The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Albany, New York 12234

Summary of New York State Test Equating Procedures: 2002 – 2005

Rationale for Equating

This document provides an overview of the procedures employed to equate New York State tests from 2002 to 2005. Although the same procedures apply to each grade and content area, examples provided in this document are from Mathematics grade 4 because that test has recently received focused attention.

The equating procedures used in the New York State testing program comply with standards for scales, norms, and score comparability as outlined in Chapter 4 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, 1999).

For test security and validity reasons, alternate test forms containing different questions are administered to New York State students each year. If the identical set of questions were administered across years, then students might be able to prepare for those specific questions and obtain scores that are higher than their actual achievement level. The alternate test forms cover the same content and are designed to be similar in terms of test difficulty and other technical characteristics.

A common and widely-accepted statistical process called test equating is employed to ensure that students would be expected to get the same reported score regardless of which alternate test form they took. “Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably” (Kolen & Brennan, 2004, p.2). Equating procedures are employed in almost all testing programs that use more than one test form. Although test forms in the same testing program are developed using the same content and technical specifications, they are similar but not entirely identical in difficulty. Equating adjusts for small differences in difficulty among the test forms. This is accomplished by placing alternate test forms onto a common score scale for score reporting purposes. This score scale is used over multiple years so that reported scores can be compared across years and across alternate forms. Equating adjustments were made each year for each New York State test form.

A reported score (also called a scale score) is different from a raw score. A raw score is simply the number of points obtained on the test by a student; that is, the number of multiple choice questions answered correctly plus the number of points earned on open-ended items. Scale scores derived from the equating process are designed to accurately reflect student's achievement level regardless of which test form was taken, whereas raw scores reflect performance only on the particular test form taken and do not generalize to other test forms. This is precisely why equating is performed and scale scores are reported.

Item Response Theory Models

Psychometric (statistical) models are employed in testing programs for various purposes such as to evaluate items and test forms, assemble tests, report scores, and equate alternate test forms. A common family of psychometric models called Item Response Theory (IRT) is used in New York State. The overall model is based on a mathematical functional relationship between students' ability and the probability that the students will correctly answer the item.

IRT item statistics (parameters) have the advantage over classical item statistics (such as percent of students that answered the item correctly, called a "p-value") in that they generally are not dependent upon student ability. For example, a low item p-value could result from the item being truly difficult or because only low-ability students took the item (and hence scored poorly on it). Similarly, a high item p-value could result from either the item being less difficult or because only high-ability students took the item. IRT, on the other hand, gets purer measures of item difficulty and student ability by disentangling their effects on item performance. IRT takes into account the fact that not all items provide the same amount of information in determining how much a student knows or can do. Computer programs that implement IRT models use actual student data to estimate the characteristics of the items on a test, called "parameters." The parameter estimation process is called "item calibration." IRT allows comparisons among items and student scale scores, even those from different test forms, by using a common scale for all items and students.

IRT models differ relative to the number of parameters estimated. For the New York State tests, the three-parameter logistic model (Lord & Novick, 1968; Lord, 1980) was used to calibrate multiple choice items, and the two-parameter partial credit model (Muraki, 1992; Yen, 1993) was used to calibrate constructed response items.

In the three-parameter logistic (3PL) model, the probability that a student with ability θ responds correctly to item i is

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\theta - b_i)]} ,$$

where a_i is the item discrimination, b_i is the item difficulty, and c_i is the probability of a correct response by a very low-scoring student. The discrimination parameter is an index of how well an item differentiates between high-performing and low-performing students. An item that cannot be answered correctly by low-performing students, but can be answered correctly by high-performing students, will have a high discrimination value. The difficulty parameter is an index of how easy or difficult an item is. The higher the difficulty parameter is, the harder the item. The guessing parameter is the probability that a student with very low ability will answer the item correctly.

The two-parameter partial credit 2PPC model is a special case of Bock's (1972) nominal model. The 2PPC or "generalized" partial credit model states that the probability of an examinee with ability θ having a score (k - 1) at the k-th level of the j-th item is

$$P_{jk}(\theta) = P(x_j = k-1 \mid \theta) = \frac{\exp Z_{jk}}{\sum_{i=1}^{m_j} \exp Z_{ji}} , \; k = 1 \ldots m_j ,$$

where

$$Z_{jk} = A_{jk}\theta + C_{jk} .$$

The m_j denotes the number of score levels for the j-th item, and typically the highest score level is assigned $(m_j – 1)$ score points. For the special case of the 2PPC model used here, the following constraints were used:

$$A_{jk} = \alpha_j (k-1),$$

and

$$C_{jk} = -\sum_{i=0}^{k-1} \gamma_{ji},$$

where $\gamma_{j0} = 0$, and α_j and γ_{ji} are free parameters to be estimated from the data. Each item has $(m_j –1)$ independent γ_{ji} parameters and one α_j parameter; a total of m_j parameters are estimated for each item.

The IRT model parameters were estimated using PARDUX software specifically designed for this purpose. The software estimates parameters simultaneously for MC and CR items using marginal maximum likelihood procedures implemented via the EM (expectation-maximization) algorithm (Bock & Aitkin, 1981; Thissen, 1982). Simulation studies have compared PARDUX with MULTILOG (Thissen, 1991), PARSCALE (Muraki & Bock, 1991), and BIGSTEPS (Wright & Linacre, 1992). PARSCALE, MULTILOG, and BIGSTEPS are among the most widely known and used IRT programs. PARDUX was found to perform at least as well as these other programs (Fitzpatrick, 1990; Fitzpatrick, 1994; Fitzpatrick and Julian, 1996).

Field Test Designs

In September 2001, two grade 4 Mathematics field test forms (called Form B and Form C) were administered to representative samples of grade 5 students. Grade 5 students were administered grade 4 forms because the field test administration was very early in the school year. These forms were constructed to be similar to the operational test form in regard to test content and format. The 2000 operational test form (called Form A) was also administered to a representative sample of students. The three forms were spiraled (administered in alternating fashion) at the classroom level with each student taking either operational (anchor) form or one of the field test forms.

In 2002, three Grade 4 Mathematics field test forms (Form A, Form B, and Form C) were administered to three representative samples of Grade 4 New York State (NYS) students. Each form contained both multiple-choice and constructed-response items and was matched to the operational test form in regard to test content and format. The three field test forms were spiraled at the classroom level and administered within the 2002 operational test testing window. Each student took one field test form, and the field test data for these students were matched to their 2002 operational data. The field test forms were separately equated to the 2002 operational form using the multiple choice items contained in 2002 operational form as the anchor set.

In 2003, the Grade 4 Mathematics field test consisted of 23 short field test forms administered to 23 representative samples of Grade 4 students in a census field testing. Six of these forms consisted of 15 multiple-choice items each, and the remaining forms contained either 3 or 4 constructed- response items each. Sets of field test forms were spiraled at the classroom level. The 2003 field test forms were administered within the 2003 operational testing window. The field test data were matched to 2003 operational data at the student level. The multiple choice items contained in 2003 operational test served as anchors in the 2003 field test equating.

Table 1 depicts the composition of the samples selected to take the three 2002 field test forms and is illustrative of the samples obtained for the 2001 and 2003 field test forms. The samples were chosen to approximately represent the Grade 4 students in New York State. Additional subsamples acquired to assess differential item functioning (DIF) included 11 additional schools. The data from the DIF subsamples were added in order to provide an adequate number of Asian American students for the DIF analyses. The ethnicity data were recorded by the teachers on the answer documents for each student.

DIF analyses were conducted to flag items that did not behave the same in different ethnic and gender groups of students, after controlling for student ability. If answering the item requires skills or knowledge that are not intended to be measured by the test, then the possibility of DIF is increased, and inclusion of the item in the operational test might adversely affect the validity of the intended test score interpretation. Items flagged for statistically significant DIF at the field test stage were closely examined for potential bias by multiple content reviewers. Based on the reviews, items that appeared to be problematic (e.g., potentially biased) were not selected to appear in the operational test. If no reason could be generated as to why the item was flagged statistically, then the item was eligible for operational item selection if needed. Three statistical DIF methods were employed at the field test stage: standardized mean difference, Mantel-Haenszel (Zwick, Donoghue, & Grima1993), and Linn-Harnisch IRT method (Linn & Harnisch, 1981). As a further check and confirmation, DIF analyses were also conducted on the items when they appeared in the operational forms.

Table 1
Composition of the Acquired 2002 Field Test Samples

	Form A	Form B	Form C	Totals
Representative	1750	1737	1709	5196
DIF	472	452	437	1361
Totals	2222	2189	2146	6557
	Form A	Form B	Form C	Totals
Section 504	39	45	22	106
IEP	208	168	150	526
	Form A	Form B	Form C	Totals
White	789	760	754	2303
African Am.	227	241	253	721
Hispanic	275	286	267	828
Asian Am.	178	203	169	550
Missing	753	699	703	2155
Total	2222	2189	2146	6557

Field Test Items and Operational Test Forms

The 2002, 2003, 2004 and 2005 operational test forms were built using items from the pool of items that have parameters on the same (operational) score scale. The 2002 operational test form was built entirely from items field tested in 2001 and equated to the 2000 operational test form. The 2003 operational test form was built using items field tested in 2001 and 2002. Items field tested in 2002 were equated to the 2002 operational test form. The 2004 and 2005 forms were developed using items field tested in 2001, 2002 and 2003. The 2003 field test items were equated to the 2003 operational test. A summary of these relationships between field test items and operational forms is provided in Table 2a and Table 2b.

Table 2a
Operational Forms and Field Test Items

Year Operational Form was Administered	Year that Items on the Operational Form were Field Tested
2002	2001
2003	2001, 2002
2004	2001, 2002, 2003
2005	2001, 2002, 2003

Table 2b
Equating of Field Test Items to Operational Forms

Year that the Field Test Items were Administered as Field Test Items	Administration Year of Operational Form used to Equate Field Test Items
2001	2000
2002	2002
2003	2003

All equated field test items were evaluated in terms of the following psychometric properties:

- Convergence status – An item was flagged if parameters could not be estimated for it in the calibration process (i.e., that did not converge). These items were excluded from the item pool.
- Item fit – a statistical index indicating the appropriateness of using the 3PL or 2PPC model for the item; misfit items were flagged and further investigated
- Item difficulty – also called p-value. Items with very high (greater than 0.90) or very low (less than 0.30) p-values were flagged and further investigated.
- Item discrimination – a statistical index (called point biserial correlation) of how well the item differentiates high- and low-scoring students. Multiple choice items with low (less

than 0.15) point-biserial correlation for the correct answer or positive point biserial correlation on one or more distractors were flagged and further investigated.

- Omission rates – The rate at which students did not respond to the item. Items with omission rates higher than 5% were flagged.
- Differential Item Functioning – item flagged by a statistical DIF method described in Field Test Design section were flagged and reviewed by content experts.

Some items flagged based on these psychometrical properties were still included in the pool and were selected for operational use. It was not always possible to avoid including flagged items in an operational test form because of the size of the item pool and the need to match the test content blueprint. The content experts minimized the number of flagged items on operational test forms.

Equating Procedures

From 2002 to 2005, an IRT pre-equating design was used. In IRT pre-equating, the new items or test forms can be equated to existing test forms and placed onto the operational scale immediately following field testing (prior to operational administration). The item parameters for the field test items are obtained by an item calibration and then placed onto the operational scale by equating to an existing operational test form. These parameters are then on the same scale as the operational tests. In the pre-equating design, a large pool of calibrated and equated field test items is maintained and the new test forms are built by sampling items from this pool. No additional or separate equating of the operationally administered test forms is required when this design is employed.

Field test items administered in 2001, 2002, and 2003 were placed onto the operational score scale and equated to operational test forms using a set of anchor items. Anchor items were selected to be representative of the total test content.

All test items in the 2000 operational test form were used as anchors to equate the 2001 field test. In 2002 and 2003 operational multiple-choice items representative of the test content were used as anchors to equate the 2002 and 2003 field test items to the operational scale. IRT parameters for the anchor items were already on the NYS operational scale. This method of anchoring field test items to the operational scale allows for the final scoring tables to be produced before the operational test administration to meet required score-reporting timelines.

Placing new field test items onto the New York State scale was performed using a commonly-used IRT (test) characteristic curve method by Stocking & Lord (1983). Characteristic curve methods find the linear transformation ($M1_{New}$ and $M2_{New}$) that transforms the original item parameter estimates to the scale score metric and minimizes the difference between the relationship between raw scores and ability estimates (i.e., the test characteristic curve or TCC) defined by the field test form anchor item parameter estimates and that relationship defined by operational form anchor item parameter estimates. This places the transformed parameters for the new field test items onto the New York State operational scale.

The description of the entire Stocking and Lord procedure is mathematically complex, however, the relationships between the new and old linear transformation constants that are applied to the original ability metric parameters to place them onto the NYS scale are presented below to assist in understanding:

$M1_{New} = A^* M1_{Old}$

$M2_{New =} A^* M2_{Old} + B$

where $M1_{New}$ and $M2_{New}$ are the new linear transformation constants from the Stocking & Lord (1983) procedure calculated to place the new field test items onto the NYS scale. $M1_{Old}$ and $M2_{Old}$ are the transformation constants previously used to place the anchor item parameter estimates onto the NYS scale.

The A and B values are derived from the input (old) and estimate (new) values of anchor items. Anchor input or 'old' values are known item parameter estimates entered into equating. Anchor estimate or 'new' values are parameter estimates for the same anchor items re-estimated during the equating procedure. The input and estimate anchor parameter estimates are expected to have similar values. The A and B constants are computed as follows:

$$A = \frac{SD_{New}}{SD_{Old}}$$

$$B = (Mean_{New} - \frac{SD_{New}}{SD_{Old}} Mean_{Old})$$

where

SD_{New} is the standard deviation of anchor estimates in scale score metric
SD_{Old} is the standard deviation of anchor input values in scale score metric
$Mean_{New}$ is the mean of anchor estimates in scale score metric
$Mean_{Old}$ is the mean of anchor input in scale score metric

The 2001 field test forms were equated to the 2000 operational form concurrently (simultaneously) in one equating, and a single set of transformation constants was used to place the field test items onto the operational score scale. The 2002 field test forms were equated to the 2002 operational form using separate calibrations and equatings, and different sets of transformation constants were used to place the field test items onto the operational score scale. The same anchor set was used for each 2002 field test equating. The 2003 field test forms were calibrated and equated concurrently to the 2003 operational form, and a single set of transformation constants was used to place the field test items onto the operational score scale.

Anchor Set Security

In order for an equating to accurately place the items and forms onto the operational scale, it is important to keep the anchor items secure and to reduce anchor item exposure to students and teachers. Different anchor sets were used each year to minimize item exposure that could adversely affect the accuracy of the equatings. In addition, for each field test, statistical methods were employed to identify items that might have behaved differently when administered as anchor items. These analyses help uncover undesirable effects of potential item exposure. As a result of the analyses and also taking into account the benefits of maintaining the original anchor set , all anchor items were retained for equating purposes for each set of field test items.

Equating Results

Results of the Grade 4 Mathematics field test item equating are presented in Tables 3 through 5 below. The field test sample sizes (about 1300 to 2000 students per form/item) were sufficiently large to obtain stable IRT statistics. The efficacy of the equating was evaluated in part by evaluating correlations of anchor input (old) and estimate (new) values of a- and b-parameters and p-values. As a rule of thumb, the correlation between anchor a-parameter input and estimates should be at least 0.80 and the correlation between b-parameter input and estimates as well as p-values should be at least 0.90. Typically, correlations for the c-parameter estimates are not considered in equating evaluation due to the random nature of guessing behavior and difficulties of characterizing students' guessing behavior in field testing conditions differing from operational administrations.

As indicated in Tables 3 through 5, the correlations between anchor item inputs and estimates for b-parameters and p-values were greater than 0.90 in all cases. The correlations between anchor item input and estimates for a-parameters was over 0.80 in the 2002 field test equating and very close to 0.80 in the 2001 and the 2003 field test item equating. The high correlations between input values and estimates of the a- and b-parameters as well as actual and predicted p-values of the anchor items provide evidence of accurate item equating.

The 2002 field test sample size for item calibration and equating was somewhat smaller than the 2001 and 2003 sample sizes. The 2002 field test design required matching student-level field test data to operational data, and not all data could be matched successfully. In 2001, no matching of field test data to operational data was required, and the full field test samples were used for calibration and equating. In 2003, the census field test allowed for acquiring large samples of students. Even though the 2002 sample sizes were somewhat smaller than in the other years, they were still sufficiently large to conduct IRT analyses and obtain stable results.

Table 3
Equating Results for the 2001 Field Test

Field or Operational test form	Form A (OP)	Form B (FT)	Form C (FT)
Sample size	2030	1970	1868
Number of anchor items	48 (in Form A) from 2000 operational form		
Number of FT items	96 (across Forms B and C)		
a-parameter correlation between anchor item input and estimates	0.78		
b-parameter correlation between anchor item input and estimates	0.91		
p-value correlation between anchor item input and estimates	0.99		

Table 4
Equating Results for the 2002 Field Test

Field test form	Form A	Form B	Form C
Sample size	1299	1319	1303
Number of anchor items (operational)	30 from 2002 operational form		
Number of FT items	48	48	48
a-parameter correlation between anchor item input and estimates	0.85	0.83	0.84
b-parameter correlation between anchor item input and estimates	0.93	0.92	0.91
p-value correlation between anchor item input and estimates	0.92	0.92	0.92

Table 5
Equating Results for the 2003 Field Test

Field test form	23 forms
Sample size	1475 to 1884 per form
Number of anchor items (operational)	30 from 2003 operational form
Number of FT items	145 across all forms
a-parameter correlation between anchor item input and estimates	0.79
b-parameter correlation between anchor item input and estimates	0.97
p-value correlation between anchor item input and estimates	0.98

Table 6 provides means and standard deviations for the multiple-choice IRT item parameter estimates in the 2002 through 2005 test administrations. The parameter estimates presented in this table are in scale score metric. Similarity of these parameters across forms provides evidence of the comparability of the multiple-choice sections of the forms. The a- and c-parameters of the multiple-choice items from the 2002 through 2005 grade 4 Math tests were very similar. The b-parameter estimates are similar for years 2002, 2004, and 2005, and the mean b-parameter estimate for 2003 is slightly higher than the mean b-parameter estimates for other forms. Overall, the mean and standard deviations for the parameter estimates are similar across four forms.

Table 6
Summary Statistics for IRT Parameters for Grade 4 Mathematics

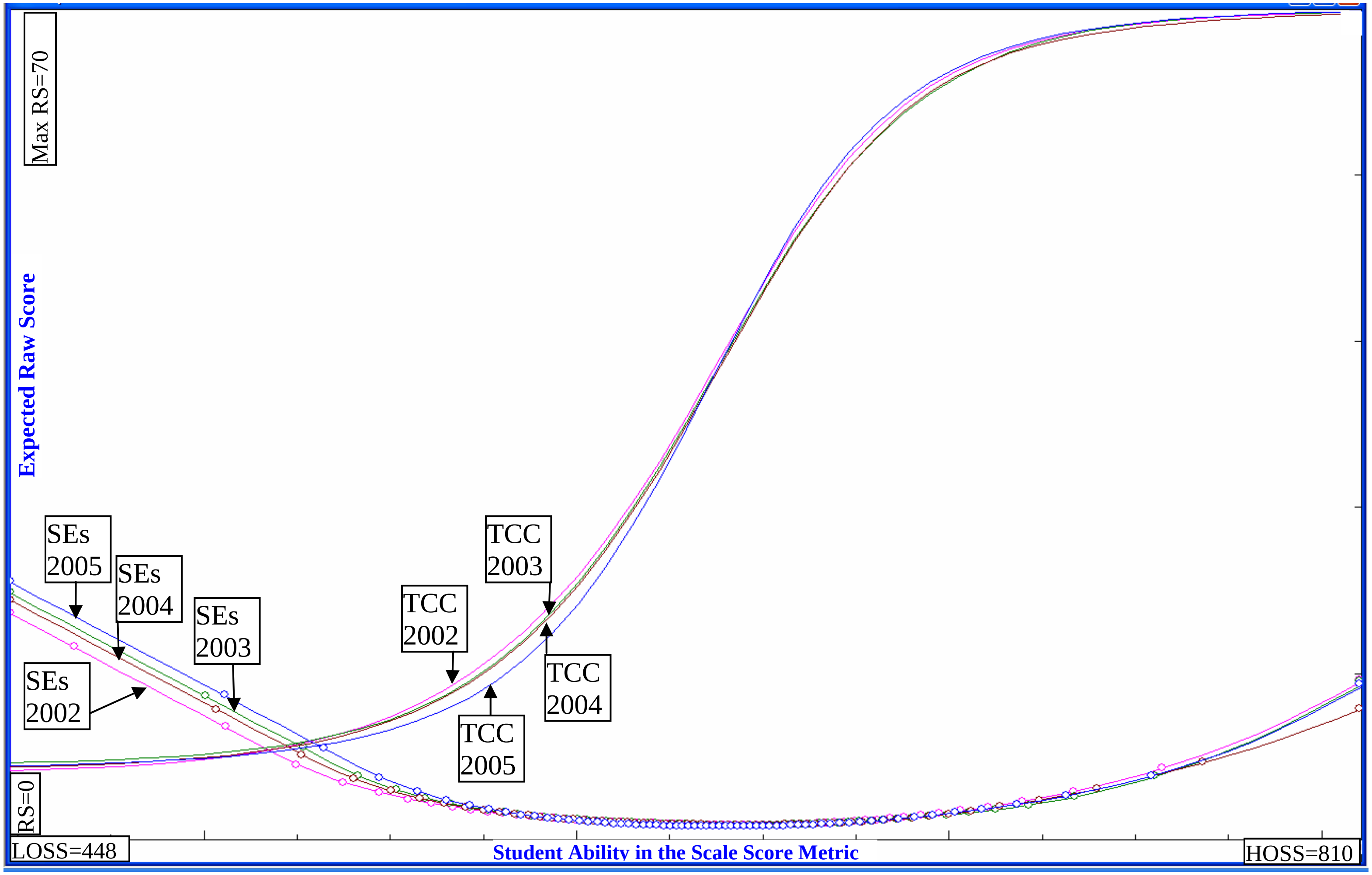
Administration year		IRT parameters		
		a	b	c
2005	mean	0.03	626.58	0.20
	SD	0.01	27.52	0.02
2004	mean	0.03	629.48	0.20
	SD	0.01	28.52	0.04
2003	mean	0.03	632.80	0.20
	SD	0.01	24.16	0.03
2002	mean	0.03	627.04	0.18
	SD	0.01	25.85	0.05

Score Comparability Across Forms

The purpose of test equating is to make statistical adjustments to account for small differences in test form difficulties so that reported scores are comparable (interchangeable) across test forms that meet the same content specifications. After forms have been equated, students are expected to receive approximately the same scale score regardless of which test form they take.

Figure 1 shows test characteristic curves (TCCs) for each operational test form (2002-2005). A TCC is a graphical overview of a form’s psychometric properties in the IRT scale score metric. The curves show the relationship between student ability on the scale score metric (X-axis) and the expected raw score on the test (in terms of proportion of the total possible points;Y-axis). The scale scores for grade 4 Math range from 448 to 810 (on X-axis) and the raw scores range from 0 to 70 (except for 2003 form which had a maximum raw score of 68; on Y-axis). Standard Error (SE) curves show the amount of measurement error at each ability level. The TCCs and SE curves for the forms are very well aligned (i.e., the curves are nearly coincident) across the ability range. The similarity of the shape and location of the TCCs and SE curves illustrates the psychometric comparability of test forms. The close alignment of TCCs and SE curves provides evidence that the forms are very similar in terms of test difficulty and discrimination, and supports the claim that scale scores from the different forms are interchangeable.

Figure 1
Operational Test Characteristic Curve and Standard Error Curve Comparison for Grade 4 Mathematics for 2002, 2003, 2004, and 2005



Note: HOSS refers to the highest obtainable scale score; LOSS refers to the lowest obtainable scale score

Concluding Remarks

As documented in this summary and in appropriate and publicly available technical reports, standard psychometric methods and industry-standard quality assurance practices were employed to equate New York State alternate forms and place them onto the same score scale. In this way, scale scores derived from the 2002-2005 test forms can be legitimately compared from one year to the next, and differences in scale scores across years reflect changes in student achievement.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association, Inc.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 443-459.

Burket, G. R. (2002). *PARDUX [Computer program, Version 1.26]*. Unpublished.

Fitzpatrick, A. R (1990). *Status Report on the results of Preliminary Analysis of Dichotomous and Multi-Level Items Using the PARMATE Program*. Unpublished manuscript

Fitzpatrick, A. R. (1994). *Two studies comparing parameter estimates produced by PARDUX and BIGSTEPS.* Unpublished manuscript.

Fitzpatrick, A. R. & Julian, M. W. (1996). *Two studies comparing the parameter estimates produced by PARDUX and PARSCALE.* Monterey, CA: Unpublished manuscript.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking (2nd Ed).* New York, NY: Springer.

Linn, R. L., and Harnisch, D. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*, pp. 109–118.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Menlo Park, CA: Addison-Wesley.

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

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parameter Scaling of Rating Data [Computer program].* Chicago, IL: Scientific Software, Inc.

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

Thissen, D. (1991). *MULTILOG user's guide, version 6*. Chicago: IL: Scientific Software International.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175-186.

Wright, B. D., & Linacre, J. M. (1992). *BIGSTEPS Rasch Analysis [Computer program].* Chicago, IL: MESA Press.

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

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 36, 233-25.