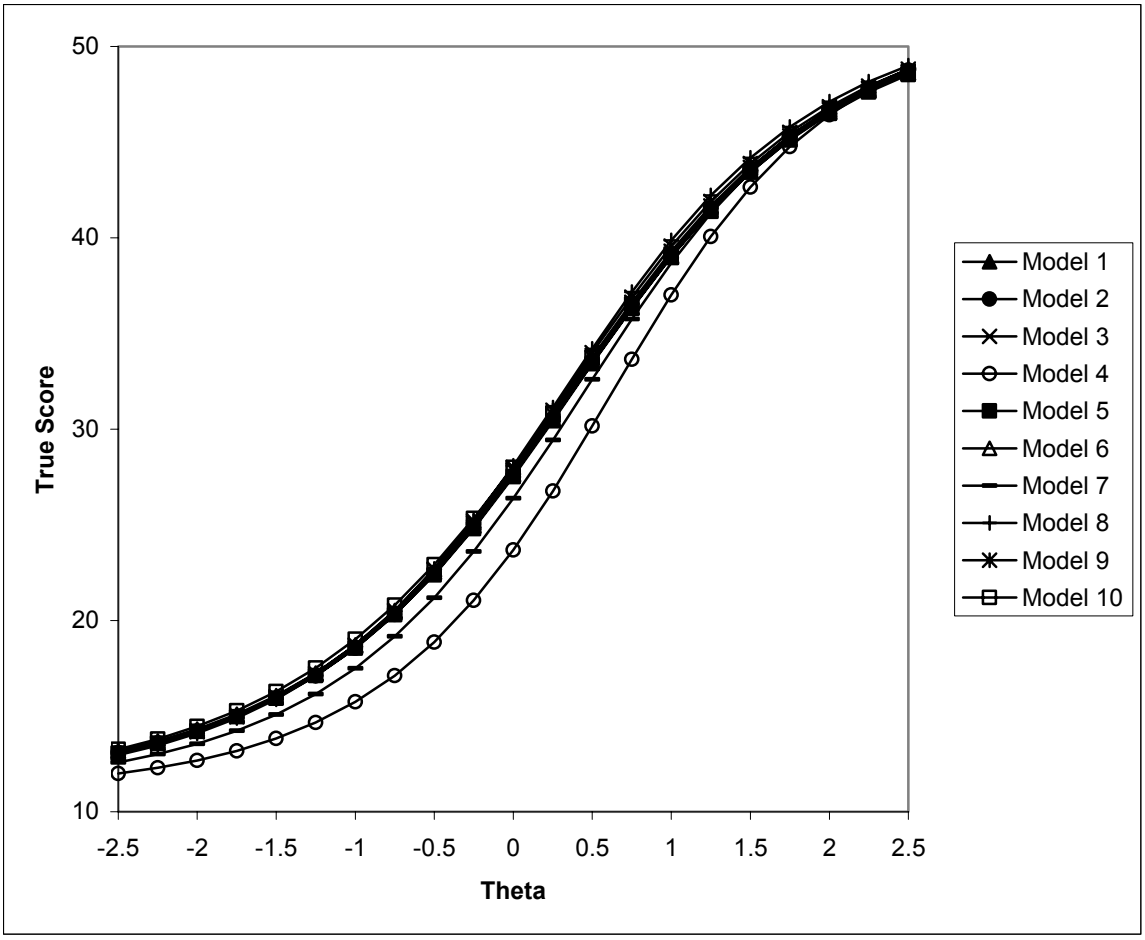
Table 7

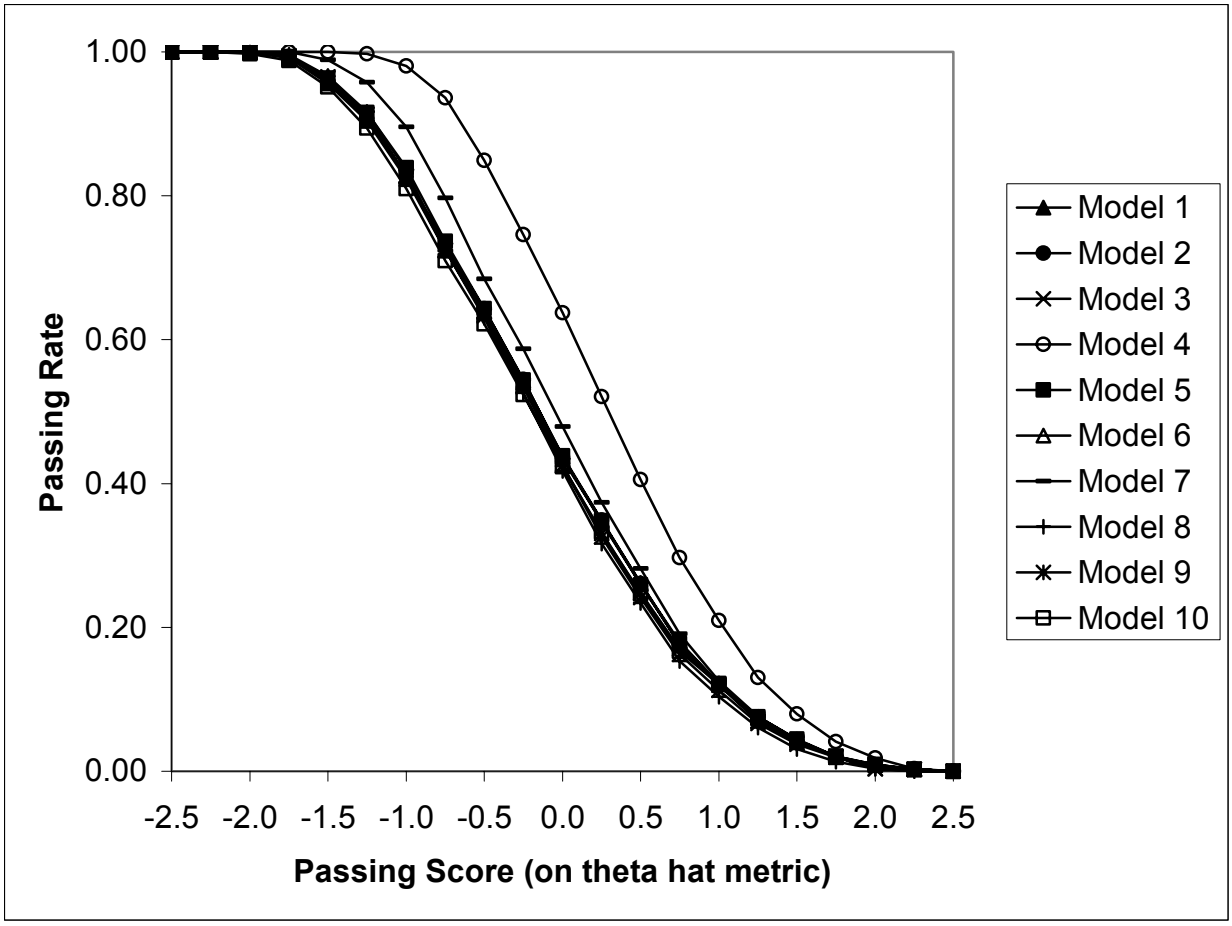
Drift Pattern for Common Items Across 7 Years

	Test Form							Mean Drift		AAD	
Item	**90X**	**91X**	**92X**	**93X**	**94X**	**95X**	**96X**	**a**	**b**	**a**	**b**
1	1.03 0.28			1.01 0.55			0.83 0.26	**0.10**	**0.01**	**0.10**	**0.28**
2		1.10 0.04			0.89 0.13			**0.21**	**-0.09**	**0.21**	**0.09**
3			0.72 0.31			0.72 0.30	0.74 0.04	**-0.01**	**0.14**	**0.01**	**0.14**
4			1.10 0.37			1.04 0.35		**0.06**	**-0.25**	**0.06**	**0.02**
5				0.62 0.80			0.85 1.05	**-0.23**	**-0.25**	**0.23**	**0.25**
6				0.71 0.40				**N/A**	**N/A**	**N/A**	**N/A**
7				1.49 0.88				**N/A**	**N/A**	**N/A**	**N/A**
8	1.86 0.58			1.32 0.41			1.49 0.68	**0.19**	**-0.05**	**0.36**	**0.22**
9		0.74 -0.68			0.78 -0.39			**-0.04**	**-0.29**	**0.04**	**0.29**
10				0.62 0.20				**N/A**	**N/A**	**N/A**	**N/A**
11	1.27 0.75			0.78 0.66			0.73 0.34	**0.27**	**0.21**	**0.27**	**0.21**
12	1.22 0.86	0.71 1.32	0.82 0.70	0.76 1.13	1.00 0.83			**0.06**	**0.01**	**0.23**	**0.45**
13		1.28 0.56			0.76 0.55		1.12 0.46	**0.08**	**0.05**	**0.44**	**0.05**
14				1.16 0.40				**N/A**	**N/A**	**N/A**	**N/A**
15				1.91 0.98				**N/A**	**N/A**	**N/A**	**N/A**
16				2.24 0.24				**N/A**	**N/A**	**N/A**	**N/A**
17				1.45 0.40				**N/A**	**N/A**	**N/A**	**N/A**
							Averages	**0.10**	**0.01**	**0.21**	**0.24**

Figure 1. True Score Functions

Figure 2. Passing Rate Functions


50
40
30
20
10
True Score
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
Theta
Model 1
Model 2
Model 3
Model 4
Model 5
Model 6
Model 7
Model 8
Model 9
Model 10


1.00
0.80
0.60
0.40
0.20
0.00
Passing Rate
-2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5
Passing Score (on theta hat metric)
Model 1
Model 2
Model 3
Model 4
Model 5
Model 6
Model 7
Model 8
Model 9
Model 10