DOCUMENT RESUME

ED 478 167 TM 035 061

AUTHOR	Karkee, Thakur; Lewis, Daniel M.; Hoskens, Machteld; Yao, Lihua; Haug, Carolyn
TITLE	Separate versus Concurrent Calibration Methods in Vertical Scaling.
PUB DATE	2003-04-00
NOTE	47p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Estimation (Mathematics); *Mathematics Tests; *Scaling; *Scoring; State Programs; *Testing Programs
IDENTIFIERS	*Calibration; Colorado Student Assessment Program; *Vertical Equating

ABSTRACT

Two methods to establish a common scale across grades within a content area using a common item design (separate and concurrent) have previously been studied under simulated conditions. Separate estimation is accomplished through separate calibration and grade-by-grade chained linking. Concurrent calibration established the vertical scale in a single step by simultaneously estimating parameters for all items at all grades. These methods, and a third hybrid method, pairwise concurrent estimation, were examined in this study, using operational test data. Data were obtained from responses to the 2002 Colorado Student Assessment Program (CSAP) mathematics assessments in grades 5 through 10. These assessments had been constructed to have at least 20 items in common between adjacent grades. The data set for each grade consisted of 10,000 cases, randomly selected from the population of approximately 54,000 students with valid scores at each grade. The 2002 CSAP mathematics scales were placed on a common scale spanning grades 5 through 10 using the three methods. Standard analyses of calibration output indicated that the separate estimation method produced consistently better results than did the concurrent of pairwise concurrent estimation. (Contains 9 figures and 19 references.) (SLD)

Separate Versus Concurrent Calibration Methods in Vertical Scaling

Thakur Karkee, Daniel M. Lewis, Machteld Hoskens & Lihua Yao
CTB McGraw-Hill

Carolyn Haug
Colorado Department of Education

This paper is dedicated to Bradley A. Hanson

Introduction

Tracking student academic progress over time has become increasingly important for large-scale assessment programs. The national focus on educational reform, the reauthorization of the Elementary and Secondary Education Act, and state and local legislation have increased the stakes associated with student and school accountability. Standardized testing to measure progress on state content standards is a primary component of state educational accountability systems. Vertical scaling—the development of a common scale across grades of a given content area—has a number of advantages over the use of unique scales for each grade. Vertical scales can provide alternate methods to measure longitudinal growth at the student level and they permit the assessment of progress for students in earlier grades toward goals in subsequent benchmark grades on the same metric.

Vertical scales have traditionally not been incorporated into state accountability systems because most states have traditionally not assessed across contiguous grades. Also, the designs required to develop and maintain vertical scales are non-trivial, often requiring additional testing time for students and typically bringing additional costs to bear on the assessment program. However, there has been a recent increase in the incorporation of vertical scales into state assessment programs. This trend is likely to continue, particularly in light of the No Child Left Behind Act of 2001 (NCLB; Public Law 107-110), which requires and provides limited federal funds for testing in grades 3 through 8 in reading and mathematics.

Creative designs and technical solutions are needed to provide efficient, economical, and psychometrically defensible assessment programs that incorporate vertical scales as a component of state educational accountability programs. In this paper we address the development of vertical scales under a common item design. That is, we address the issues with regard to establishing a vertical scale across contiguous grades when students in adjacent grades take a common set of items as part of, or in addition to, their on-grade assessment.

Two methods to establish a common scale across grades within a content area using a common item design (separate and concurrent), have been previously studied (Hanson and Béguin, 2002; Kim & Cohen, 1998) under simulated conditions. These, and a third hybrid-method are examined in the present study, using operational test data, and are briefly described below.

1. Separate. The separate calibration method is accomplished in two steps: (1) separate calibration and (2) grade-by-grade chained linking. First, the parameters for items in each grade are estimated in separate calibration runs. Of course, this produces a unique metric for each grade. A common scale is achieved through the use of the common items. The scale for one of the grades is identified as the base scale. In the second step, the two sets of item parameter estimates for the items in common between the base grade and an adjacent grade are used to estimate a scale transformation that will place the item parameter estimates from the adjacent grade onto the scale of the base grade via a standard equating technique (e.g., Stocking & Lord, 1983). This second step is repeated for each adjacent grade until all grades are on the common base scale.
2. Concurrent. The second method, concurrent calibration, establishes the common scale in a single step—the calibration phase—by simultaneously estimating parameters for all items at all grades. As noted by Hanson and Béguin (2002) and Patz and Hanson (2002), the estimation software must be specifically designed to handle multiple groups to properly perform concurrent estimation in nonequivalent group designs, such as occur over multiple grades of achievement testing.
3. Pair-wise Concurrent. This method is a hybrid of methods 1 and 2. In the first step, concurrent scaling is employed for non-overlapping pairs of adjacent grades to establish a common scale for each adjacent grades pair. In the second step, one adjacent grades pair is identified as the base scale and the remaining adjacent grades pairs are placed on the base grade via the common items using standard equating techniques as in the separate calibration method.

Each method has merit. Concurrent calibration is efficient; all grades are calibrated in a single calibration run and a common scale emerges in this step as a result of the simultaneous calibration of the items common across grades. Further, the pooling of data across grades under concurrent calibration will produce more stable parameter estimates for the common items when sample sizes are small. However, it could be argued that separate calibration is more appropriate when multidimensionality may be present in the data. The problems introduced by multidimensional data may be ameliorated under the compromise model—pair-wise concurrent estimation—as the data may be more likely to be unidimensional between two adjacent grades of a content area than across a larger grade span.

Research conducted to date has not provided a definitive answer with regard to the best method for practice. Rather, a review of the literature provides a better understanding of the complexities involved in establishing a vertical scale, and different problems associated with the different methods. Kim and Cohen (1998) found that "when the number of common items is small, linking of separate calibration runs may be preferable to concurrent calibration." However, as the number of common items increased, the different methods "tended to yield similar results." Hanson & Béguin (2002) showed that concurrent estimation generally resulted in better performance than separate estimation when the model is correctly specified. In a later study that included polytomous items Kim and Cohen (2002) found that "recovery via concurrent calibration was consistently, albeit only slightly, better than recovery from separate calibration and linking for both item and ability parameters."

In their simulation study using different estimation programs and conditions, Hanson and Béguin (2002) concluded that concurrent estimation generally resulted in lower error than separate estimation. However, they conclude that the results of their study, together with other research on the topic, "are not sufficient to recommend completely avoiding separate estimation in favor of concurrent estimation."

This may be particularly true when model misspecification due to multidimensionality is considered. In Hanson and Béguin (2002), and other studies of this type, the data were simulated from the same unidimensional model used for item parameter estimation. Béguin, Hanson, and Glass (2000) and Hanson and Béguin (2002) studied the different estimation approaches using unidimensional IRT models with data generated from a two-dimensional model. They found significant error in the non-equivalent groups condition and several cases in which there was less error when using separate estimation compared to concurrent estimation. Yen (1985) showed that the variance of the estimated item difficulties and traits will decrease, that is, the scale will shrink when a unidimensional scaling model is applied to a multidimensional test.

Hanson and Béguin (2002) note "an important further research question is how well separate and concurrent estimation perform when the model is misspecified to some degree, as would occur with real data." It is this question that is the focus of the current study—a comparison of the different methods of estimation with real data. The recent studies cited above were conducted using simulated data with two non-equivalent groups. Given the No Child Left Behind Act of 2001 (NCLB; Public Law 107-110), it is likely that vertical scales with six or more non-equivalent groups at contiguous grades will become common. It is important that real data be analyzed using the various methods to better understand how robust the various models are.

Research has not conclusively identified a single estimation procedure as the method of choice for establishing a vertical scale in practice. However, state accountability systems are currently employing vertical scales to measure growth across years as part of their operational accountability systems. This study is intended to further the knowledge base and inform the process of establishing vertical scales that result in scale stability across years within grade and across grades within year.

In the current study, we seek to extend previous research investigating separate and concurrent estimation methods by (a) using data obtained from an operational, mixed format (open-ended and selected-response), state assessment, (b) using data across a span of six contiguous grades linked via a common item design, (c) adding pair-wise concurrent estimation—a hybrid of the separate and concurrent estimation methods—to the methods examined.

Given the results of the simulation studies cited above, we do not expect to find large differences in the results of the separate or concurrent estimation methods. Because of the hybrid-nature of the pair-wise concurrent method, where differences do exist between the separate and concurrent methods, we expect the pair-wise method to produce results that depart less from the other two methods than they do from each other.

Data

The data for this study were obtained from responses to the 2002 Colorado Student Assessment Program (CSAP) mathematics assessments in grades 5 through 10. The configuration of the 2002 CSAP mathematics assessments is shown in Table 1. Mathematics content is measured by six standards: Number Sense; Patterns, Functions, and Algebra; Data Analysis, Probability, and Statistics; Geometric Concepts; Measurement; and Operation and Calculation.

Table 1. Configuration of the 2002 CSAP Mathematics Assessments

	Maximum Possible	Total No. of Items		Frequency of CR Items with the Given Number of Maximum Points			
Grade	Points	SR	CR	1	2	3	4
5	96	54	15		6	6	3
6	87	45	15		6	6	3
7	87	45	15		6	6	3
8	86	44	15		6	6	3
9	87	45	15		6	6	3
10	87	45	15		6	6	3

Common Item Design

The assessments were constructed to have at least 20 items in common between adjacent grades. This was accomplished by (a) constructing a "core" test at each grade based on the test frameworks, (b) selecting 10 items from each core test that were appropriate in terms of standards, curriculum, and difficulty for the grade above (except in grade 10 where there is no tested grade above), and (c) similarly selecting 10 items appropriate for the grade below (except in grade 5 where there is no tested grade below). In order to maintain the test framework at each grade, the set of common items was specified to be representative of the overall test in terms of standards representation, range of difficulty, and when possible, item format. Both selected-response (SR) and constructed-response (CR) items were used as common items. The common items appeared in approximately the same location in each test in which the item appeared. There were 20 items in common between grades 6 and 7, between grades 7 and 8, and between grades 8 and 9. There were 25 items in common between grades 5 and 6 and between grades 9 and 10.

Sample

The data set for each grade (5-10) consisted of 10,000 cases, randomly selected from the population of approximately 54,000 tested students with valid scores at each grade. The data set was randomly split into two sets of 5000 each. The first data set was used for calibration and the second data set was used for subsequent cross-validation analyses (item and test residuals).

Method

Item Response Theory Models

Two IRT models were used to calibrate the operational test items. The three-parameter logistic (3PL) model (Lord, 1980) was used to estimate parameters for the selected-response items. The two-parameter partial credit (2PPC) model (Muraki, 1992; Yen, 1993) was used to estimate parameters for the constructed response items.

Calibration Method and Software

The PARDUX (Burket, 2002) microcomputer software program was used for all calibrations and equatings in the present study. PARDUX can estimate IRT models for both single groups and for multiple groups. In the single group case PARDUX constrains the mean and SD of the person ability distribution equal to 0 and 1, respectively, during the item parameter estimation process to obtain model identification. In the multiple group case PARDUX constrains the mean and SD of the person ability distribution equal to 0 and 1, respectively, for one of the groups and estimates the means and SDs for all other groups during the item parameter estimation process, in order to obtain model identification (for more details, see Patz & Hanson, 2002). Marginal Maximum Likelihood is used to estimate item parameters. Maximum likelihood estimates are used to estimate the person abilities in PARDUX.

The PARDUX program can also be used for equating and linking. PARDUX uses the Stocking and Lord (1983) test characteristic curve method for equating/linking.

Assessment of Item Fit to the IRT Model

Item fit was assessed using the Q_1 statistic described by Yen (1981) for the dichotomously scored items and using a generalization of this statistic for the multi-level items. As described by Yen, Q_1 is a Pearson chi-square of the form

$$Q_{1j} = \sum_{i=1}^{I} \frac{N_{ji}(O_{ji} - E_{ji})^2}{E_{ji}} + \sum_{i=1}^{I} \frac{N_{ji}[(1 - O_{ji}) - (1 - E_{ji})]^2}{1 - E_{ji}} ,$$

where N_{ji} is the number of examinees in cell i for item j. O_{ji} and E_{ji} are the observed and predicted proportions of examinees in cell i that attain the maximum possible score on item j, where

$$E_{ji} = \frac{1}{N_{ji}} \sum_{a \in i}^{N_{ji}} P_j\left(\hat{\theta}_a\right).$$

The generalization of Q1 for multi-level items can be stated as

$$Q_{1j} = \sum_{i=1}^{I} \sum_{k=1}^{m_j} \frac{N_{ji}(O_{jki} - E_{jki})^2}{E_{jki}} ,$$

where

$$E_{jki} = \frac{1}{N_{ji}} \sum_{a \in i}^{N_{ji}} P_{jk}\left(\hat{\theta}_a\right).$$

O_{jki} is the observed proportion of examinees in cell *i* who performed at the *k-th* score level.

The chi-squared statistics are affected by sample size and extreme expectations, and their degrees of freedom are a function of the number of independent observations entering into the calculation minus the number of parameters estimated. Items with more score levels have more degrees of freedom, making it awkward to compare fit for items that differ in the number of score levels. To simplify, the following standardization of the Q_1 statistic was used:

$$Z_j \equiv \frac{Q_{1j} - df}{\sqrt{(2df)}}.$$

The Z-statistic is an index of the degree to which obtained proportions of students with each item score are close to the proportions that would be predicted by the estimated thetas and item parameters. To use this standardized statistic to flag items for potential misfit, a critical value for Z as a function of sample size is computed and item Z-values above this critical Z-value may indicate poor model fit.

Non-convergence of items. PARDUX flags an item as non-converged if it did not meet the convergence criterion. The convergence criterion is the maximum change across iterations in the gamma parameter for a constructed-response item and in the b parameter for a selected-response item. In the present study the convergence criterion was set at 0.001.

The Vertical Scaling Methods

The 2002 CSAP Mathematics scales were placed on a common scale that spans grades 5 through 10 using the three methods described previously—separate estimation, concurrent estimation, and pair-wise concurrent estimation. A graphical representation of the three estimation methods and the method used to place the three approaches on a common scale to facilitate analyses is presented in Figure 1.

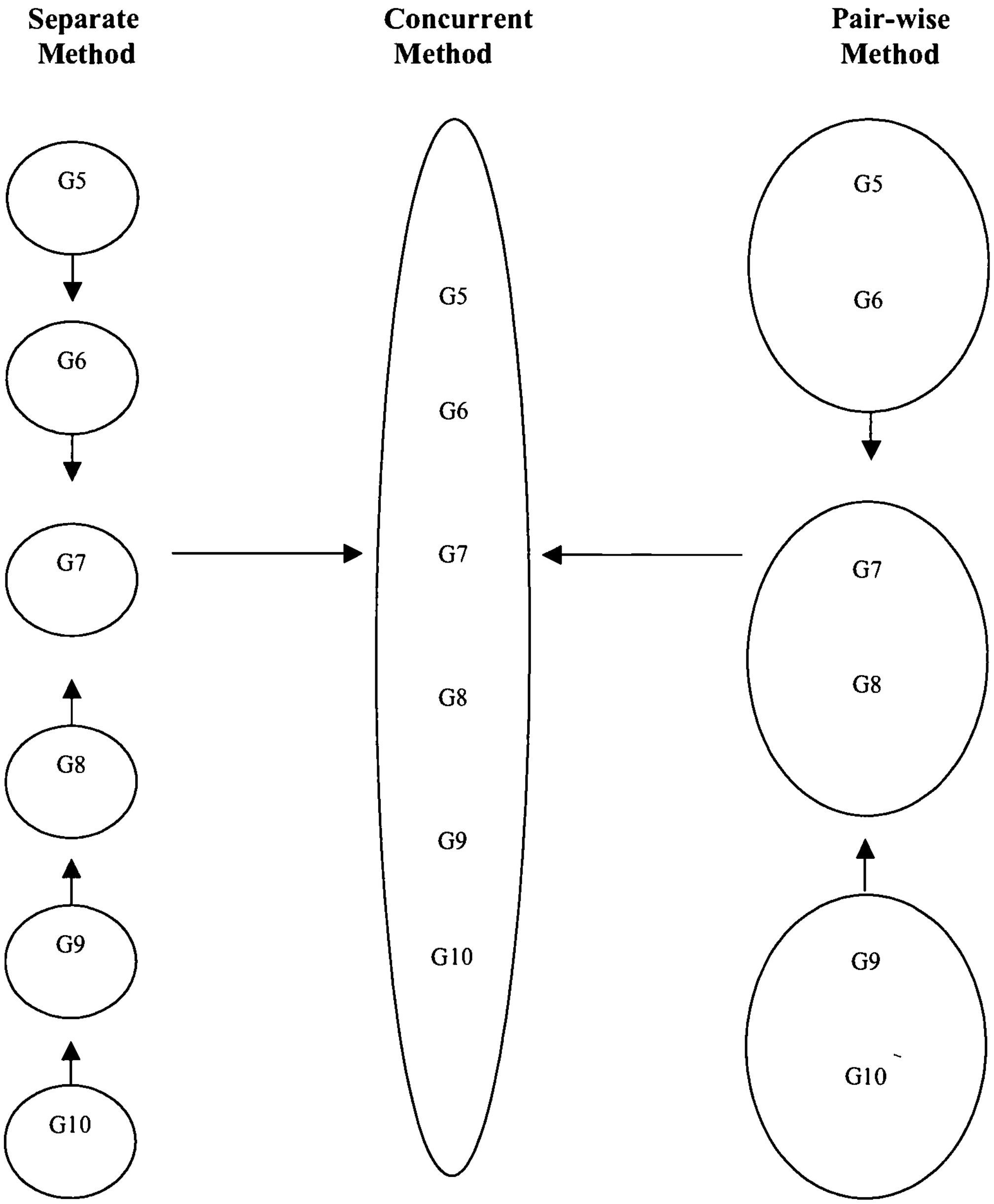
Separate estimation proceeded by first estimating parameters for each grade separately followed by a chain linking, that is, by using the common items between adjacent grades to link the scale across grades via the Stocking and Lord procedure. The result of the calibration of the Grade 7 items was considered the base scale. A chain of Stocking and Lord links created a single scale—the common items between grades 6 and 7 were used to place the grade 6 item parameters on the base scale. The parameters for items common to grades 5 and 6 were then used to link the grade 5 item parameters to the base scale. Links between grades 8, 9, and 10 were similarly established to place the item parameters across the grades on the base scale as indicated for separate estimation in Figure 1.

The concurrent calibration was implemented by estimating the item parameters for all grades simultaneously via the multiple group capability of PARDUX.

The pair-wise estimation method is a hybrid of the concurrent and separate methods. First, concurrent calibration was implemented for pairs of adjacent grades. That is, the parameters for items in grades 5 and 6 were simultaneously estimated using the combined data set for students in these grades. Similarly, the items in grades 7 and 8 were concurrently estimated, as were the items in grades 9 and 10. The result of the concurrent calibration of the grades 7 and 8 item parameters was considered the base scale, as illustrated by Figure 1. The parameters for items common to grades 8 and 9 were used to link the grades 9 and 10 parameters to the base scale. Similarly, the parameters for items common to grades 6 and 7 were used to link the grades 5 and 6 parameters to the base scale.

Because the results of each estimation and linking method produce a unique vertical scale, a final step was required to make the three vertical scales (separate, concurrent, and pair-wise concurrent) comparable for subsequent analyses. This was accomplished by considering the concurrent estimation results as the base vertical scale. The Stocking and Lord equating procedure was implemented using the grade 7 item parameters, as indicated in Figure 1. That is, the concurrent estimation parameters for all items in the grade 7 assessment were used as anchors to obtain transformation constants to place the grade 7 separate estimation item parameters on the base (concurrent estimation) scale. These constants were applied to the separate estimation parameter estimates across the grades. This procedure was repeated to link the pair-wise concurrent estimation vertical scale on the base (concurrent estimation) scale. As a final step, each scale was transformed to an operational scale with a linear transformation via a multiplicative constant of 60 and additive constant of 550.

Figure 1. Graphical representation of the Separate, Concurrent, and Pair-wise Methods of Calibration and Linking



Evaluation and Comparison Criteria

A strong rationale for the use of simulation as a means of studying concurrent and separate estimation methods in previous studies (e.g., Hanson and Béguin, 2002; Kim & Cohen, 1998) is that objective criteria exist to evaluate the results (e.g., the accuracy of the recovery of input parameters used to generate the simulated data). This is not the case when operational data is used; item parameters and students' true scores are not known. However, in the current study, the results of the implementation of the different estimation and linking methods used to establish vertical scales are compared in terms of several common criteria.

First, the results are compared using common methods of evaluating calibration results. These include an examination of (a) non-convergence of items, (b) model fit, and (c) differential item functioning.

Second, the properties of the parameter and ability estimates can be observed and compared for the three estimation methods after they are placed on the base (concurrent) scale as described previously. Following the practice of Kolen and Brennan (in preparation) the vertical scales derived from the three methods can be compared in terms of specific properties including the pattern of grade-to-grade growth (means), grade-to-grade variability, and separation of grade distributions.

Third, a comparison of expected versus observed performance is conducted for the three methods using the cross-validation data set by computing item and test residuals for each method. Note that the parameters used in this analysis are those derived from the 5000 case calibration data set. Ability estimates and the subsequent comparisons of observed versus expected performance on items are based on the 5000 case cross-validation data set.

The item parameters for the three methods are applied to the cross-validation data set to obtain student ability estimates ($\hat{\theta}$). The theta ($\hat{\theta}$) and parameter estimates are then applied to compare predicted versus observed item responses on the cross-validation data set, as described below.

Let Y_{ij} be the observed score (see Data Vector 1, below) for the jth student on the ith item. Further, let $\hat{Y}_{ij}/\hat{\theta}_j, \hat{\beta}_i$ be the predicted score (see Data Vector 2) for the jth student with ability $\hat{\theta}_j$ on the ith item with parameters ($\hat{\beta}_i$: a_i, b_i, c_i, α_i, γ_{i1}, γ_{i2},…).

Data Vector 1
(Observed Scores)

$$\begin{pmatrix} y_{11}, y_{21}, \ldots\ldots\ldots\ldots y_{i1} \\ y_{12}, y_{22}, \ldots\ldots\ldots\ldots y_{i2} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{1j}, y_{2j} \ldots\ldots\ldots\ldots y_{ij} \end{pmatrix}$$

Data Vector 2
(Predicted Scores)

$$\begin{pmatrix} \hat{y}_{11}, \hat{y}_{21}, \ldots\ldots\ldots\ldots \hat{y}_{i1} \\ \hat{y}_{12}, \hat{y}_{22}, \ldots\ldots\ldots\ldots \hat{y}_{i2} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \hat{y}_{1j}, \hat{y}_{2j}, \ldots\ldots\ldots\ldots \hat{y}_{ij} \end{pmatrix}$$

The item-score residual for item i is calculated as the square root of the sum over students of the square of the difference between the observed and predicted score for each item.

$$residual_{item.i} = \sqrt{\sum_j (y_{ij} - \hat{y}_{ij} / \hat{\theta}_j, \beta_i)^2} \text{ , where j = 1, 2, ..., 5000.}$$

The test-score residual, $\text{residual}_{\text{test.g}}$, provides the mean item residual for the items in the grade g assessment.

$$residual_{test.g} = \frac{\sqrt{\sum_i \sum_j (y_{ij} - \hat{y}_{ij} / \hat{\theta}_j, \beta_i)^2}}{q} \text{, where i = 1, 2, ..., t indexes the set of t test}$$

items and q is the maximum obtainable number correct score for the set of t items in grade g.

The residual analyses are conducted for the complete set of items comprising each grade's assessment as described above. A second, similar, but more stringent residual analysis was also conducted. In this case, the ability estimate for student j, $\hat{\theta}_j$ ', is based on student j's responses to (and parameter estimates associated with) only the common items on a given assessment. The item residuals are computed for the unique (non-common) items only, that is, when Data Vector 1 and Data Vector 2 consist of observed and expected scores, respectively, for only the unique items in the grade's assessment. That is, $residual_{item(unique).i} = \sqrt{\sum_j (y_{ij} - \hat{y}_{ij} / \hat{\theta}_j ', \beta_i)^2}$, where θ_j ' is the ability estimate for student j based on the common items for the grade, j = 1, 2, ..., 5000, and item i is a unique item.

$$residual_{test(unique).g} = \frac{\sqrt{\sum_k \sum_j (y_{ij} - \hat{y}_{ij} / \hat{\theta}_j ', \beta_i)^2}}{m} \text{, where k = 1, 2, ..., n indexes the}$$

set of n unique items and m is the maximum obtainable number correct score for the set of unique items in grade g.

Results

Analysis of Calibration Results

A pre-calibration analysis of item statistics was conducted. A review of the item statistics indicated that the data was accurate. In particular, a review of the common item statistics indicated that the items were functioning as would be expected. Item p-values were higher for the common items in the higher of the two adjacent grades in which the item appeared for all but 5 common items across the six grades. One of the items for which the p-value was higher for the lower grade did not converge. Three of the remaining five items were within 0.01 in p-value for the two grades and the other two were within 0.03.

Convergence and Model Fit. Non-converging items and items flagged by PARDUX as potentially exhibiting poor fit (as described previously) are indicated in Table 2; mean fit values (Z-values) for the items by method are shown in Table 3. Note that 17 different items either did not converge or were flagged for poor fit for one or more of the three methods. These items are labeled item 1 to item 17 in Table 2 to allow the items to be tracked across methods or, in the case of common items, across grades. These item numbers do not indicate position within the actual assessments. Common items are indicated in bold. Selected-response items are indicated by an asterisk superscript.

Table 2. Non-Converging Items and Items Flagged as Potentially Exhibiting Poor Model Fit

	Non-Converged Items			Items Flagged for Potential Poor Fit		
Grade	Concurrent	Pair-wise	Separate	Concurrent	Pair-wise	Separate
5				**1, 2**	**1, 2,** 3*	1, 3*
6				**1, 2,** 4	**1, 2,** 4	**2,** 4
7				0	0	0
8	**17***		**17***	**5***, 6, 7, **8**, 9	6, 7, **8**, 9	7, **8**, 9
9	**17***	**17***		**5***, **8**, 10, **12**, **13***	**8**, 10, **12**, **13***	**8**, 11
10				**12**, **13***, 14*, 15, 16,	**12**, **13***, 14*, 16,	14*

* Indicates a selected-response item
Common items are indicated in bold
The item numbers are a table index only; they do not indicate test position.

Table 3. Mean Z-value Across Methods by Grade

Grade	Concurrent	Pair-wise	Separate
5	2.7	3.1	2.0
6	4.2	3.8	2.6
7	2.9	2.3	1.9
8	5.1	4.1	3.7
9	5.0	3.9	2.8
10	3.5	3.5	2.1

As indicated in Table 2, one item (labeled item 17 in Table 2) did not converge for some of the estimation conditions. Item 17 is a common item to the grades 8 and 9 assessments. It did not converge (a) in the concurrent estimation, (b) in grade 9 for the pair-wise concurrent estimation (in the grade 9-10 adjacent grade pair), or in grade 8 in the separate estimation. It *did* converge (a) in grade 8 for the pair-wise concurrent estimation (in the grade 7-8 adjacent grade pair) and (b) in grade 9 in the separate estimation. The item is extremely difficult and is one of the common items for which the p-value for the lower grade (0.14 in grade 8) was higher than that of the upper grade

(0.12 in grade 9) with a p-value of about 0.13 for grade 8 students and 0.12 for grade 9 students.

The data in Table 3 indicate that the separate estimation method resulted in better average fit for the items in every grade. The pair-wise estimation resulted in better average fit than the concurrent method in all grades except grade 5.

The separate calibration method flagged fewer items as potentially exhibiting poor fit than did the concurrent or pair-wise methods. Nine different items were flagged under the separate method and thirteen and fourteen different items were flagged under the pair-wise and concurrent methods, respectively. Most of the items flagged for model fit are constructed-response items; 7 of the 9 items (78%) in the separate calibration, 11 of the 14 items (79%) in the concurrent calibration, and 10 of the 13 items (77%) in the pair-wise concurrent calibration were flagged for misfit. The items flagged by the separate calibration are generally also flagged by the concurrent and pair-wise calibration; the only exception was item 11 in grade 9, which was only flagged for model fit under the separate method. Figures 2 and 3 consist of observed and expected percent of maximum score curves for item 1 in grades 5 and 6, respectively. Item 1 is a 2-point constructed-response item. The smooth curve in Figures 2 and 3 represent the expected percent of maximum score across the ability range. The jagged curve represents the observed percent of maximum score across the ability range.

Item 1 was flagged for poor fit by all three methods in grade 5. As illustrated in Figure 2, item 1 showed poorer fit for the concurrent and pair-wise estimations than for the separate calibration in grade 5. Student performance was generally underestimated across the ability distribution by the concurrent and pair-wise concurrent methods for grade 5 students (Figure 2) and was overestimated in the region where the item provided the most information in grade 6 (Figure 3).

Figure 2. Observed and Expected Percent of Maximum Score Curves for Item 1 in Grade 5

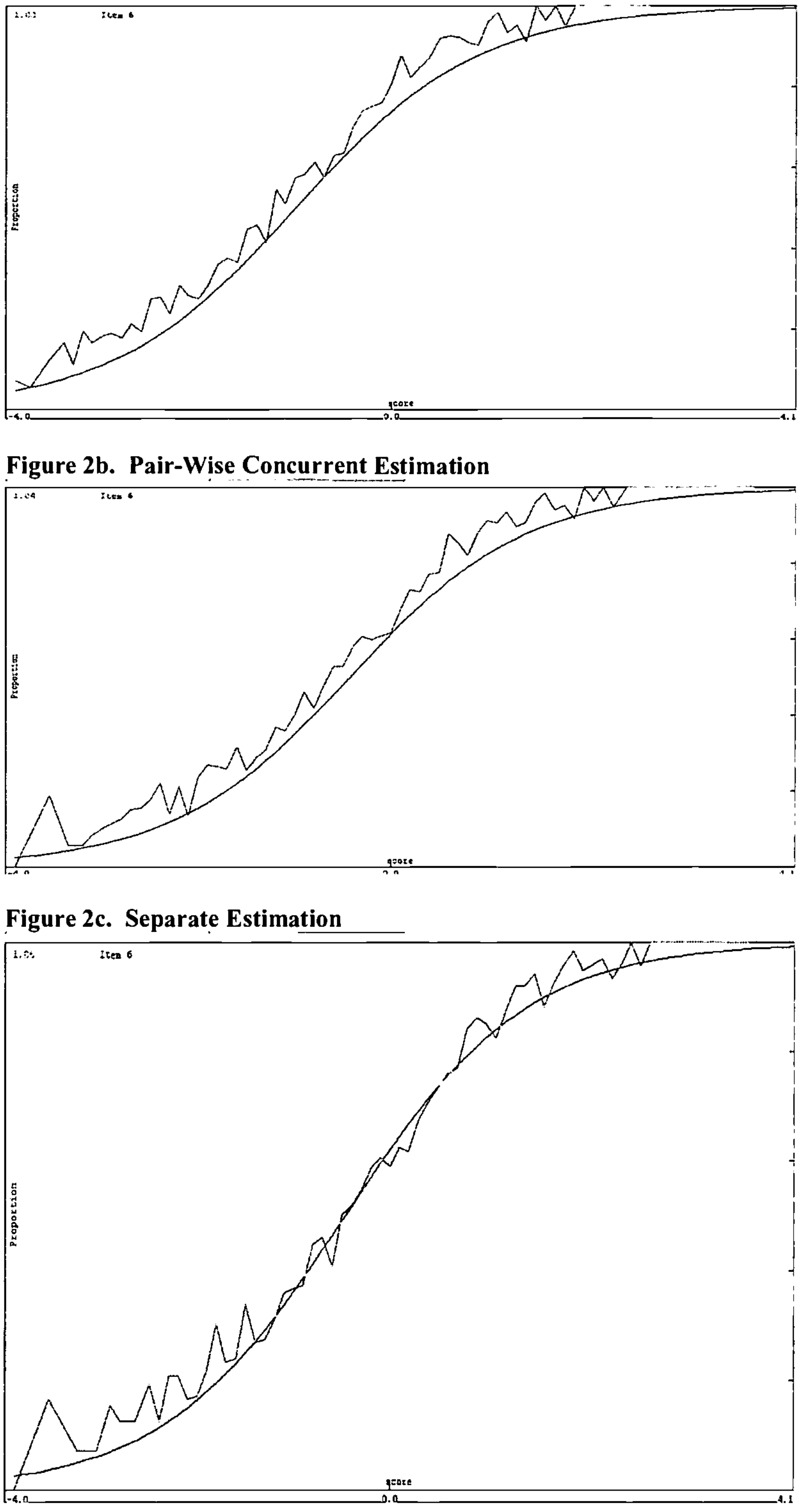


Figure 2a. Concurrent Estimation

Figure 2b. Pair-Wise Concurrent Estimation

Figure 2c. Separate Estimation

Figure 3. Observed and Expected Percent of Maximum Score Curves for Item 1 in Grade 6

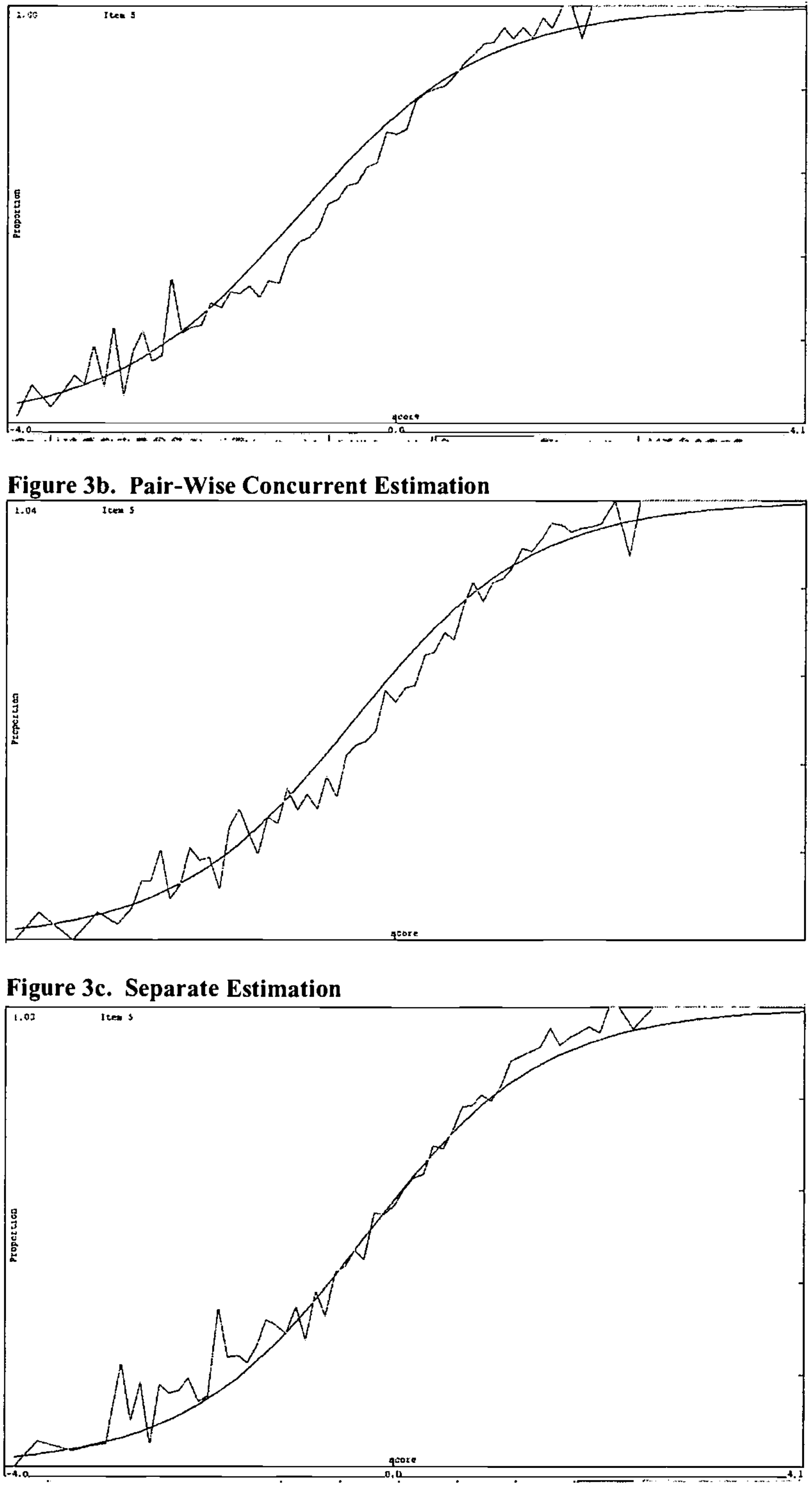


Figure 3a. Concurrent Estimation

Figure 3b. Pair-Wise Concurrent Estimation

Figure 3c. Separate Estimation

Differential Item Functioning. Differential item functioning was evaluated using a variation of the Linn and Harnisch (1981) method. The items flagged for differential item functioning (DIF) by the three methods are indicated in Table 4. As a rule of thumb, DIF analyses are not conducted when there are fewer than 50 cases per focal group in a given grade, as occurred for Native Americans in grade 5. There were generally similar case counts for each focal group across grades. There were an average of 59 Native Americans in grades 6-10, and averages of 140 Asian, 282 Black, 1011 Hispanic, 3491 White, 2503 Male, and 2491 Female students in grades 5-10.

Table 4. Items Flagged for Differential Item Functioning

				Concurrent			
	Ethnicity					Gender	
Grade	Native American	Asian	Black	Hispanic	White	Male	Female
5	ND	$\mathbf{1}^{+}$, 2^{*-}, 3^{*+}		$\mathbf{1}^{+}$			
6		4^{*-}, 5^{+}		$\mathbf{1}^{-}$			
7		6^{*-}, 8^{*-}				7^{-}, $\mathbf{9}^{+}$	7^{+}, 10^{+}
8	13^{*+}	11^{*-}	12^{-}			$\mathbf{9}^{+}$	$\mathbf{9}^{-}$
9		14^{+}, 17^{*-}, 18^{*-}, 19^{-}	15^{*+}, 16^{*+}, 20^{-}				
10		21^{-}, 23^{-}, $\mathbf{24}^{*+}$	$\mathbf{24}^{*+}$	22^{+}			
				Pair-wise			
5		$\mathbf{1}^{+}$, 2^{*-}, 3^{*+}		$\mathbf{1}^{+}$			
6		5^{+}		$\mathbf{1}^{-}$			**1-**
7	25-	6*-, 8*-				7^{-}, $\mathbf{9}^{+}$	7^{+}, 10^{+}
8	13^{*+}		12^{-}				$\mathbf{9}^{-}$
9		14^{+}, 17^{*-}, 19^{-}	16^{*+}, 20^{-}				
10		21^{-}, 23^{-}, $\mathbf{24}^{*+}$	$\mathbf{24}^{*+}$	22^{+}			
				Separate			
5							
6							
7		6^{*-}, 8^{*-}				$\mathbf{7}^{-}$	$\mathbf{7}^{+}$
8			12^{-}			$\mathbf{9}^{+}$	$\mathbf{9}^{-}$
9	26^{*+}	17^{*-}, 19^{-}					
10		21^{-}, 23^{-}	27^{-}				

* indicates a selected-response item
-DIF Flagged as disfavoring the group
+ DIF Flagged as favoring the group
Common items are indicated in bold
ND: No DIF analyses performed. Fewer than 50 members of the focal group.
The item numbers are a table index only; they do not indicate test position.

Note that 25 different items were flagged for DIF by one or more of the three methods. These items are labeled item 1 to item 26 in Table 4 to allow the items to be tracked in the table across methods or, in the case of common items, across grades. These item numbers do not indicate position within the actual assessments; they do not necessarily indicate the same items as were indexed in Table 2. Common items are indicated in bold.

Selected-response items are indicated by an asterisk superscript. A superscript plus (minus) sign indicates the DIF was favoring (disfavoring) the focal group.

The results indicate that fewer items were flagged for DIF under separate calibration than concurrent or pair-wise estimation methods. Eleven different items were flagged for DIF for one or two focal groups under separate estimation, twenty-one different items were flagged for DIF for one or two focal groups under pair-wise concurrent estimation, and twenty-four different items were flagged for DIF for one or two focal groups under concurrent estimation.

Item 1 in Table 4 is also item 1 in Table 2; this item tended to underestimate overall performance in grade 5 and overestimate performance in grade 6 for the full sample of students. As indicated by Table 4, this relationship holds for Hispanic students relative to the rest of the population in each grade, that is, in grade 5 Hispanic student performance is underestimated by the item parameters and overestimated in grade 6 for the concurrent and pair-wise concurrent methods. The common items for which DIF is manifest do not necessarily do so in both grades in which the item appears. Items 1 and 24 are common items flagged for DIF favoring Asian students in grades 5 and 10, respectively, but not in the adjacent grades (6 and 9, respectively) for both concurrent and pair-wise concurrent methods. Item 9 is a common item flagged for DIF disfavoring females in grade 8 but not in grade 7 for all three methods.

Properties of the parameter and ability estimates

Recall that the results of each estimation and linking method produced a unique vertical scale. To facilitate the comparison of parameter and ability estimates across the three methods, the concurrent estimation results were treated as the base vertical scale and the separate and pair-wise concurrent methods were placed on the base scale via the Stocking and Lord (1983) equating procedure using the grade 7 item parameters, as described previously. The transformation constants used to (a) link each grade or adjacent grade pair to the base scale for the separate or pair-wise concurrent estimation methods and (b) to place the separate and pair-wise methods onto the base (concurrent) vertical scale can be seen in Appendix A.

In the next parts of this section, we compare the results within and across grade(s) and method(s). We evaluate the vertical scale properties by following the practice of Kolen and Brennan (in preparation); we observe the pattern of grade-to-grade growth, grade-to-grade variability, and separation of grade distributions. The calibration data-set response vectors were scored using item parameters on the base vertical scale metric for each method. The lowest and highest obtainable scale scores (same across methods) and observed score means and standard deviations derived from the calibration data set for each method are presented in Table 5. The column labeled "Change in Mean from Previous Grade" indicates the scale score increase or decrease in the mean of the given grade from the mean of the adjacent grade below. The column labeled "SDM" is the standardized difference in means and presents the "Change in Mean from Previous Grade" in terms of the pooled standard deviation units.

Table 5. Observed Score Means and Standard Deviations (Calibration Data Set)

			Separate				Concurrent				Pair-wise			
Grade	LOSS	HOSS	Mean	SD	Change in Mean from Previous Grade	SDM	Mean	SD	Change in Mean from Previous Grade	SDM	Mean	SD	Change in Mean from Previous Grade	SDM
5	220	800	501	67			500	66			496	74		
6	240	830	516	67	15	0.22	515	67	16	0.23	515	72	19	0.26
7	280	860	529	66	13	0.20	528	66	13	0.20	528	67	13	0.19
8	310	890	544	64	15	0.23	545	65	16	0.24	545	64	17	0.26
9	340	920	555	66	10	0.16	554	64	9	0.14	555	68	9	0.14
10	370	950	572	65	17	0.26	568	64	14	0.22	572	66	17	0.26

SD=Standard Deviation, SDM=Standardized Difference in Means
LOSS = lowest obtainable scale score for the grade; HOSS = highest obtainable scale score for the grade

Within grade differences. As indicated in Table 5 and represented graphically in Figure 4a, the within grade means for grades 6 through 9 differ by at most one point across methods. In grade 5, the mean scale score for the pair-wise concurrent method (496) is four points lower than for the concurrent method (500) and five points lower than for the separate method (501). In grade 10, the mean scale score for the concurrent method (568) is four points lower than the common mean of the other two methods (572).

Also indicated in Table 5 and represented graphically in Figure 4b, the variability of the score distributions were more similar for the separate and concurrent methods than that of the pair-wise concurrent method. The standard deviations for the separate and concurrent methods were within 2 points of each other at each grade. The pair-wise method was similar to the other methods in grades 7 through 10, but was 5 points greater in grade 6 and up to 8 points greater in grade 5.

The pattern of grade-to-grade growth. As indicated in Table 5 and represented graphically in Figure 4a, the mean scale score increases with grade for each method, as expected, with similar but slightly different patterns of growth. Growth trends, as measured by the change in the mean from the previous grade, are quite similar except at the end grades. The pair-wise concurrent method shows 3 and 4 points more growth from grade 5 to grade 6 than the concurrent and separate methods. The concurrent method shows 3 fewer points of growth from grade 9 to grade 10 than do the other two methods. In terms of scale score separation across grades, the pair-wise method showed the most overall growth; the difference between the grade 10 and grade 5 mean score is 75 points as opposed to a difference of 70 and 68 points for the separate and concurrent methods. The pair-wise method also had the least consistency of growth. These growth trends across methods are also reflected in the SDM.

Figure 4a. Calibration Sample Mean

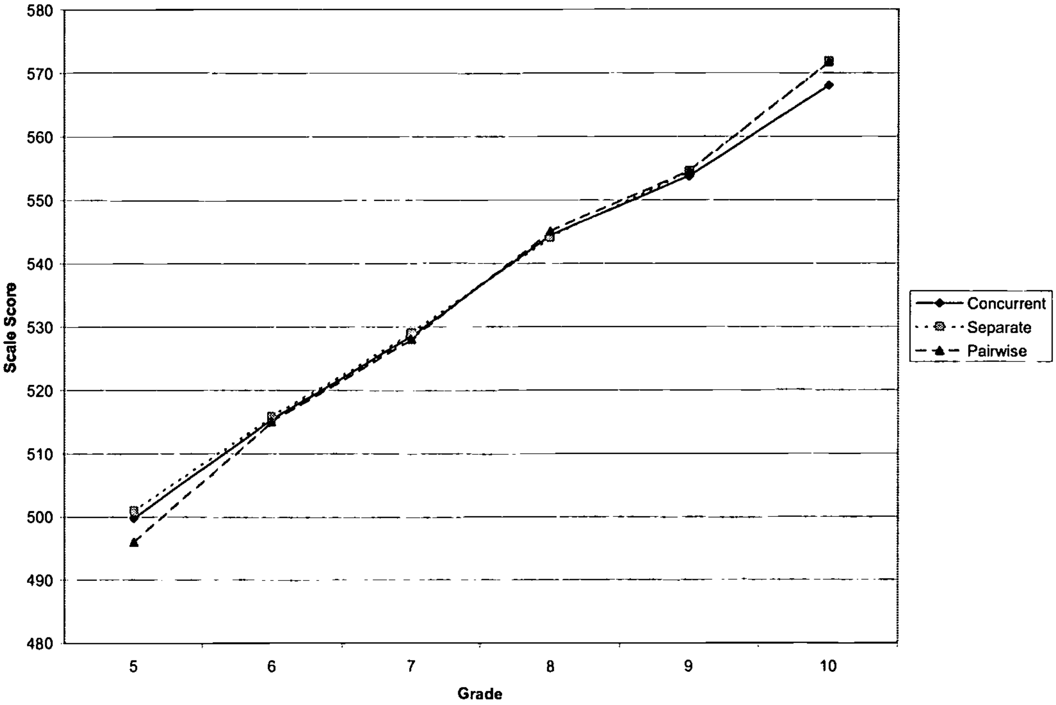
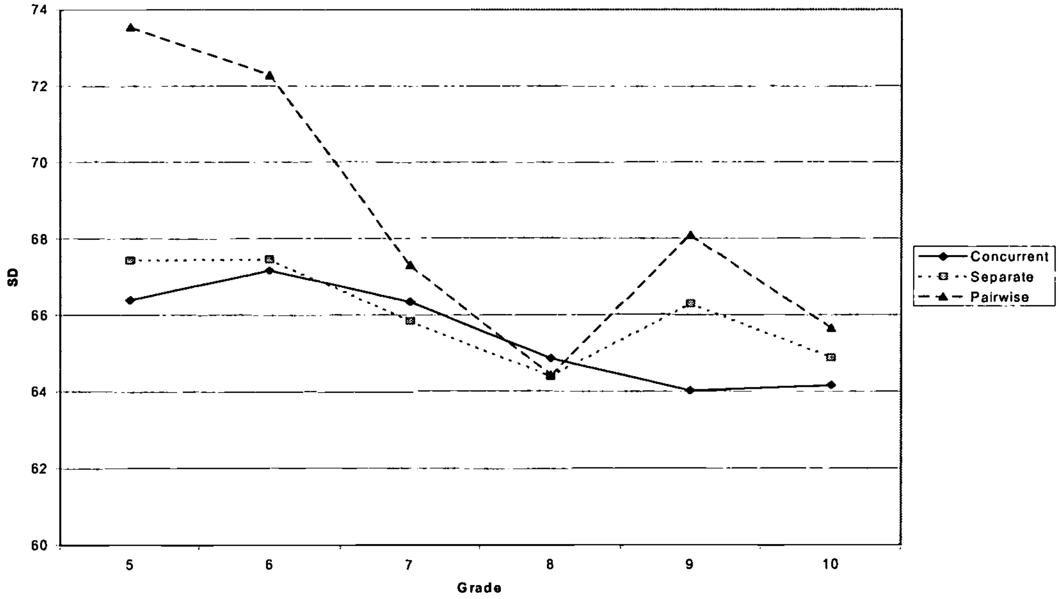


Figure 4b. Calibration Sample Standard Deviation



The pattern of grade-to-grade variability. As indicated in Table 5 and represented graphically in Figure 4b, the pattern of grade-to-grade variability is reasonably flat for the separate and concurrent methods; as indicated in Table 5, the standard deviations remain within 3 points across the grades for these two methods, ranging between 64 and 67. The variability of the assessments decreases from grade 5 to grade 8, but increases, then decreases in grades 9, and 10; the pair-wise method has less consistent standard deviations across the grades, ranging between 64 and 74. The standard deviations for grades 5 and 6 for the pair-wise method tend to be outliers, at 74 and 72, respectively, with 67 being the maximum standard deviation across the grades for the other two methods and 68 being the maximum standard deviation in the remaining 4 grades for the pair-wise method.

The scale scores for each method at the 10th, 25th, 50th, 75th, and 90th percentiles are provided in Table 6. An examination of the data in Table 6 allows one to monitor growth not only at the mean but also across the ability distribution. The concurrent and separate estimation results are quite similar for grades 5 through 8; the scale scores at each of the five percentile ranks are within 2 points at each grade for these two methods. These two methods are also similar at and below the 50th percentile in grades 9 and 10, but diverge somewhat in the upper range of the scale, with the separate estimation results being 5 and 8 points higher than the concurrent method at the 90th percentile in grades 9 and 10, respectively. The pair-wise method tends to produce lower scales scores at the below the median and higher scale scores above the median in grades 5 and 6 than do the other two methods. All three methods are quite similar in grades 7 and 8. In grades 9 and 10, the pair-wise method produces results similar to the separate method across the ability range.

Table 6. Scale Scores at Specified Percentiles by Estimation Method

				Percentile		
Grade	**Method**	**10**	**25**	**50**	**75**	**90**
Grade 5	Concurrent	416	459	503	543	579
	Pair-wise	402	451	500	544	584
	Separate	416	459	504	545	581
Grade 6	Concurrent	431	475	519	559	594
	Pair-wise	425	472	518	563	600
	Separate	432	475	519	560	595
Grade 7	Concurrent	440	491	535	574	605
	Pair-wise	439	491	535	574	605
	Separate	441	491	535	574	606
Grade 8	Concurrent	462	507	551	590	619
	Pair-wise	465	509	551	589	619
	Separate	464	507	549	588	618
Grade 9	Concurrent	470	518	563	598	626
	Pair-wise	466	517	564	601	631
	Separate	469	515	563	601	631
Grade 10	Concurrent	488	537	579	611	639
	Pair-wise	489	538	581	616	647
	Separate	491	538	580	616	647

The separation of grade distributions. Figure 5 presents the cumulative frequency distributions for all three methods across the six grades. Figures 6a, 6b, 6c, 6d, 6e, and 6f present the cumulative frequency distributions for the three methods for grades 5, 6, 7, 8, 9, and 10, respectively. As indicated by Figures 5 and 6 each of the three methods results in reasonable and similar separation of the grade distributions.

Figure 5. Cumulative Frequency Distributions for the Three Methods Across the Six Grades

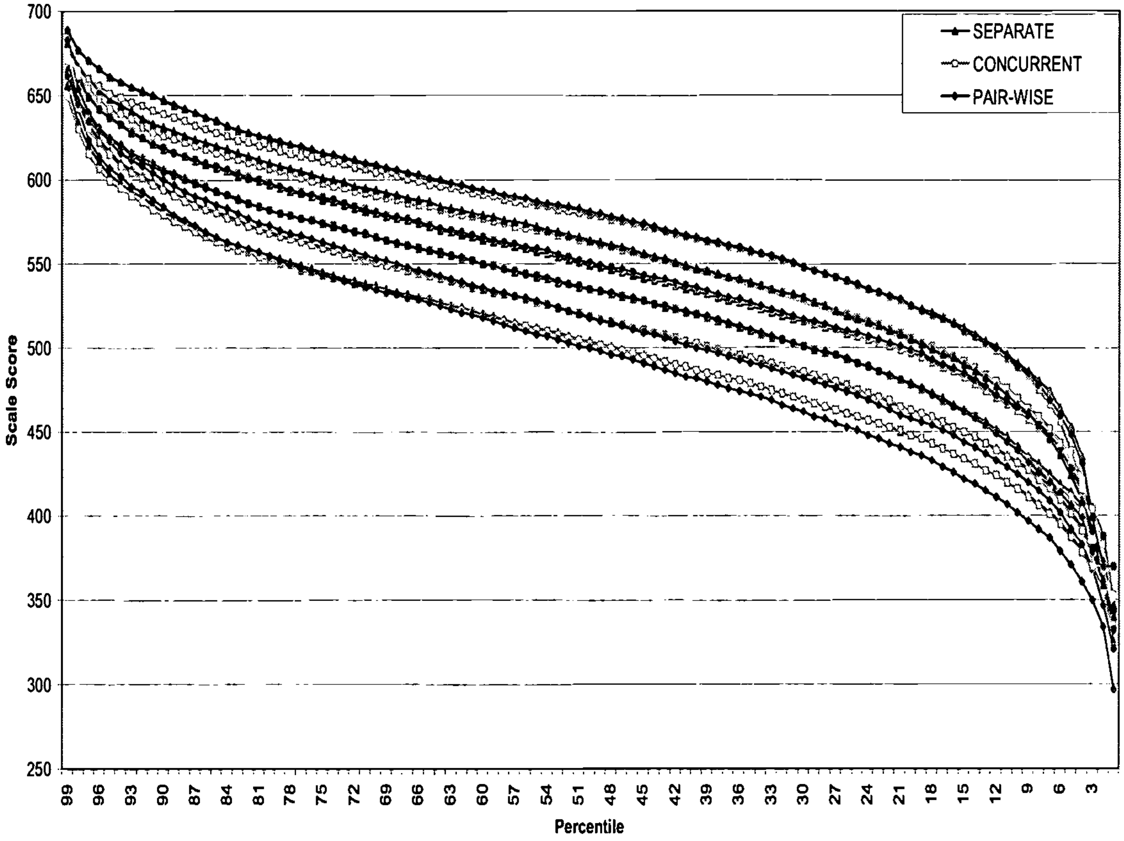


Figure 6. Cumulative Frequency Distributions for the Three Methods by Grade

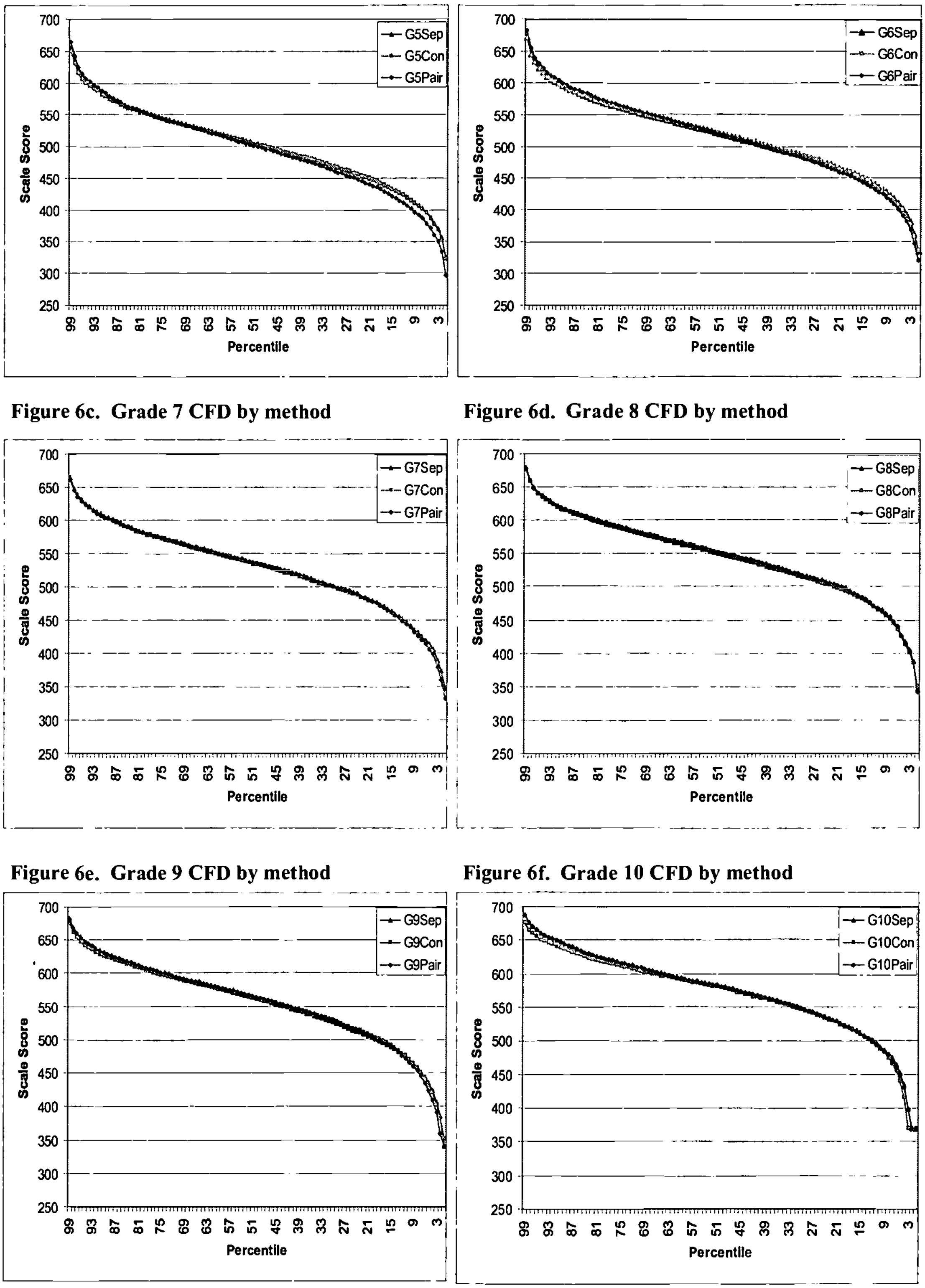


Figure 6a. Grade 5 CFD by method

Figure 6b. Grade 6 CFD by method

Figure 6c. Grade 7 CFD by method

Figure 6d. Grade 8 CFD by method

Figure 6e. Grade 9 CFD by method

Figure 6f. Grade 10 CFD by method

Item and Test Residuals

Item-score Residuals. Recall that item-score residuals were computed two ways. Residual.$_{item}$ was computed, as indicated previously, and summarized here. For a given estimation method, each complete (full test) response vector in the cross-validation sample was scored to obtain an estimated theta for that response vector. For a given item, theta and the items' parameters are applied to determine the expected score on that item, given theta. Residual.$_{item}$ is computed as the square root of the sum (over response vectors) of the squares of the difference between the observed and expected score for the item.

Residual.$_{item(unique)}$ was computed, as indicated previously, and summarized here. For a given estimation method, each partial response vector, consisting of only the responses to the unique (non-common) items, was scored to obtain an estimated theta for that response vector. For a given unique item, the estimated theta and the items' parameters are applied to determine the expected score on that item, given theta. Residual.$_{item(unique)}$ is computed as the square root of the sum (over response vectors) of the squares of the difference between the observed and expected score for the item.

The numbers of times the minimum value of residual.$_{item}$ and residual.$_{item(unique)}$ occurred for each method is shown in Table 7. Table 7 indicates that for four of the six grade the separate method had the greatest number of occurrences of the minimum value of residual.$_{item}$ and residual.$_{item(unique).}$ The minimum value of residual.$_{item}$ and residual.$_{item(unique)}$ occurred once each for the concurrent and pair-wise concurrent methods.

Table 7. Number of Times the Minimum Item Residual Occurred for Each Method

		Residual.$_{item}$				**Residual.$_{item(unique)}$**		
Grade	**Total No. of Items**	**Concurrent**	**Pair-wise**	**Separate**	**Total No. of Items**	**Concurrent**	**Pair-wise**	**Separate**
5	69	15	33*	22	44	2	44*	2
6	60	12	20	30*	15	6	0	10*
7	60	24*	16	21	20	5	10	11*
8	59	21	13	27*	20	13*	0	7
9	60	20	20	22*	15	2	7	8*
10	60	20	23	25*	35	12	12	14*

Note: The minimum value occurred for multiple methods for some grades.

* indicates the method for which the minimum value of the residual occurred most often at the given grade

Test-score Residuals. Recall that residual.$_{test}$ and residual.$_{test(unique)}$ are related in a fashion analogous to that described for residual.$_{item}$ and residual.$_{item(unique)}$. The data in Table 8 indicate that the both residual.$_{test}$ and residual.$_{test(unique)}$ are nearly identical across methods with each grade. The value of the test residuals are not comparable across grades because of the different numbers of items in each grade's assessment.

Table 8. Test-Score Residuals

Residual$_{.test}$

	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10
Concurrent	4.88	4.56	3.95	4.43	4.38	3.77
Pair-wise	4.87	4.54	3.96	4.44	4.38	3.77
Separate	4.89	4.56	3.99	4.42	4.39	3.77

Residual$_{.test(unique)}$

	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10
Concurrent	3.84	5.60	6.00	6.15	5.09	4.39
Pair-wise	3.78	5.60	6.00	6.17	5.08	4.39
Separate	3.84	5.59	6.00	6.16	5.09	4.40

Parameter consistency across the methods

The parameter estimates from the three methods were compared for consistency across methods. To allow the comparison of parameters across models (2PPC and 3PL), the discrimination and difficulty parameters calibrated under the 3PL model were placed on the metric of the constructed response items calibrated under the 2PPC metric using the following conversions:

$\alpha = 1.7a$
$\gamma = 1.7\ ab$

Table 9 provides (a) descriptive statistics for the various item parameter estimates for each method and grade and (b) comparisons of the parameter estimates between methods. The first two columns of Table 9 indicate the parameter described and number of items on the given grade's assessment that are associated with the given parameter (the numbers of multiple-choice items for the a-, b-, and c-parameters and constructed-response items for the γ- and α-parameters). Note that the properties of all the γ-parameters, that is, the gammas for all levels (γ_1, γ_2, γ_3, γ_4) of all constructed-response items in the grade are summarized together.

The next six columns provide the means and standard deviations for the parameters for each of the three estimation methods. For example, the first row of numerical entries in Table 9 indicates that the mean and standard deviation of the a-parameter for the 54 selected-response items on the 5th grade assessment are 0.022 and 0.006, respectively.

Table 9 also provides comparisons of the parameter estimates between the methods. The correlation of the parameter estimates between methods is provided first. For example, the first row of numerical entries in Table 9 indicates that the correlation between fifth grade pair-wise concurrent and concurrent a-parameters is 0.98.

Table 9. Item Parameter Comparison for the Total Test, Mathematics

								Pair-wise and Concurrent			Concurrent and Separate			Pair-wise and Separate		
	No. of	Pair-wise		Concurrent		Separate				SDR =			SDR =			SDR =
Parameters	Items	μ_p	σ_p	μ_c	σ_c	μ_s	σ_s	r	SDM	σ_p / σ_c	r	SDM	σ_c / σ_s	r	SDM	σ_p / σ_s
Grade 5																
a	54	0.022	0.006	0.024	0.007	0.023	0.007	0.98	-0.38	0.89	0.96	0.13	1.11	0.96	-0.26	0.99
b	54	9.76	3.39	11.14	3.84	10.71	3.30	0.97	-0.38	0.88	0.96	0.12	1.17	0.95	-0.28	1.03
c	54	0.213	0.069	0.206	0.104	0.212	0.083	0.80	0.08	0.67	0.71	-0.06	0.83	0.69	0.02	1.24
α	15	0.013	0.002	0.014	0.002	0.014	0.002	1.00	-0.71	0.92	0.97	-0.04	0.83	0.97	-0.68	0.76
γ	42	5.47	1.33	6.20	1.38	6.23	1.46	1.00	-0.54	0.96	0.98	-0.02	0.95	0.98	-0.54	0.91
Grade 6																
a	45	0.022	0.005	0.024	0.006	0.024	0.006	0.97	-0.36	0.85	0.96	0.05	1.05	0.98	-0.32	0.89
b	45	10.78	3.09	11.89	3.65	11.71	3.41	0.99	-0.33	0.85	0.98	0.05	1.07	0.99	-0.29	0.91
c	45	0.208	0.083	0.208	0.102	0.206	0.078	0.92	0.00	0.82	0.86	0.02	1.07	0.93	0.02	1.31
α	15	0.014	0.003	0.015	0.003	0.015	0.003	1.00	-0.34	0.90	0.99	0.03	1.03	0.99	-0.31	0.93
γ	42	6.57	2.45	7.13	2.62	7.10	2.54	1.00	-0.22	0.94	1.00	0.01	1.03	1.00	-0.21	0.97
Grade 7																
a	45	0.026	0.010	0.025	0.010	0.025	0.009	0.99	0.03	1.02	0.98	0.05	1.07	0.99	0.08	1.09
b	45	13.37	6.13	13.17	6.04	12.92	5.63	0.99	0.03	1.01	0.99	0.04	1.07	0.98	0.08	1.09
c	45	0.232	0.093	0.214	0.095	0.211	0.078	0.85	0.20	0.98	0.81	0.04	1.19	0.79	0.26	1.22
α	15	0.017	0.007	0.017	0.007	0.017	0.006	1	0.01	1.02	1	0.01	1.04	0.99	0.02	1.06
γ	42	9.36	4.07	9.32	4.01	9.22	3.86	1.00	0.01	1.02	1.00	0.02	1.04	1.00	0.03	1.06

P=Pair-wise, C=Concurrent, S=Separate, r=correlation coefficient, SDM=standardized difference in means, SDR=standard deviation ratio

Table 9. Item Parameter Comparison for the Total Test, Mathematics (Continued)

								Pair-wise and Concurrent			Concurrent and Separate			Pair-wise and Separate		
	No. of	Pair-wise		Concurrent		Separate				SDR =			SDR =			SDR =
Parameters	Items	μ_p	σ_p	μ_c	σ_c	μ_s	σ_s	r	SDM	σ_p / σ_c	r	SDM	σ_c / σ_s	r	SDM	σ_p / σ_s
Grade 8																
a	44	0.027	0.011	0.027	0.011	0.027	0.011	0.99	0.05	1.01	0.98	0.00	1.02	1.00	0.06	1.03
b	44	14.97	6.67	14.64	6.68	14.56	6.43	0.99	0.05	1.00	0.98	0.01	1.04	0.99	0.06	1.04
c	44	0.201	0.088	0.193	0.104	0.199	0.072	0.88	0.08	0.84	0.71	-0.07	1.23	0.81	0.02	1.46
α	15	0.019	0.008	0.019	0.008	0.019	0.007	0.99	-0.02	1.01	0.99	0.02	1.04	1.00	0.00	1.05
γ	42	10.70	4.52	10.70	4.54	10.70	4.30	0.99	0.00	0.99	0.99	0.00	1.06	1.00	0.00	1.05
Grade 9																
a	45	0.029	0.009	0.030	0.009	0.028	0.009	0.98	-0.18	0.95	0.96	0.27	1.07	0.97	0.09	1.02
b	45	16.53	5.90	17.41	6.14	15.95	5.74	0.99	-0.15	0.96	0.99	0.25	1.07	0.99	0.10	1.03
c	45	0.228	0.086	0.217	0.101	0.205	0.084	0.84	0.11	0.85	0.81	0.13	1.03	0.94	0.27	1.21
α	15	0.020	0.006	0.020	0.006	0.020	0.005	0.98	-0.08	0.97	0.97	0.11	1.13	0.99	0.02	1.10
γ	42	12.10	3.41	12.42	3.44	12.06	3.17	0.98	-0.09	0.99	0.98	0.11	1.08	0.99	0.01	1.08
Grade 10																
a	45	0.030	0.009	0.034	0.010	0.030	0.009	0.99	-0.36	0.89	0.97	0.38	1.15	0.98	0.02	1.02
b	45	17.83	5.63	19.78	6.33	17.70	5.49	1.00	-0.33	0.89	0.98	0.35	1.15	0.99	0.02	1.02
c	45	0.230	0.097	0.247	0.104	0.228	0.100	0.97	-0.17	0.93	0.93	0.18	0.97	0.96	0.01	1.04
α	15	0.021	0.008	0.022	0.009	0.020	0.008	1.00	-0.20	0.87	0.99	0.23	1.13	1.00	0.03	0.98
γ	42	13.63	5.04	14.56	5.68	13.51	5.06	1.00	-0.17	0.89	0.99	0.19	1.12	1.00	0.02	1.00

P=Pair-wise, C=Concurrent, S=Separate, r=correlation coefficient, SDM=standardized difference in means, SDR=standard deviation ratio

Table 9. Item Parameter Comparison for the Total Test, Mathematics (Continued)

								Pair-wise and Concurrent			Concurrent and Separate			Pair-wise and Separate		
	No. of	Pair-wise		Concurrent		Separate				SDR =			SDR =			SDR =
Parameters	Items	μ_p	σ_p	μ_c	σ_c	μ_s	σ_s	r	SDM	σ_p / σ_c	r	SDM	σ_c / σ_s	r	SDM	σ_p / σ_s
a	46	0.026	0.008	0.027	0.009	0.026	0.008	0.98	-0.20	0.94	0.97	0.15	1.08	0.98	-0.06	1.01
b	46	13.87	5.13	14.67	5.45	13.93	5.00	0.99	-0.18	0.93	0.98	0.14	1.10	0.98	-0.05	1.02
c	46	0.219	0.086	0.214	0.102	0.210	0.082	0.88	0.05	0.85	0.81	0.04	1.05	0.85	0.10	1.25
α	15	0.017	0.006	0.018	0.006	0.018	0.005	1.00	-0.22	0.95	0.99	0.06	1.03	0.99	-0.15	0.98
γ	42	9.64	3.47	10.05	3.61	9.80	3.40	0.99	-0.17	0.96	0.99	0.05	1.05	0.99	-0.11	1.01

P=Pair-wise, C=Concurrent, S=Separate, r=correlation coefficient, SDM=standardized difference in means, SDR=standard deviation ratio

The second comparison provides the standardized difference between the means (SDM) for each pair of methods in pooled standard deviation units, that is, $SDM_{m_1m_2} = \frac{\mu_{m_1} - \mu_{m_2}}{\sqrt{\frac{\sigma^2_{m_1} + \sigma^2_{m_2}}{2}}}$, where m_1 is estimation method 1 and m_2 is estimation method 2. For example, the SDM_{pc} for the fifth grade pair-wise and concurrent a-parameters is 0.38.

The third comparison, the standard deviation ratio (SDR), is the ratio of the standard deviation of method 1 to that of method 2, that is, $SDR_{m_1m_2} = \frac{\sigma_{m_1}}{\sigma_{m_2}}$, where m_1 is estimation method 1 and m_2 is estimation method 2. For example, the SDR_{pc} for the fifth grade pair-wise and concurrent a-parameters is 0.89.

The descriptive and comparative statistics are provided for the a-, b-, c-, α-, and γ-parameters for each grade. The last set of entries in Table 9 provides the average (across grades) for each of the statistics and comparisons described above.

The results in Table 9 indicate that the correlations were high for the item difficulty (b for selected-response and γ for constructed-response items) and discrimination (a for selected-response and α for constructed-response) parameters; the correlations for the difficulty parameters ranged from 0.95 to 1.0 and the correlations for the discrimination parameters ranged from 0.96 to 1.0 across all method-pairs. The correlations for the pseudo-guessing parameter were generally lower than for the other parameters, ranging from 0.69 to 0.97. The pseudo-guessing correlations were generally highest between the pair-wise concurrent and concurrent methods. The average correlation across grades for the a-, b-, α-, and γ- parameter was 0.97 or higher for each comparison. The average correlation for the c-parameter ranged from 0.81 (concurrent and separate) to 0.88 (pair-wise and concurrent).

An examination of the standardized difference in means (SDM) for each method pair indicates that the parameter estimates for the a-, b-, α-, and γ- parameters were similar for the three methods in grades 7 and 8, with SDM values below 0.1 for each method pair in these grades. In grades 5 and 6, the SDM for the a-, b-, α-, and γ- parameters were relatively large for the pair-wise and concurrent comparison and for the pair-wise and separate comparison but not for the concurrent and separate comparison. The SDM for the a-, and α-parameters ranged in absolute value from 0.26 to 0.71 for the comparisons involving the pair-wise method in grades 5 and 6. The SDM for the b-, and γ-parameters ranged in absolute value from 0.21 to 0.54 for the comparisons involving the pair-wise method in grades 5 and 6.

In grades 9 and 10, the SDM for the a-, b-, α-, and γ- parameters tended to be larger for the comparisons involving the concurrent method than for the pair-wise and separate comparison than for the pair-wise and separate comparison.

The c-parameter estimates tended to be more stable across methods than the other parameters, with most of the SDM below 0.20 and a maximum SDM of 0.27.

The standard deviation ratios (SDR) produced results similar to that of the SDM in that the SDR for grades 7 and 8 indicated that variability of the parameter estimates in these grades were similar across method.

Scatter plots comparing parameter estimates from the different estimation methods are provided in Appendix B.

Discussion

The present study compared and analyzed three estimation and linking methods for establishing vertical scales. Data were sampled from an operational mathematics testing program utilizing a common item design in grades 5 through 10. The first method, separate estimation, is accomplished in two steps: (1) separate calibration and (2) grade-by-grade chained linking. The second method, concurrent estimation, establishes the vertical scale in a single step—the calibration phase—by simultaneously estimating parameters for all items at all grades. The third method, pair-wise concurrent estimation, is a hybrid of methods 1 and 2. In the first step, concurrent estimation is employed for items in non-overlapping pairs of adjacent grades to establish a common scale for each adjacent grades pair. In the second step, the scale for one adjacent grades pair is identified as the base scale and the remaining adjacent grades pairs are placed on the base grade via the common items via chained linking as in the separate calibration method.

Standard analyses of calibration output indicated that the separate estimation method produced consistently better results than did the concurrent or pair-wise concurrent estimation methods in terms of convergence of items, model fit, and differential item functioning analyses. Only one item did not converge—an item common to grades 8 and 9. It is notable that the item converged in the grades 7-8 pair-wise estimation but not for the grade 8 separate estimation because the grades 7-8 pair-wise estimation did not contribute any additional data on the item itself than existed for the grade 8 separate estimation. This indicates that the construct for grade 8 was altered by the simultaneous calibration with grade 7 items. The item characteristic curves (ICCs) and observed scores for this item for the grade 7-8 pair-wise concurrent estimation and for the grade 8 separate estimation (using the parameters from the last iteration) are provided in Appendix C.

The separate method resulted in better overall fit than did the other two methods. The mean fit statistic (z) was lower for the separate estimation than for the other two methods in each grade. In grades six through 10 the pair-wise estimation method had a lower mean fit statistic than did the concurrent method, but the opposite was true in grade 5. The separate method had only 9 items flagged for possible misfit and the pair-wise concurrent and concurrent methods had 13 and 14 items, respectively. Between 77% and 79% of the items flagged for misfit for each of the three methods were constructed-response items. Thus, it seems that the separate method results in better fit, as one would expect if some multidimensionality were present across grades, but none of the three methods seemed differentially more or less robust to item fit with regard to item format (selected- versus constructed-response).

Differential item functioning analyses indicated a similar pattern of results for the three methods—the separate method resulted in fewer items flagged for differential item functioning. The pair-wise concurrent and concurrent methods flagged about twice as many items for DIF than did the separate method. There were 11, 21, and 24 items flagged for DIF under the separate, pair-wise concurrent, and concurrent methods, respectively. This is not unexpected, given the pattern of model fit results. In this case, 64%, 62%, and 50% of the items flagged for DIF were constructed-response items for the separate, pair-wise concurrent, and concurrent methods, respectively. This indicates that (a) greater proportions of items flagged for DIF are selected-response items regardless of method than were flagged for model fit and that (b) the concurrent method differentially flags greater proportions of selected-response items for DIF than the other two methods. One might expect the proportions of selected- and constructed-response items flagged for DIF to be similar for model fit and differential item functioning (because DIF is essentially model misfit assessed by focal group). This may indicate that multidimensionality may be manifest differentially by focal group with regard to item format. The efficacy of scaling selected- and constructed-response items together using traditional IRT models is based on the assumption of unidimensionality. It has been argued by some that selected- and constructed-response items should be scaled separately; others argue the robustness of the IRT models and estimation methods to modest transgressions of dimensionality. Although the results of one study are not sufficient to generalize, it should be noted that their may be an interaction of item type and focal group with regard to issues of model misspecification that may be important to consider when identifying an estimation method for vertical scaling.

The properties of the parameter and ability estimates were compared for the three estimation methods using several criteria. With regard to ability estimates, the methods produced similar aggregate results except at the end grades. In grade 5, the pair-wise method tended to produce lower scale scores at and below the 50th percentile. In grade 10, the concurrent method tended to produce lower scale scores above the 50th percentile. In terms of grade-to-grade growth measured by the change in the mean scale score from grade-to-grade, this resulted in (a) more growth from grade 5 to grade 6 for the pair-wise concurrent method, (b) less growth from grade 9 to grade 10 for the concurrent method, and (c) similar growth from grades 6 to 7, 7 to 8, and grade 8 to grade 9 for the three methods. The growth trends in the current study depart somewhat from what might be expected—previous research might have lead one to expect a pattern of decreasing grade-to-grade growth, with means better fitting a quadratic curve than the more linear growth trends observed in the present study (Kolen and Brennan, in progress). Given that the differences are not large—the mean score for the pair-wise method is 4 less than occurs for the other two methods in grade 5 and the mean score for the concurrent method is 4 or 5 less than for the other two methods in grade 10—the differences in ability estimates do not indicate a significant preference for one method over another.

The pattern of grade-to-grade variability was very similar for the concurrent and separate methods but was somewhat different for the pair-wise method, where there was greater variability in grades 5 and 6 than for the other 4 grades under any of the methods. Grades 5 and 6 were calibrated concurrently under the pair-wise concurrent method and linked to the base scale (the grade 7-8 adjacent grade pair) using the items common to grades 6 and 7. Initial attempts at investigating the cause of the differences in variability did not provide a rationale for the apparent anomaly, and further investigation is warranted and will continue.

The pattern of grade-to-grade variability was generally flat for the concurrent and separate methods. This is not unexpected; Kolen and Brennan (in preparation) indicate that previous research has shown trends of increasing, decreasing, and homogenous variances for different tests under different scaling methods.

Differences in the separation of grade distributions between the methods was unremarkable.

The results of the item and test residuals indicated minor differences overall in the ability of the item parameters to predicted performance on the items. Although the results of the item residuals supported the efficacy of the separate method over the other two methods, the test residual results did not reinforce this.

The comparison of item parameter estimates between the methods reflected the differences observed previously in terms of the ability estimates. The parameter comparisons indicated that the estimates for the pair-wise method diverged from the other two in grades 5 and 6. This was mirrored by (a) standard deviations for the pair-wise method that were quite different than occurred for the other two methods and (b) lower scale scores at the 10th, 25th, and 50th percentiles for the pair-wise method than occurred for the other two methods. Also, the comparisons indicated that the grade 10 estimates for the concurrent method diverged somewhat from those of the other two methods; this divergence is reflected by lower scale scores at the 75th and 90th percentiles for the concurrent method than occurred for the other two methods.

The tendency for the parameter estimates to be most similar at grades 7 and 8 and to diverge in the lower and upper grades is likely due to the fact that the three methods were linked through the grade 7 parameters. That is, in order to facilitate comparison, the parameter estimates from the separate and pair-wise methods were transformed via the Stocking and Lord procedure (1983) using the grade 7 concurrent item parameters as the anchor set. One would differences between the parameters across method would be expected to be less near the anchor grade (7) and increase with distance from the anchor grade. In the current study, the estimation and linking methods are confounded. Other approaches were possible, such as using the concurrent parameter estimates for all items in all grades as the anchor set.

Our expectations to *not* find large differences in the results of the separate or concurrent estimation methods were only partially fulfilled. The differences between these two methods for model fit and differential item functioning were non-trivial. However, there were relatively minor differences between the two methods with regard to patterns of grade-to-grade growth and variability, and separation of grade distributions. The correlations between the parameters for the two methods also tended to be high.

The expectation that the hybrid-nature of the pair-wise concurrent method would result in the pair-wise method producing results that were more similar to the other two methods than they were to each other was also only partially supported. The pair-wise method fell between the separate and concurrent methods in terms of model fit and differential item functioning analyses and tended to be more similar to the concurrent method than the separate method in this regard. However, the pair-wise method was the outlier with regard to its pattern of grade-to-grade variability. In terms of grade-to-grade growth, the separate method tended to be more similar to the other two methods than they were to each other. The separate method produced results that were similar to at least one of the other methods at all grades, whereas the pair-wise mean diverged from the other two in grade 5 and the concurrent mean diverged in grade 10. The parameter estimate comparisons mirrored these observations.

Given that the issues of model-data fit are central to the validity of the inferences made from test scores, these differences may be the most salient. The results of one study are not sufficient to generalize, however, if further research indicates the tendency for greater misfit for concurrent estimation methods with regard to the general population and relevant subgroups, then the separate method may be preferred over the concurrent method.

References

Béguin, A. A., Hanson, B. A., & Glass, C. A. W. (2000). *Effect of multidimensionality on separate and concurrent estimation in IRT equating.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, L.A.

Burket, G. (2002). PARDUX [Computer program]. Unpublished. Monterey, CA: CTB McGraw-Hill.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, Vol. 26 No. 1, 3-24.

Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, Vol. 26 No. 1, 25-41.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under Item response theory. *Applied Psychological Measurement*, Vol. 22No. 2 131-143.

Kolen, M. J. & Brennan, R. L. (in preparation). Test equating: Methods and practices (2nd Ed.). New York: Springer-Verlag.

Linn, R. L., & Harnisch, D. (1981). Interaction between item content and group membership in achievement test items. *Journal of Educational Measurement*, 18, 109-118.

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

Muraki, E. (1992). A generalized partial credit model: Application of an M algorithm. *Applied Psychological Measurement, 16*, 159-176.

Patz, R. J., & Hanson, B. A. (2002). *Psychometric issues in vertical scaling.* Paper presented at the annual meetings of the National Council on Measurement in Education, New Orleans, LA.

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

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika*, 50, 399-410.

Appendix A

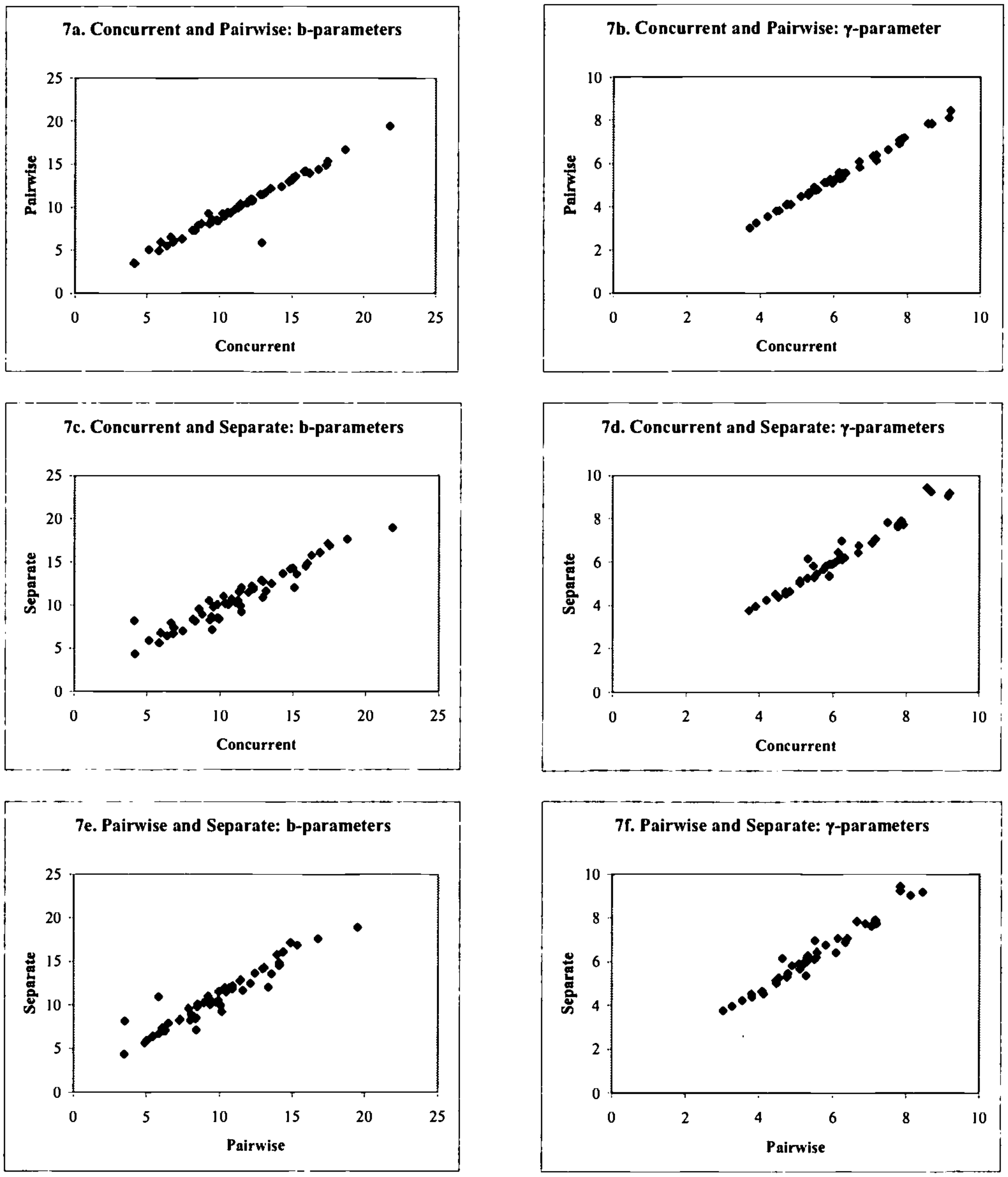
Transformation Constants Used to Link the Separate and Pair-Wise Methods to the base (Concurrent) Scale

	Multiplicative (M1) and additive (M2) constants to link the initial estimation results to the base scale for each method. [(0,1) metric]		Multiplicative (M1) and Additive (M2) constants to link the initial vertical scales to the concurrent Scale		Multiplicative (M1) and Additive (M2) constants to transform the concurrent scale to an arbitrary operational (working) scale	
Separate Method: Base scale = Grade 7						
Grade	**$M1^C$**	**$M2^C$**	**$M1^* =$ 0.96873 x $M1^C$**	**$M2^* =$ 0.96873 x $M2^C$ - 0.19931**	**M1= 60 x $M1^*$**	**M2 = 60 x $M2^*$ + 550**
Grade 5	1.02	-0.55	0.99	-0.73	59.33	505.99
Grade 6	1.04	-0.29	1.01	-0.48	60.35	521.08
Grade 7	1.00	0.00	0.97	-0.20	58.12	538.04
Grade 8	0.97	0.25	0.94	0.04	56.62	552.42
Grade 9	0.98	0.41	0.95	0.20	57.13	561.90
Grade 10	0.94	0.69	0.91	0.47	54.64	577.92
Pair-wise Method: Base scale = Grade 7-8 Adjacent Grade Pair						
Grade	**$M1^C$**	**$M2^C$**	**$M1^*$ = 0.92211 x $M1^C$**	**$M2^* =$ 0.92211 x $M2^C$ + 0.08352**	**M1 = 60 x $M1^*$**	**M2 = 60 x $M2^*$ + 550**
Grade 5	1.09	-0.44	1.01	-0.32	60.47	530.78
Grade 6	1.09	-0.44	1.01	-0.32	60.47	530.78
Grade 7	1.00	0.00	0.92	0.08	55.33	555.01
Grade 8	1.00	0.00	0.92	0.08	55.33	555.01
Grade 9	1.01	0.46	0.93	0.51	55.69	580.52
Grade 10	1.01	0.46	0.93	0.51	55.69	580.52

Appendix B

Scatter Plots: Comparing Parameters for the Estimation Methods

Figure 7. Grade 5: Scatter Plots of Difficulty Parameters by Method-Pairs



7a. Concurrent and Pairwise: b-parameters

7b. Concurrent and Pairwise: γ-parameter

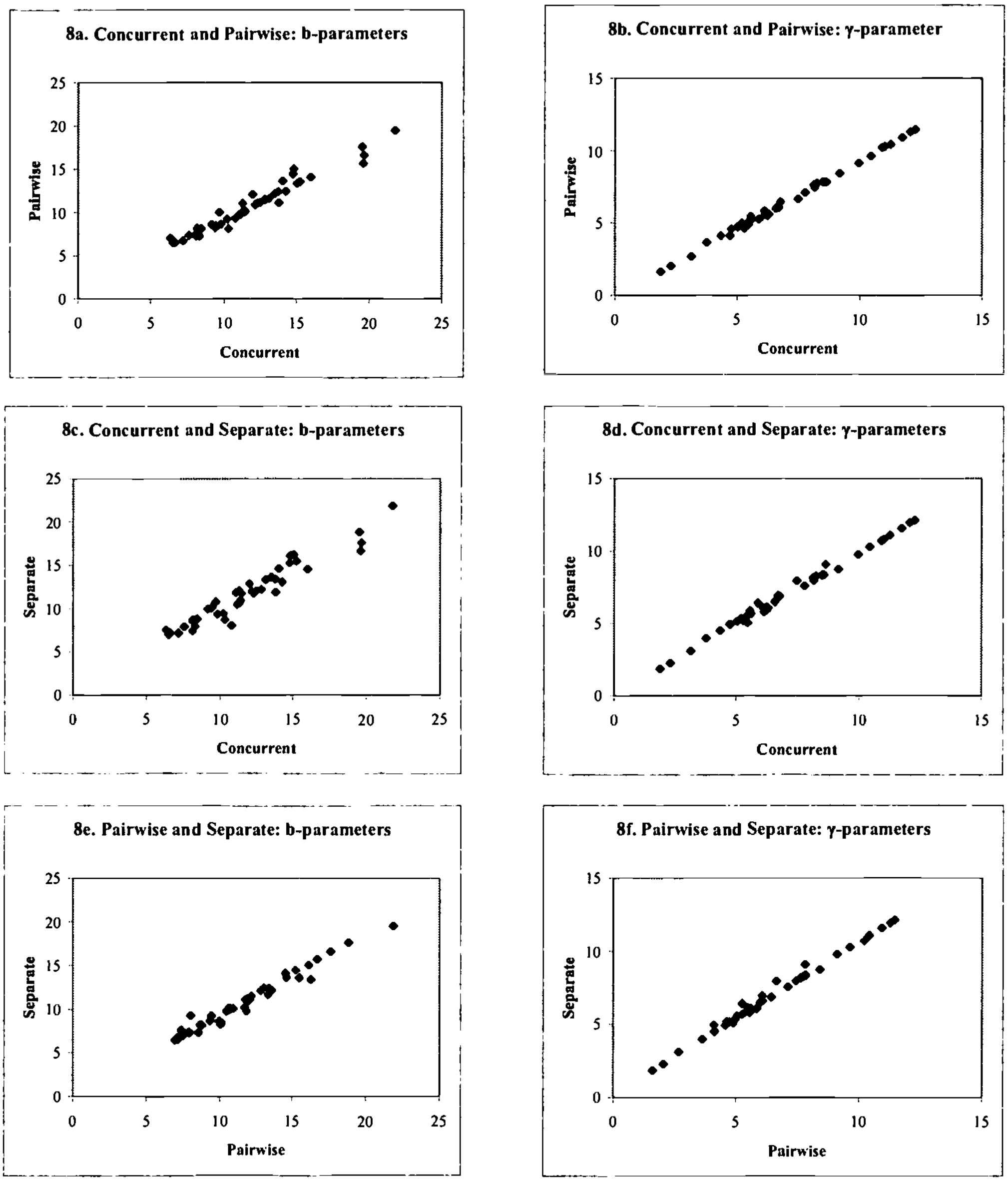
7c. Concurrent and Separate: b-parameters

7d. Concurrent and Separate: γ-parameters

7e. Pairwise and Separate: b-parameters

7f. Pairwise and Separate: γ-parameters

Figure 8. Grade 6: Scatter Plots of Difficulty Parameters by Method-Pairs



8a. Concurrent and Pairwise: b-parameters

8b. Concurrent and Pairwise: γ-parameter

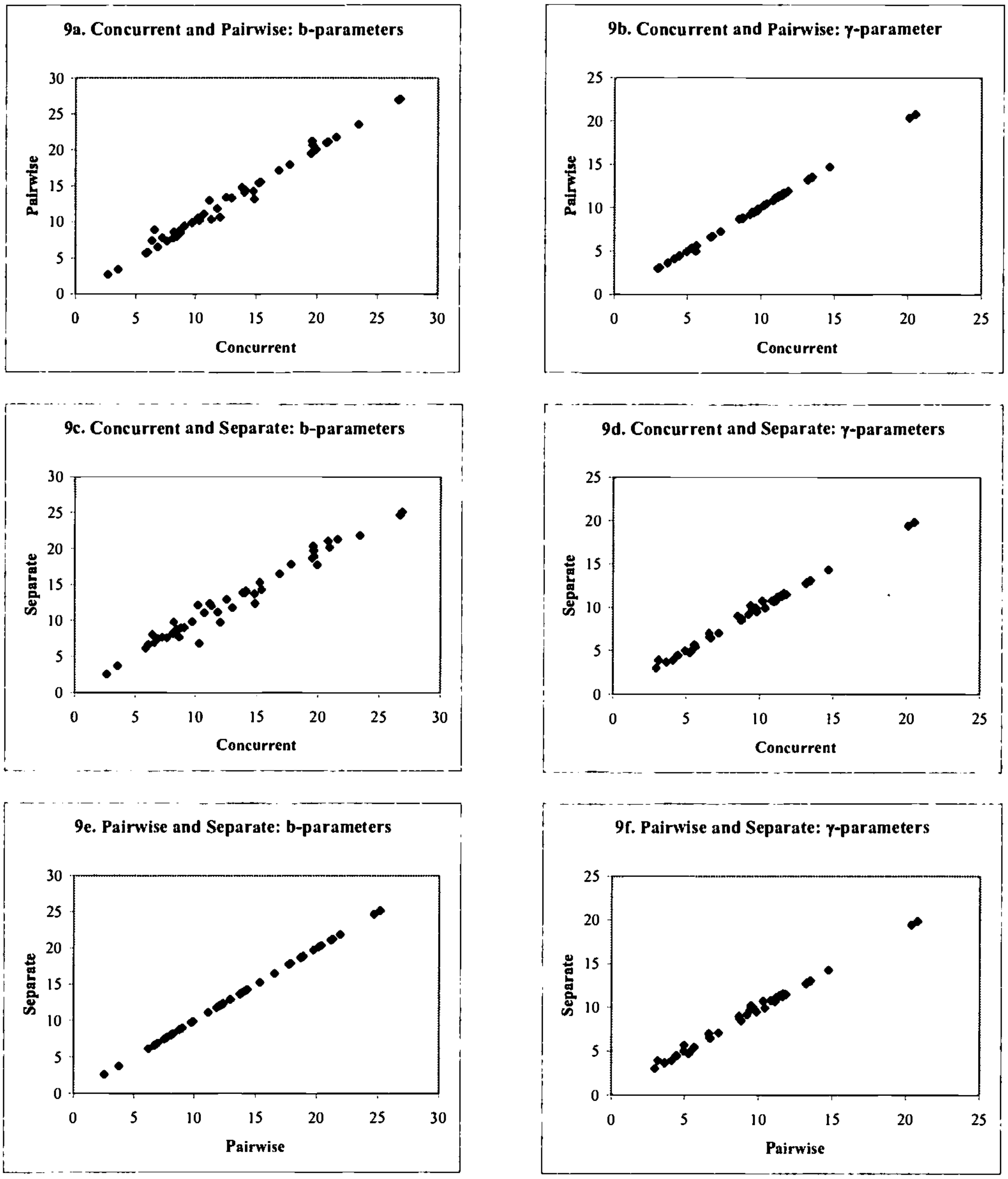
8c. Concurrent and Separate: b-parameters

8d. Concurrent and Separate: γ-parameters

8e. Pairwise and Separate: b-parameters

8f. Pairwise and Separate: γ-parameters

Figure 9. Grade 7: Scatter Plots of Difficulty Parameters by Method-Pairs



9a. Concurrent and Pairwise: b-parameters

9b. Concurrent and Pairwise: γ-parameter

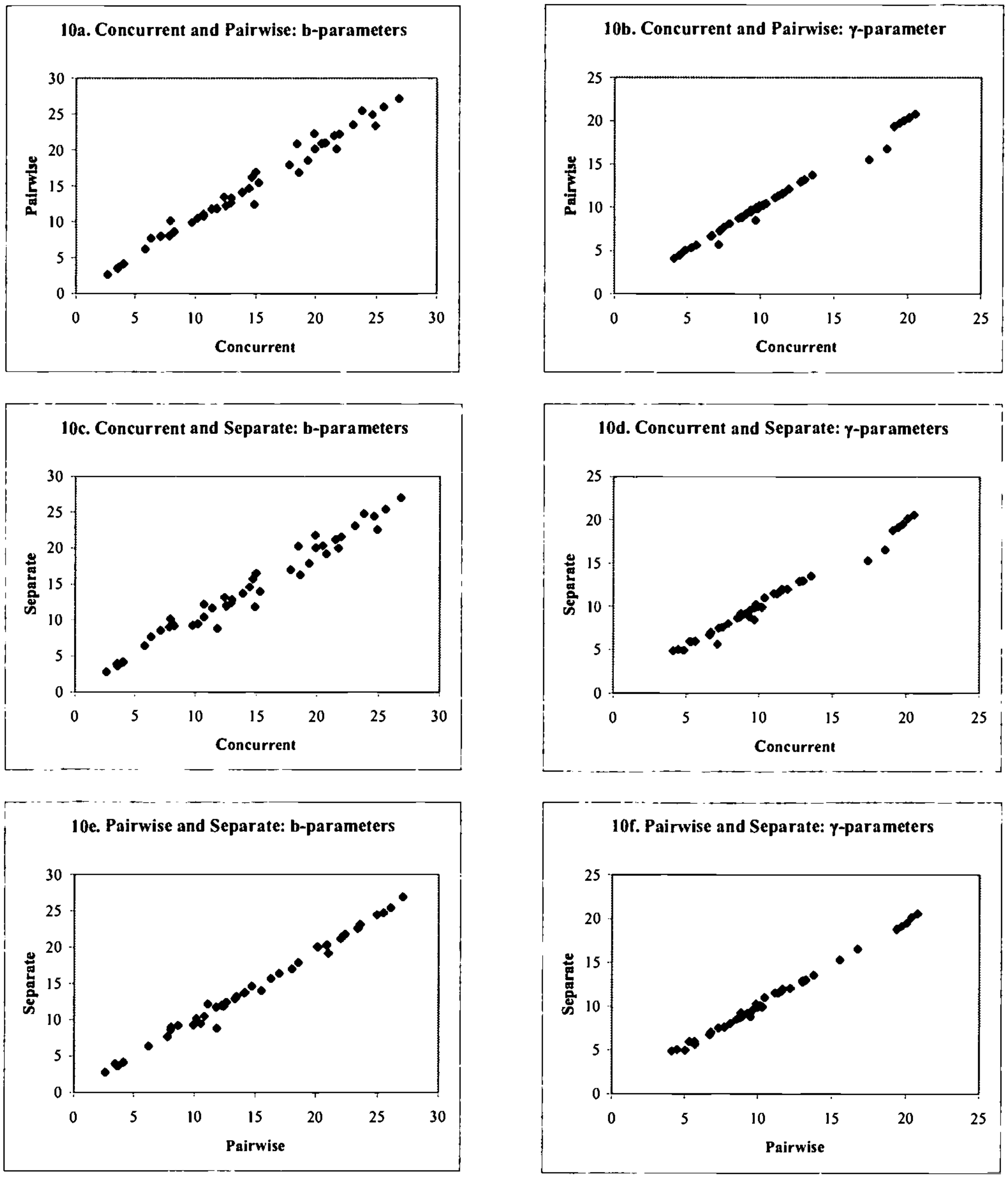
9c. Concurrent and Separate: b-parameters

9d. Concurrent and Separate: γ-parameters

9e. Pairwise and Separate: b-parameters

9f. Pairwise and Separate: γ-parameters

Figure 10. Grade 8: Scatter Plots of Difficulty Parameters by Method-Pairs



10a. Concurrent and Pairwise: b-parameters

10b. Concurrent and Pairwise: γ-parameter

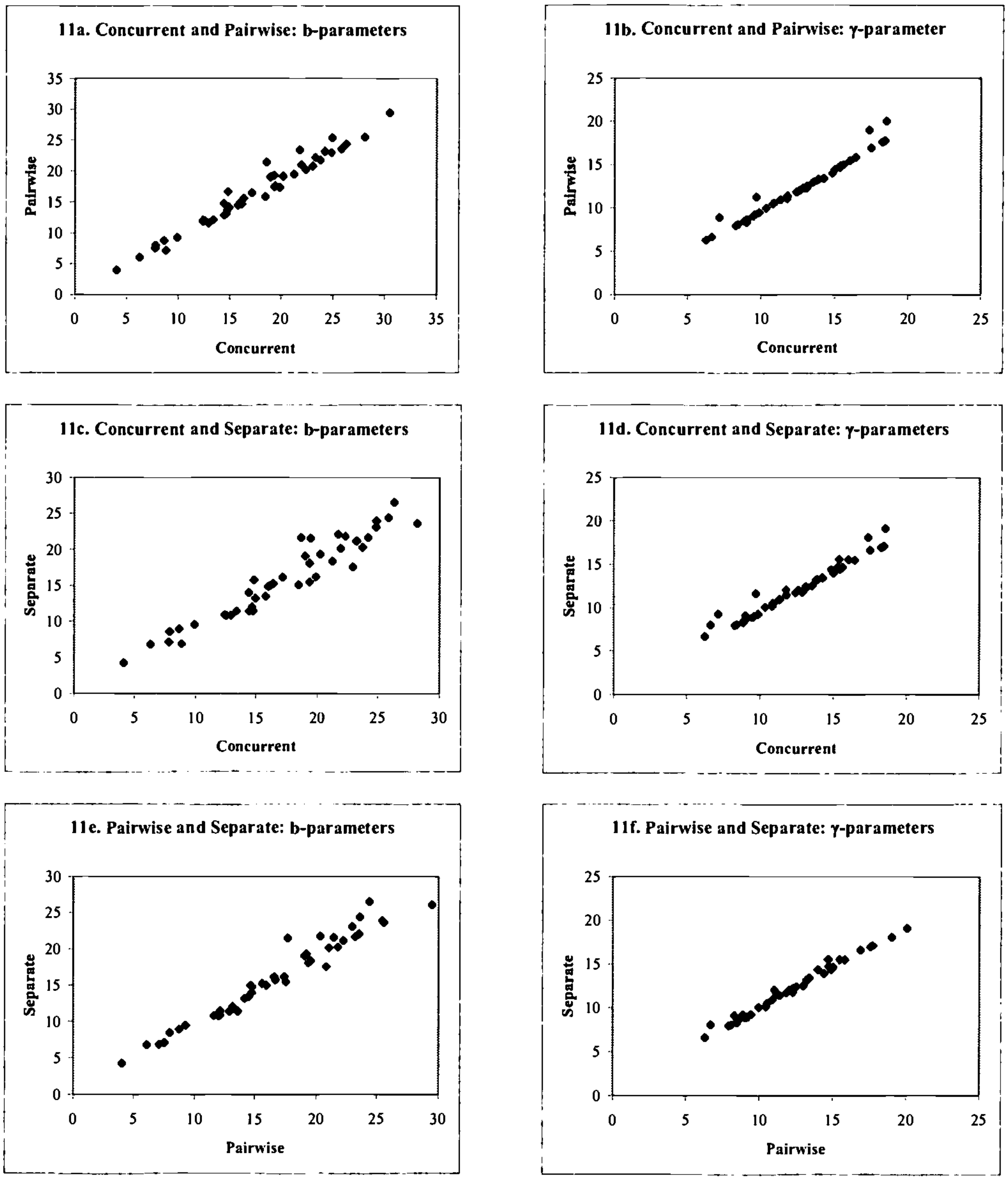
10c. Concurrent and Separate: b-parameters

10d. Concurrent and Separate: γ-parameters

10e. Pairwise and Separate: b-parameters

10f. Pairwise and Separate: γ-parameters

Figure 11. Grade 9: Scatter Plots of Difficulty Parameters by Method-Pairs



11a. Concurrent and Pairwise: b-parameters

11b. Concurrent and Pairwise: γ-parameter

11c. Concurrent and Separate: b-parameters

11d. Concurrent and Separate: γ-parameters

11e. Pairwise and Separate: b-parameters

11f. Pairwise and Separate: γ-parameters

Figure 12. Grade 10: Scatter Plots of Difficulty Parameters by Method-Pairs

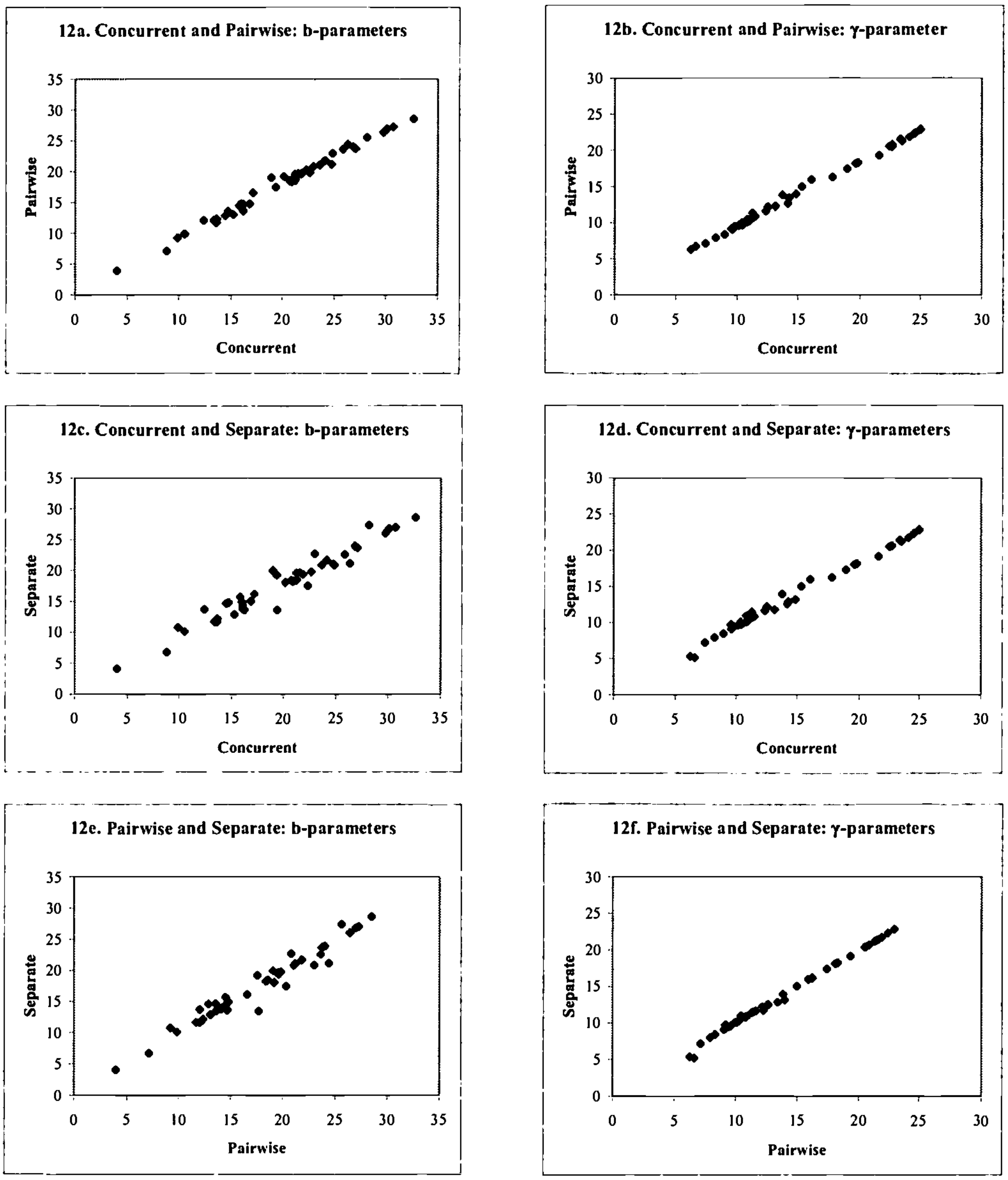
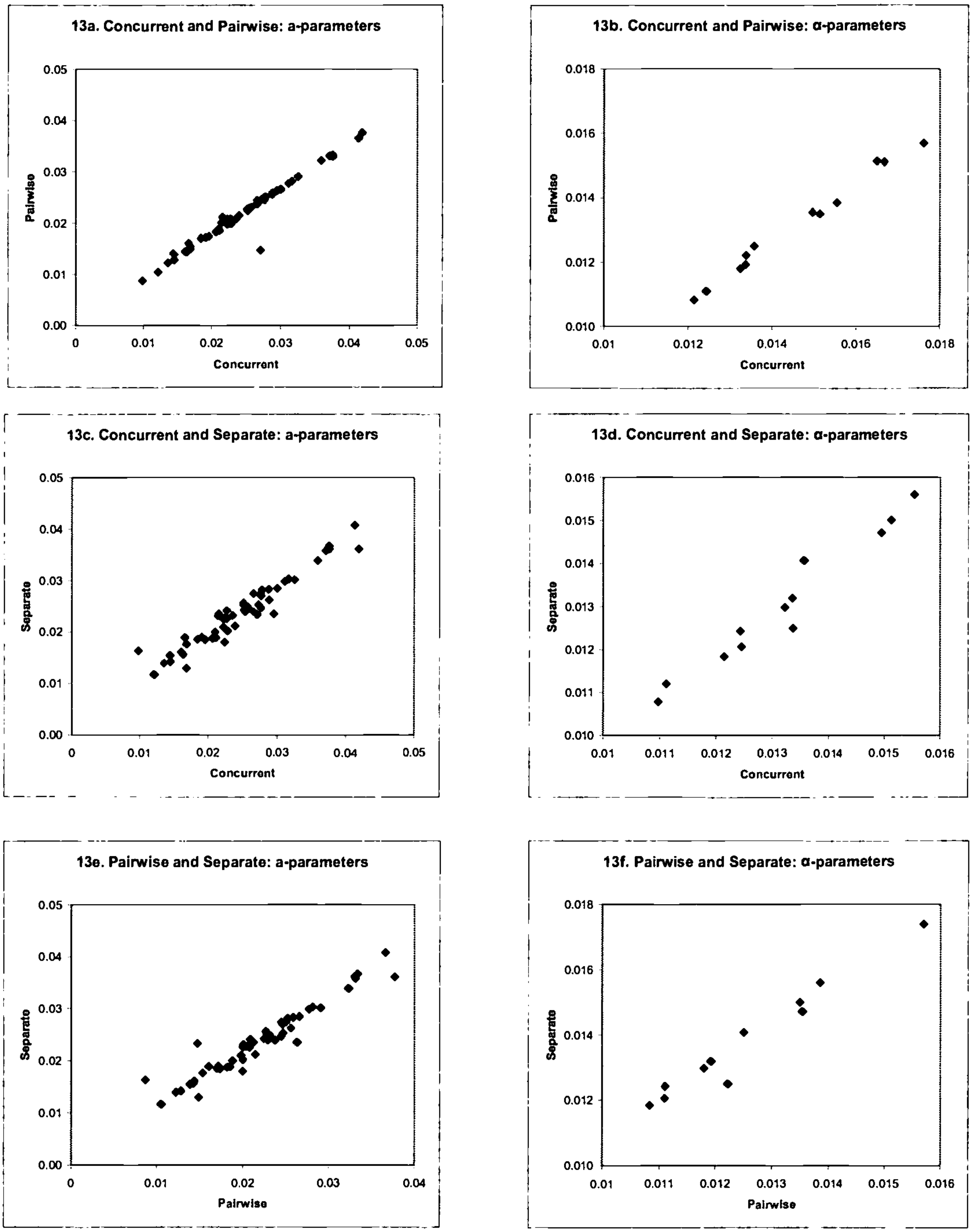


Figure 13. Grade 5: Scatter Plots of Discrimination Parameters by Method-Pairs



13a. Concurrent and Pairwise: a-parameters

13b. Concurrent and Pairwise: α-parameters

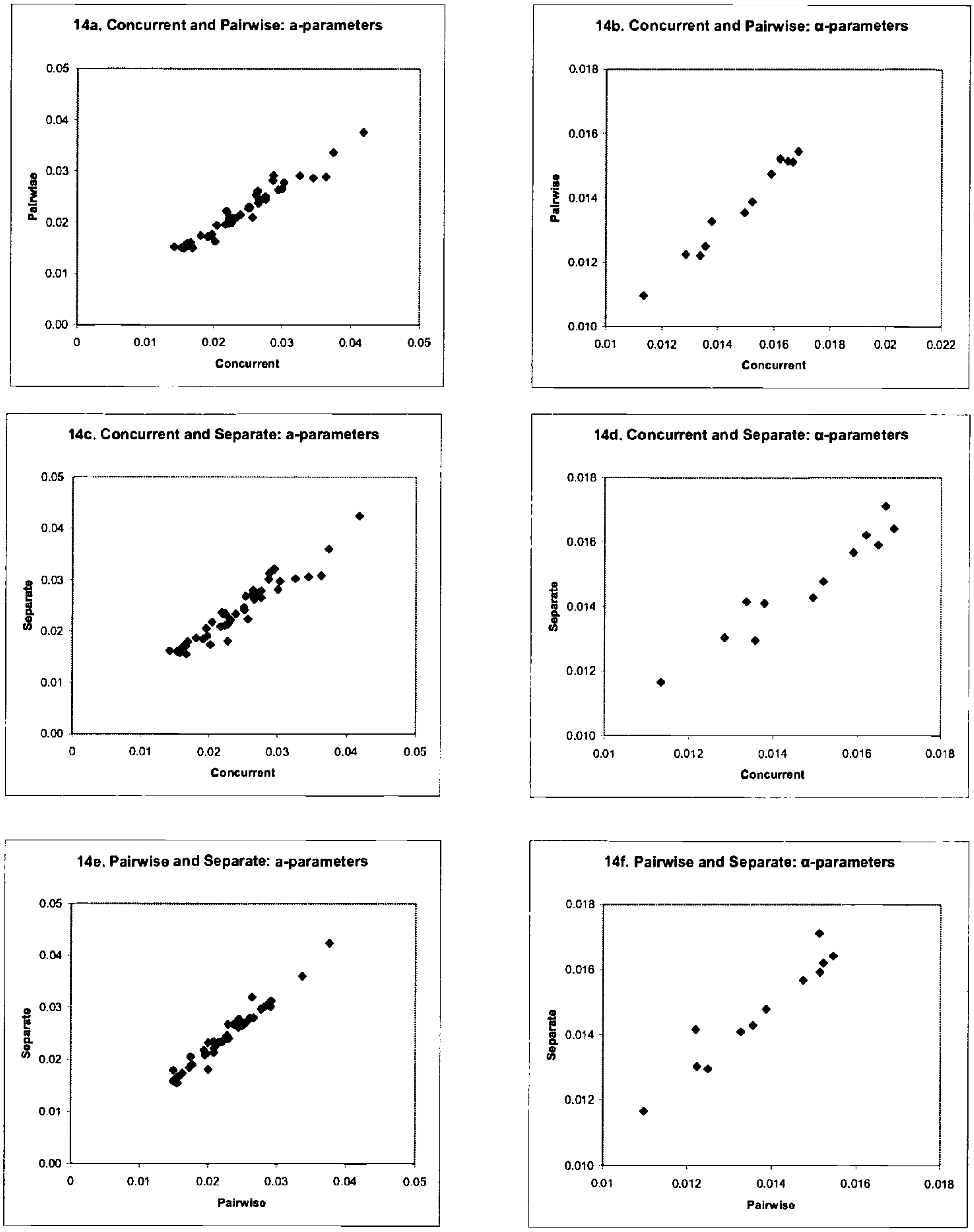
13c. Concurrent and Separate: a-parameters

13d. Concurrent and Separate: α-parameters

13e. Pairwise and Separate: a-parameters

13f. Pairwise and Separate: α-parameters

Figure 14. Grade 6: Scatter Plots of Discrimination Parameters by Method-Pairs



14a. Concurrent and Pairwise: a-parameters

14b. Concurrent and Pairwise: α-parameters

14c. Concurrent and Separate: a-parameters

14d. Concurrent and Separate: α-parameters

14e. Pairwise and Separate: a-parameters

14f. Pairwise and Separate: α-parameters

Figure 15. Grade 7: Scatter Plots of Discrimination Parameters by Method-Pairs

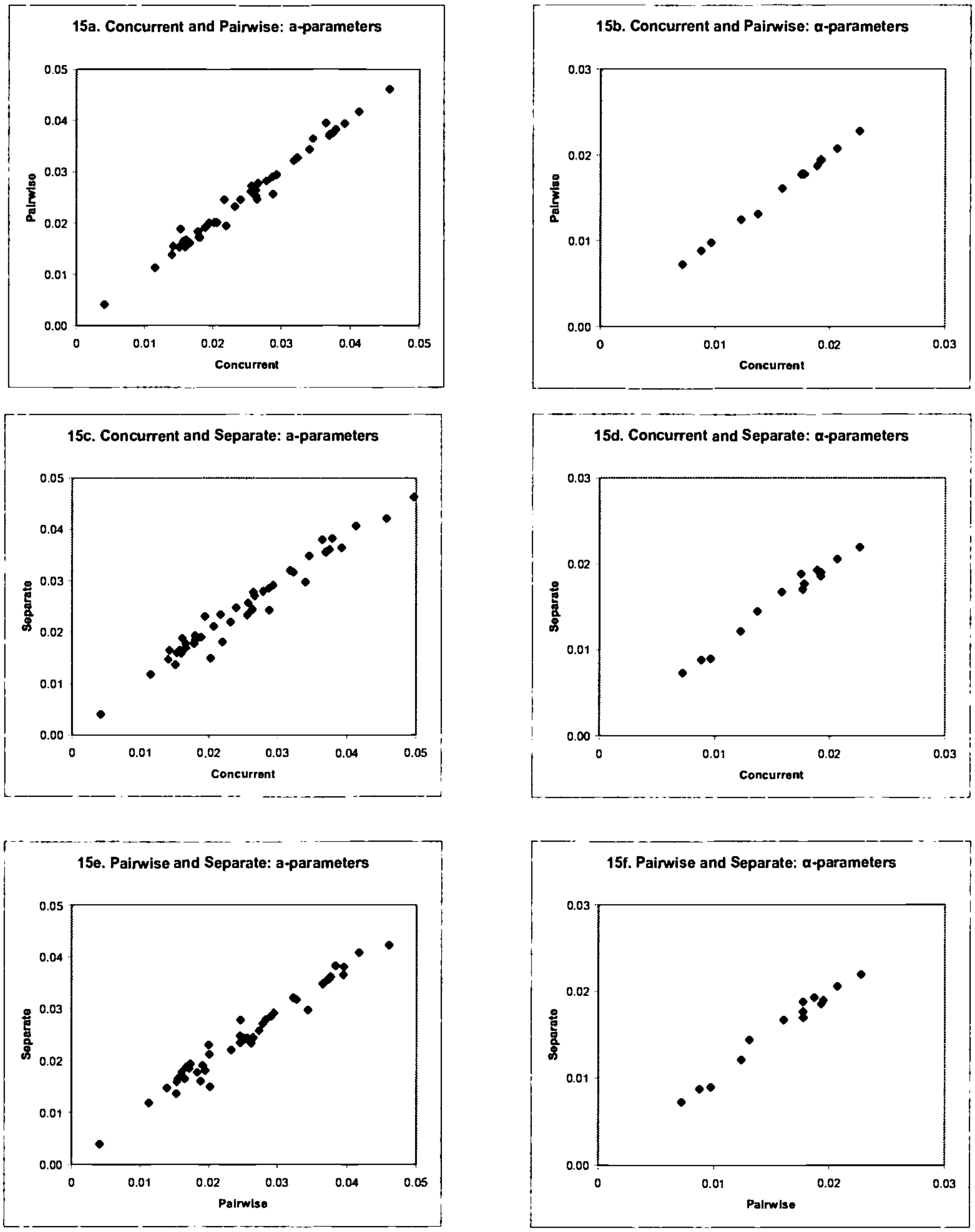
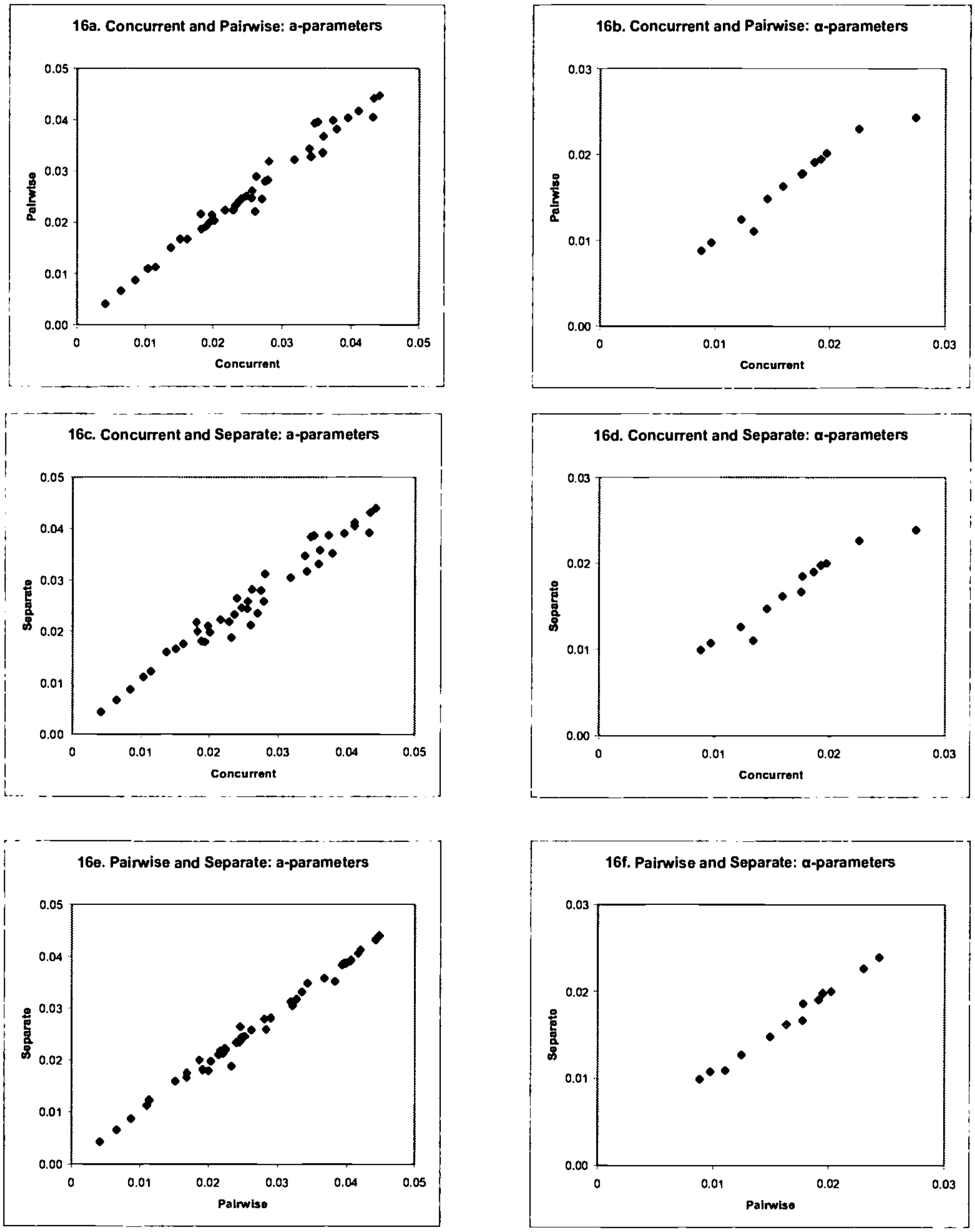


Figure 16. Grade 8: Scatter Plots of Discrimination Parameters by Method-Pairs



16a. Concurrent and Pairwise: a-parameters

16b. Concurrent and Pairwise: α-parameters

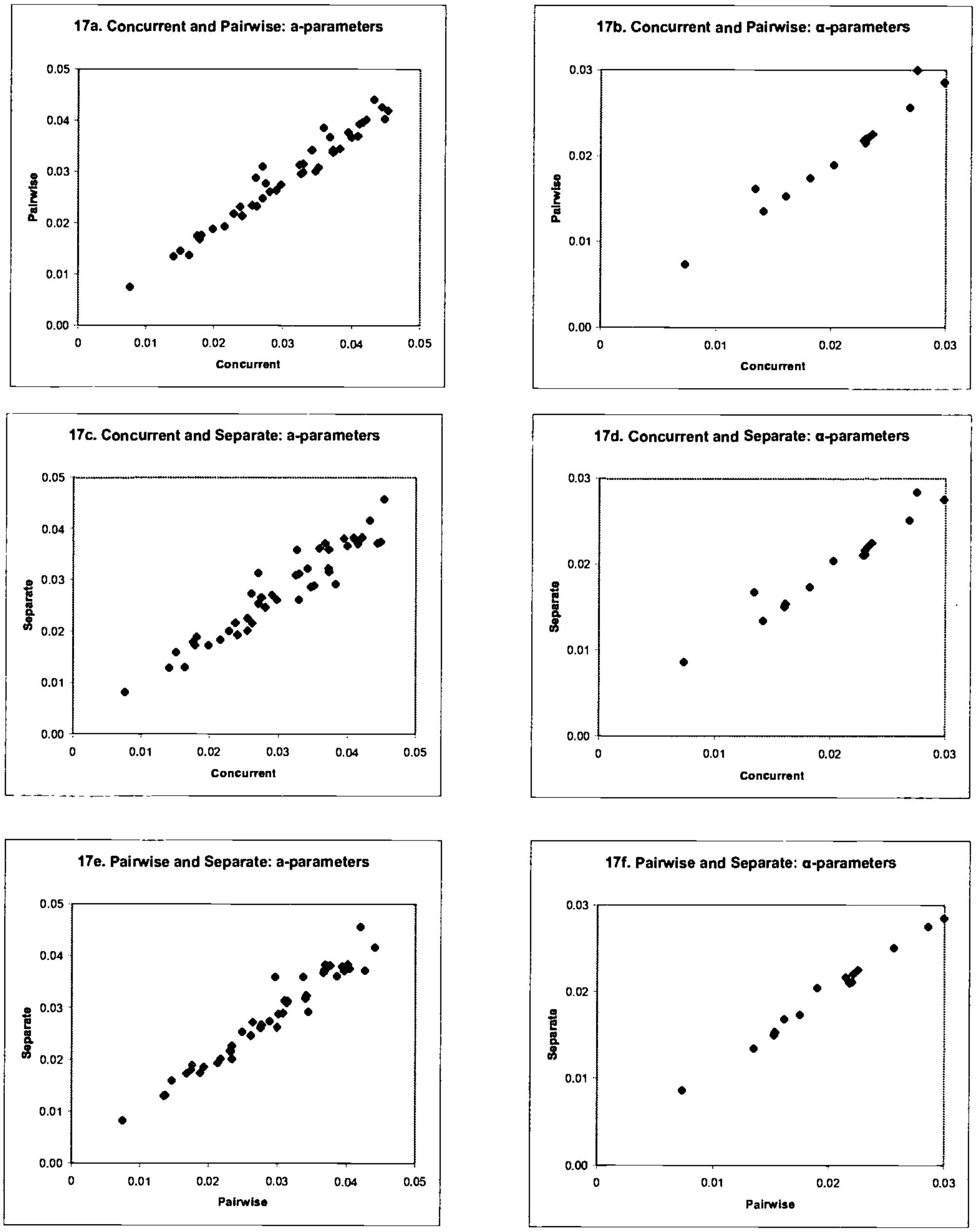
16c. Concurrent and Separate: a-parameters

16d. Concurrent and Separate: α-parameters

16e. Pairwise and Separate: a-parameters

16f. Pairwise and Separate: α-parameters

Figure 17. Grade 9: Scatter Plots of Discrimination Parameters by Method-Pairs



17a. Concurrent and Pairwise: a-parameters

17b. Concurrent and Pairwise: α-parameters

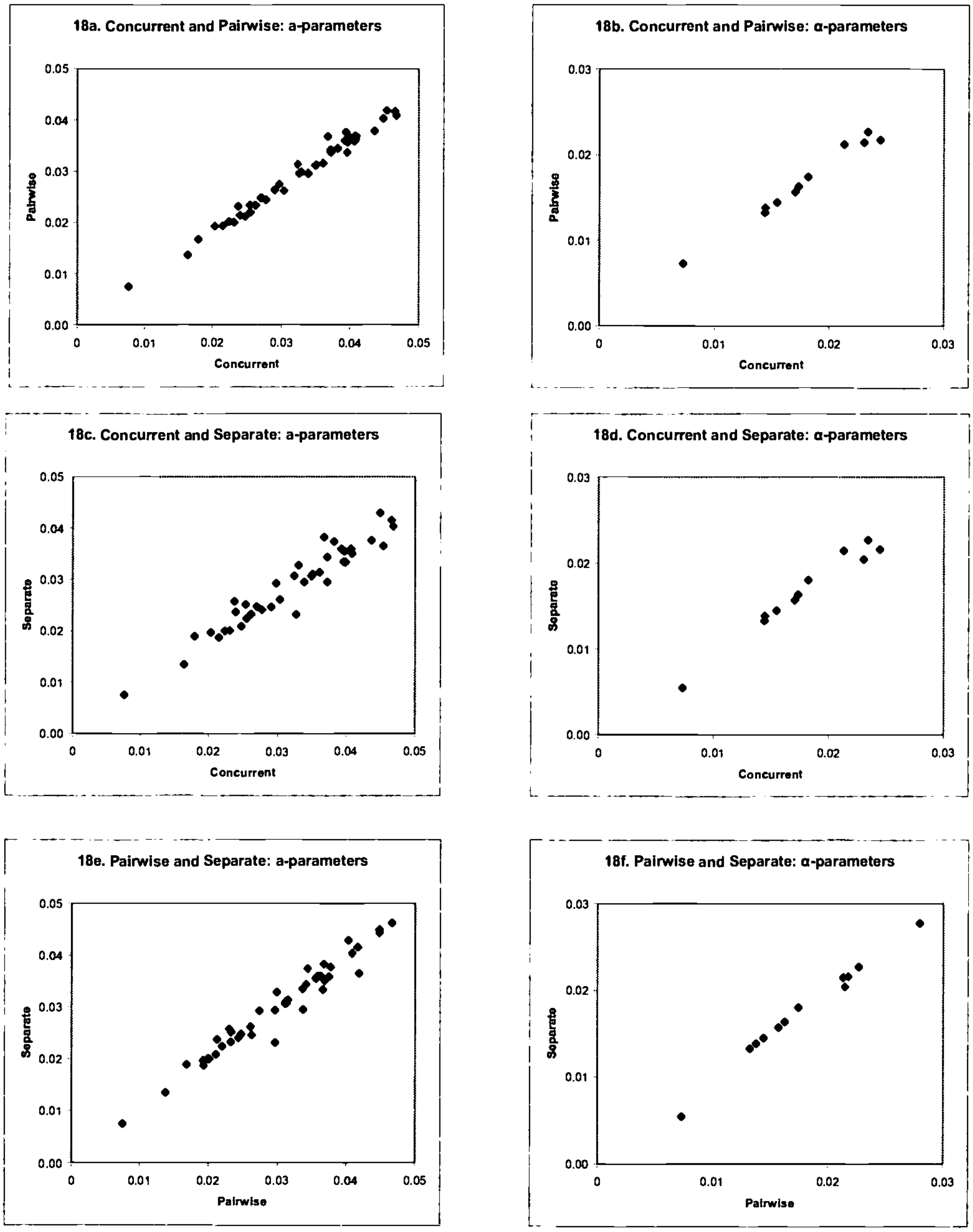
17c. Concurrent and Separate: a-parameters

17d. Concurrent and Separate: α-parameters

17e. Pairwise and Separate: a-parameters

17f. Pairwise and Separate: α-parameters

Figure 18. Grade 10: Scatter Plots of Discrimination Parameters by Method-Pairs



18a. Concurrent and Pairwise: a-parameters

18b. Concurrent and Pairwise: α-parameters

18c. Concurrent and Separate: a-parameters

18d. Concurrent and Separate: α-parameters

18e. Pairwise and Separate: a-parameters

18f. Pairwise and Separate: α-parameters

Appendix C

Figure 19. Item Characteristic Curves (ICCs) for the Non-Converging Item

Figure 19a. Pair-wise Concurrent Method: Grades 7 and 8

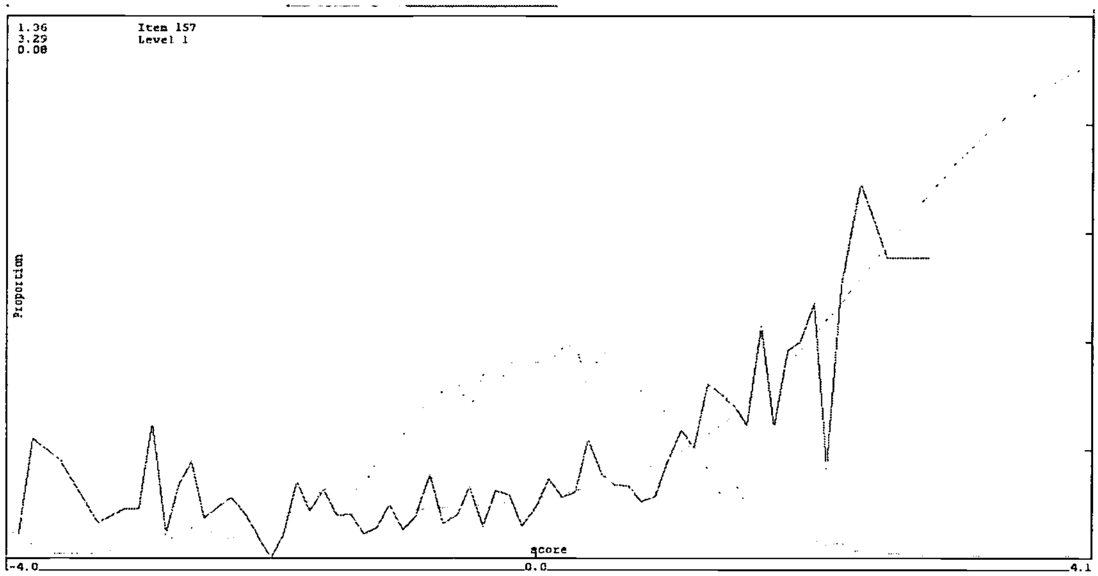
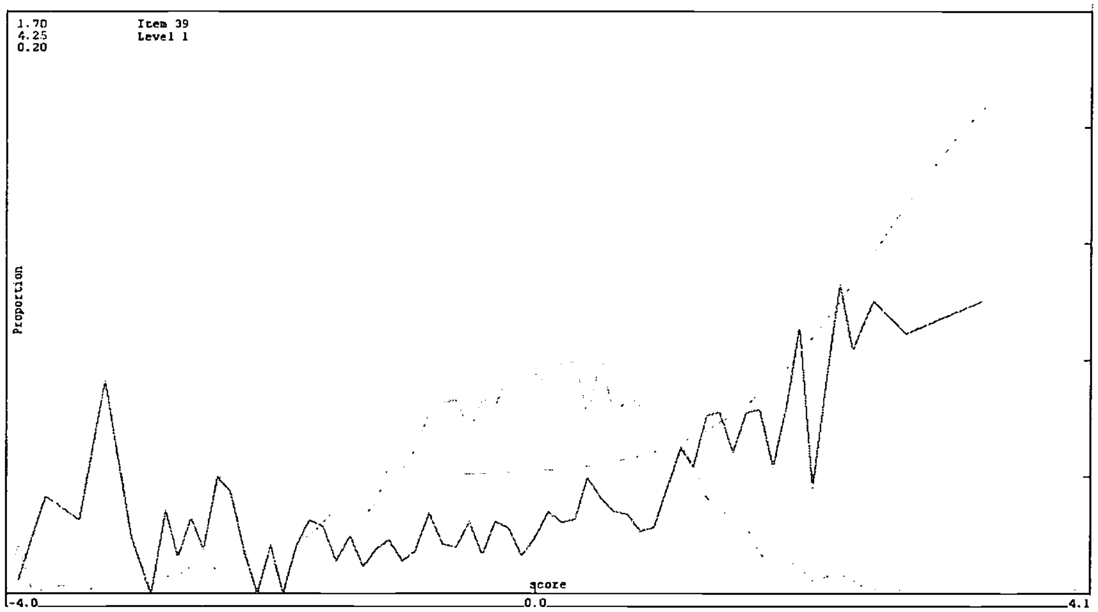


Figure 19b. Separate Method: Grade 8



III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org