Research Report

A Didactic Approach to the Use of IRT True-Score Equating

Alina A. von Davier

Christine Wilson

Research & Development

December 2005
RR-05-26

A Didactic Approach to the Use of IRT True-Score Equating

Alina A. von Davier and Christine Wilson
ETS, Princeton, NJ

December 2005

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

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

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

Abstract

This paper discusses the assumptions required by the item response theory (IRT) true-score equating method (with Stocking & Lord, 1983; scaling approach), which is commonly used in the nonequivalent groups with an anchor data-collection design. More precisely, this paper investigates the assumptions made at each step by the IRT approach to calibrating items and equating tests, and discusses the approaches that one might take for checking whether these assumptions are met for a particular data set. We investigated two types of tests: tests that consist of multiple-choice items only, and tests that consist of both multiple-choice and free-response items. Real data from the AP® Calculus AB exam are used to illustrate the application of the IRT true-score equating method as well as for the comparisons.

Key words: Population sensitivity, test equating, item response theory (IRT), IRT true-score equating method, observed-score equating methods, 3PL

Acknowledgments

The authors would like to thank Neil Dorans, Ming-mei Wang, David Wright, and Dan Eignor for their comments and suggestions during the project. The authors also thank Krishna Tateneni for contributing to the initial planning of the study and to the analysis of dimensionality of the tests and Kim Fryer for editorial support.

Introduction

Test equating methods are used to produce scores that are exchangeable across different test forms. Item response theory (IRT; Cook & Petersen, 1987; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Petersen, Cook, & Stocking, 1983; Petersen, Kolen, & Hoover, 1989; Thissen, Wainer, & Wang 1994; and many others) has provided alternative ways to approach test equating.

This paper discusses the assumptions required by the IRT true-score equating method (with Stocking & Lord, 1983, scaling approach) that is commonly used in the nonequivalent groups with an anchor (NEAT) data-collection design. This study represents the first set of investigations we did in the context of a larger study that focused on analyzing the population sensitivity of the IRT equating functions (von Davier & Wilson, in press).

The goals of this paper are to:

1. investigate the assumptions made at each step of the IRT approach to calibrating items and equating tests with multiple-choice (MC) items only versus tests with both MC and free-response (FR) items
2. discuss the assumptions made by the observed-score equating methods used in this study
3. illustrate the steps we took to check the IRT assumptions
4. compare the equating results obtained by using the IRT true-score equating method with the results obtained by the traditional observed-score equating methods (only for the MC tests)

Real data from the AP® Calculus AB exam were used to illustrate both the new method and the comparisons.

Notations, Assumptions, Models, and Methods

In this section, we introduce our notation and explicitly present the assumptions that underlie the data-collection design, the IRT model, and the equating methods. We also describe the particular IRT models used in this study.

In the NEAT design, X and Y are the *operational test forms* given to two samples from the two *test administrations* (populations) P and Q, respectively; and V is a set of common items,

the *anchor test*, given to both samples from *P* and *Q*. The anchor test score, *V*, can be either a part of both *X* and *Y* (the internal anchor case) or a separate test (the external anchor case). The subscripts *P* and *Q* will indicate the populations.

The data structure for the NEAT design is illustrated in Table 1 (see also von Davier, Holland, & Thayer, 2004a).

Table 1

Description of the Data-Collection Design

	X	*V*	*Y*	
P	✔	✔	✔	*X*, *V* observed on *P*
Q		✔	✔	*Y*, *V* observed on *Q*

Note that Table 1 describes the data-collection procedure and does not refer to the tests scores to be used in the observed-score equating.

The analysis of the NEAT design usually makes two assumptions, which we combined into Assumption 1 (see also von Davier et al., 2004a).

Assumption 1. There are two populations of examinees, *P* and *Q*, who can each take one of the tests and the anchor test. Two samples are independently and randomly drawn from *P* and *Q*, respectively.

The usual IRT assumptions for fitting IRT models and for calibrating items from the tests and the anchor are presented in Assumption 2.

Assumption 2. We assume that the tests to be equated (*X* and *Y*) and the anchor (*V*) are unidimensional and measure the same construct. For all items in these tests, the unidimensionality, local independence, and monotonicity assumptions hold (see Hambleton et al., 1991, for example).

Hence, IRT models rely on the assumptions of monotonicity, unidimensionality, and local independence at the item level; these models express the probability of a response, $z_{ni,}$ of a given person, n ($n = 1, \ldots, N$), to a given item, i ($i = 1, \ldots, I$), as a function, f, of the person's ability (latent), θ_n, and a possibly vector-valued item parameter, $\boldsymbol{\beta}_i$; for example:

$$P_{ni} = P(X = z_{ni}) = f(z_{ni}, \theta_n, \boldsymbol{\beta}_i).$$

In the case of the well-known three-parameter logistic (3PL) model (Lord & Novick, 1968), the vector $\boldsymbol{\beta}_i$ consists of the slope, the difficulty, and the guessing parameter; that is: $\boldsymbol{\beta}_i^{\,t} = (a_i, b_i, c_i)$.

The 3PL model, which serves as the standard example of an item response model (IRM) in this paper, is given by

$$P(zni = 1 \mid \theta n, ai, bi, ci) = ci + (1 - ci)\, logit\text{-}1[ai\,(\theta n - bi)], \qquad (1)$$

where z_{ni} denotes the answer of the person n to the item j, $\text{logit}^{-1}(\cdot) = \exp(\cdot) / [1 + \exp(\cdot)]$, and a's, b's, and c's are the item parameters, θ is the person parameter (ability or competency of interest), and $P(z_{ni} = 1 \mid \theta_n, a_i, b_i, c_i)$ is the conditional probability of a correct answer of the person n to the item i (see Hambleton et al., 1991; or Lord, 1980, for details).

The generalized partial-credit model (GPCM; Muraki & Bock, 1997) for the polytomous items (with $m + 1$ categories, for example) is based on the assumption that each probability of choosing the k^{th} category over the $(k\text{-}1)^{\text{th}}$ category follows a dichotomous model (with k between 0 and m+1):

$$P(z_{nik} = 1 \mid \theta_n, a_i, b_{ik}) = \frac{\exp[\sum_{\nu=0}^{k} a_i\,(\theta_n - b_{i\nu})]}{\sum_{\lambda=0}^{m_j} \exp[\sum_{\nu=1}^{\lambda} a_i\,(\theta_n - b_{i\nu})]}, \qquad (2)$$

where $b_{i0} = 0$ (arbitrarily fixed to 0). The threshold parameters in the partial-credit model, b_{ik}, are the intersection points between the probability curves P_{nik} and $P_{nik\text{-}1}$ (see Muraki & Bock, 1997).

The GPCM is the item response model used to fit the data containing the FR items in this study.

Table 1 shows that in the NEAT design **X** is not observed in the population Q, and **Y** is not observed in the population P. To overcome this feature, all linking methods developed for the NEAT design (both observed-score and IRT methods, which are also called true-score methods) must make additional assumptions that do not arise in the other linking designs. The assumption that the IRT models make for the NEAT design is given in Assumption 3.

Assumption 3. If the model fits the data in each of the (two) populations, then the item parameters of the common items are population invariant (up to a linear transformation).

If the IRT calibration were to be carried out separately on the two samples from the two different populations *P* and *Q*, two sets of parameter estimates for the anchor test would be obtained. All IRT models have an intrinsic lack of identifiability of the item and person parameters that is usually addressed by imposing some restrictions on the item or person parameters (there are various ways to implement restrictions, and consequently different software use different approaches). This leads to a need for placing the item parameters on the same scale even in the absence of equating. In other words, even if the same test instrument would have been given to the two groups and if the calibration was done separately in the two groups then, the item parameters still need to be placed on the same scale by some sorts of linear transformations. Hence, for scaling and equating purposes in a NEAT design and assuming that Assumption 3 holds, the two separate parameter estimates of the anchor in the two groups need to be placed on the same scale. There are various methods for the scale transformation: the mean-mean, mean-sigma methods, or the characteristic curve methods such as the Stocking and Lord method (1983) and the Haebara method (1980).

In this study we use the 3PL model for MC items and the generalized partial-credit model for the FR items. The characteristic curve method (Stocking & Lord, 1983) is used to place the separately estimated parameters onto a common scale. Next, the true-score equating method is used to obtain equivalent scores on *X* and *Y* (see Kolen & Brennan, 2004, or Petersen, Kolen, & Hoover, 1989, for a detailed description of the method; see also the description of the process of IRT true-score equating). The IRT equating method requires that the tests are number-right scored, which is an implicit assumption that there are no omits. This is captured by Assumption 4.

Assumption 4. IRT equating assumes that there are no omitted responses.

IRT true-score equating also introduces Assumption 5.

Assumption 5. The relationship between the true scores generalizes to the observed scores.

Although there is no theoretical reason why Assumption 5 might hold, studies indicate that the resulting true-score conversion is similar to the conversion of the observed scores (see Kolen & Brennan, 2004, for a discussion of the IRT true-score equating method).

Hence, the study of the population sensitivity of the IRT true-score equating function relies on the five assumptions already given.

The observed-score equating functions investigated in this study (chain equipercentile and Tucker) also make assumptions in order to overcome the missing-data-by-design condition, a feature of the NEAT design. We will not give any computational detail for the two observed-score equating methods, since they are well known. We give the assumptions for the two methods in order to emphasize that all equating methods require some (nontestable) assumptions to be fulfilled. The assumptions and the formulas for the chain equipercentile equating function and for the Tucker linear equating function are given in Kolen and Brennan (2004); von Davier, Holland, and Thayer (2004a, 2004b); and von Davier (2003).

Assumption 6 (Chain Equating). The linking functions, from X to V and from V to Y, are population invariant.

***Assumption 7 (Tucker Equating)*.** The linear regressions of X on V and of Y on V are population invariant. The conditional variances of X given V and of Y given V are population invariant.

By looking carefully at the Assumptions 3, 6, and 7, you can see that each use a particular type of population invariance assumption. Hence, both the true-score and the observed-score equating methods make similar types of assumptions.

Data

In this section, we describe the data used to investigate whether the five assumptions for achieving the IRT equating function hold.

The data were from the 2001 and 2003 administrations of the AP Calculus AB exam. In these data sets, there were 145,415 examinees in the 2001 administration and 163,142 examinees in the 2003 administration. These data contain the examinees who took the regular operational forms of the AP Calculus AB exam in 2001 and 2003, respectively. The operational data (i.e., the data on which the equating is conducted operationally) contain subsamples from each of these larger samples.

This AP exam uses a NEAT design, with the year 2003 test being linked back to the 2001 test. The anchor test, V, is an internal anchor within the MC component of the test. The MC sections consist of 45 items each; the (internal) anchor has 15 items.

Each particular AP Calculus AB exam has a composite score, which is a weighted sum of scores from MC and FR parts. For the AP Calculus AB exam, the FR section contains six FR

items, each with 10 possible score categories (from 0 to 9). Operationally, AP equates only the MC scores, not the composite scores.

For this particular exam, the correlation between the FR and the MC scores was .86 for 2001 and .87 for 2003. The correlation of the MC scores with the composite was .96 for 2001 and .97 for 2003. Reliability for the MC scores was .90 for 2001 and .89 for 2003. Reliability for the composite scores was .94 for both 2001 and 2003.

Operationally, the tests were scored using rounded formula scoring. In order to satisfy Assumption 4 and achieve the IRT conversion, the data files were rescored with a software package (ETS, 2004) using the number-correct scoring method. In this study, we focused only on the raw score equating and not on the scale conversion.

The effect size for the difference between 2003 and 2001 for all examinees is (6.58 – 6.30)/3.975 = 0.070, or 7% (3.975 is the average of 4.02 and 3.93).

The differences reflected in the summary statistics for the common items suggest that the examinees from 2003 were slightly more able than those from 2001. See Table 2.

Table 2

Summary Statistics of the Observed Frequencies of X and V for a Sample of Examinees From Population P and From Population Q for the AP Calculus AB Exam, MC Items Only

	Total in 2003 (*P*)		Total in 2001 (*Q*)	
N	163,142		145,415	
	X	*V*	*V*	*Y*
Mean	19.29	6.58	6.30	18.56
SD	11.12	4.02	3.93	10.66
Skewness	.09	.09	.16	.10
Kurtosis	− .90	− .94	− .87	− .85

The correlation between the test *X* (MC items only) and (internal) anchor test *V* in *P* is 0.9087, and it is 0.9278 between *Y* (MC items only) and *V* in *Q*.

Table 3 suggests (also taking into account the information from Table 2 and the high correlation between the MC and FR items) that the FR section was more difficult in 2003 than in 2001 (these are observed data; no equating has been carried out yet). This information suggests that we should expect a larger difference between the tests' characteristic curves (TCCs) when

the FR items are included than when only the MC items are investigated, and we should expect the IRT equating to adjust for this difference in difficulty (if Assumptions 2 and 3 hold).

Table 3

Summary Statistics of the Observed Frequencies of the FR Section for a Sample of Examinees from Populations P and Q for the AP Calculus AB Exam

	FR_P = 2003	FR_Q = 2001
N	163,142	145,415
N items	6	6
Mean	17.94	23.55
SD	11.77	13.47

IRT True-Score Equating: Assumptions Check

Since IRT equating is not employed operationally for this assessment, we first checked if Assumption 2 holds. This assumption is necessary for applying an IRT model in the equating process. (See Cook, Dorans, Eignor, & Petersen, 1985; Cook & Eignor, 1991; Cook & Petersen, 1987; Jodoin & Davey, 2003; Petersen, Cook, & Stocking, 1983; Petersen, Kolen, & Hoover, 1989; and Wainer, Thissen, & Wang, 1994, for a detailed discussion of the robustness of the IRT equating function.) Although the unidimensionality and local independence assumptions might not hold strictly, the IRT models might be robust enough to be used in practical situations (Cook, et al., 1985; Cook & Petersen, 1987; Thissen, et al., 1993). We investigated the dimensionality of the two tests as well as of the individual anchors (see Hattie, 1985) from different perspectives: test construction, as well as characteristics and fit of the IRT models. The several investigations were carried out for both datasets from the two administrations.

Analysis of the Dimensionality of the Tests

1. We looked carefully at the items: The items in both the MC and the FR sections did not share elements (passages, for example) and did not mechanically depend on each other.

2. We checked if the MC and the FR items were intended to measure the same construct by consulting the documents available for this exam (see Wainer et al., 1993).

3. We determined that the MC and FR correlated highly, as mentioned above (0.86 in 2001 and 0.87 in 2003).

4. We used factor analysis to investigate the dimensionality of the tests (MC + FR) given to both administrations. Exploratory factor analysis (Browne, Cudeck, Tateneni, & Mels, 1999) was carried out (no constraints beyond those required for model identification were imposed). The polychoric correlation matrix was estimated. The data was assumed to be multivariate normal. The ordinary least square (OLS) discrepancy function was minimized. First, 1-factor models were fit to both tests from the two administrations. The root mean square error of approximation (RMSEA) showed a reasonable good fit for the 1-factor model in both cases. The first eigenvalue was 17 in the sample from 2001 and 19 in the sample from 2003 (see the scree plots in Figures 1 and 2). The proportions of the variances accounted for by the first factor were 17/51 = 0.33 or 33% for the MC + FR in the sample from 2001 and 19/51= 0.37 or 37% for 2003. Since three eigenvalues were larger than 1 (the Kaiser-Guttman rule), we considered the issue of fitting at least a 2-factor model (the third eigenvalue in both groups was only slightly larger than 1).

5. For the 2-factor model, the factor-loading matrix was rotated (both orthogonally and obliquely) to search for an interpretable solution (simple structure). The results were inconclusive. The fit as measured by the RMSEA was slightly better for the 2-factor model, but neither loading was interpretable or high enough to support choosing a 2-factor model. It appears that the items did indeed measure one factor (the first factor seems to explain most of the variance).

6. We investigated if the dimensionality assumption was consistent for the test forms (MC + FR) across administrations: For example, the dimensionality of the two test forms should also have been similar in the secondary dimensions (see Jodoin & Davey, 2003). We found that the same factorial structure existed in the two test forms; the same number of eigenvalues larger than 1 existed in both administrations and they were of similar size (except the value of the first eigenvalues described above; see Figures 1 and 2).

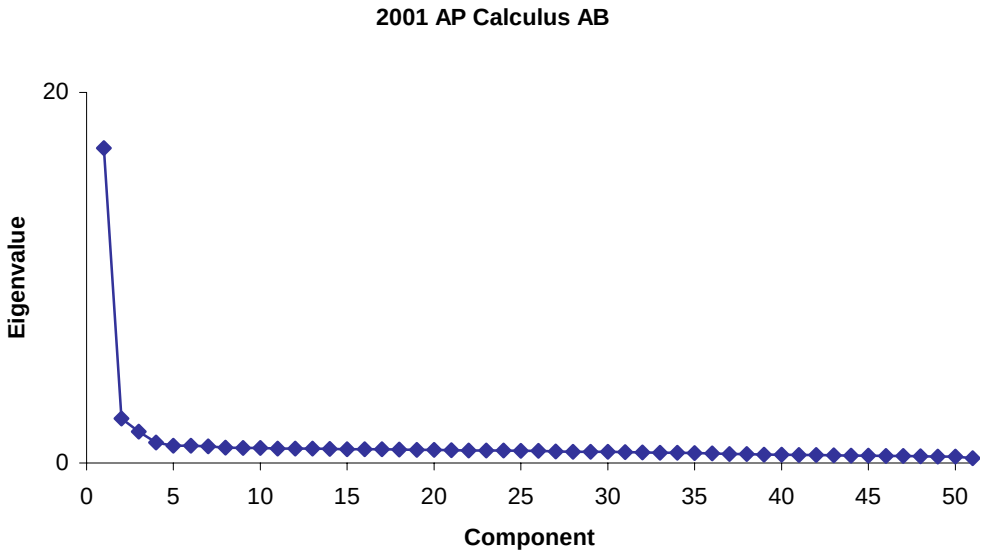

Figure 1. **Scree plot for the 2001 AP Calculus exam (MC + FR items).**

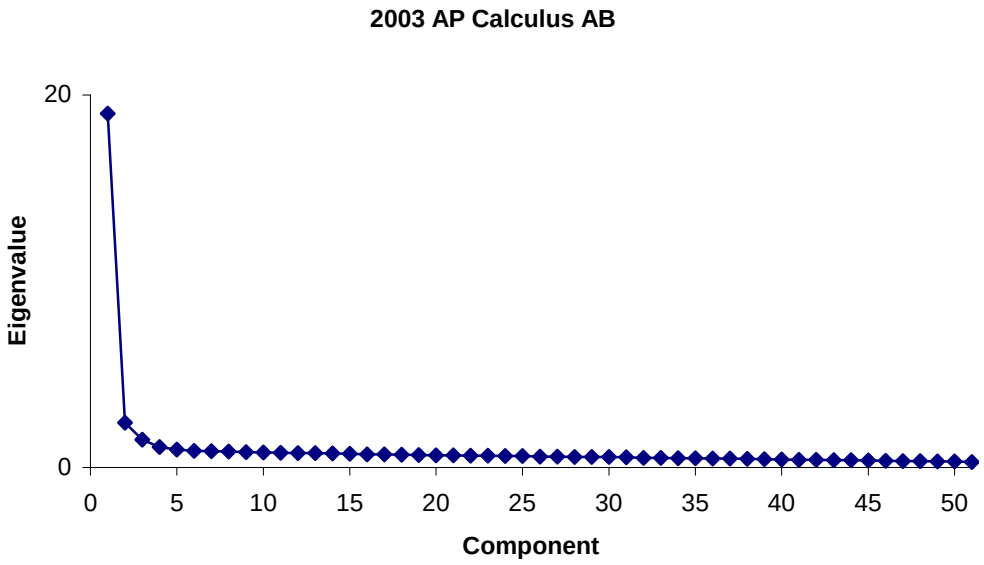

Figure 2. **Scree plot for the 2003 AP Calculus exam (MC + FR items).**

7. Calculus AB includes some items that allow test takers to use a calculator, and some items that do not. We did not find any pattern in the factorial structure that is consistent with the use of a calculator.

8. Calculus AB has two separately timed sections. We did not find any serious speededness issues in the last items from each of the two sections.

9. We checked for evidence of differential item functioning (DIF) in the MC item sets and in the FR item sets with respect to the largest subpopulations, males and females (see Ackerman, 1992; Dorans & Schmitt, 1991; Holland & Thayer, 1988). We used the Mantel-Haenszel index (M-H; Holland & Thayer, 1988) for investigating DIF in the MC items and the *overall index* (PolyStand) for investigating DIF in the FR items (Dorans & Tateneni, 1993). There were no DIF items in the common set of items; there were some DIF items in the tests; however, the value of the M-H index was low, and we didn't notice any pattern in the sign of those M-H index values. Two items showed high DIF values in the FR section; these were also the most difficult items. We will look carefully at the fit of the IRT model to these items later in the study.

Analysis of the Fit of the IRT Model

1. We fitted the 3PL model to the MC items, and the GPCM to the FR items. We investigated the item characteristic curves (ICCs) plots provided by PARSCALE (Muraki & Bock, 1997) in GENASYS (ETS, 2004). It appeared that there were no outliers among the items in each of the two administrations.

2. We compared the differences between the item parameter estimates for the MC items when calibrated alone with those estimates of the MC item when calibrated together with the FR items. The estimates changed slightly, but they were mostly inside the plus/minus two standard error band around the estimates from the calibration for the MC items only (the standard errors are very small in this situation, where the sample sizes are very large). The ICCs for the MC items from the two studies appeared to be similar.

3. We compared the estimates of the parameters of the posterior distribution for the ability from the two types of items (the prior distributions for the abilities in the two groups were set to be normal distributions, with means 0 and standard deviation 1). Again, the changes in the estimates were small (about 0.01 for the means and 0.04 for the variances).

4. We also investigated the ICCs for the FR items (see an example of the ICC for a FR item in Figure 3; the figure shows the probabilities associated with each category and the thresholds, denoted with D, for each of the 10 categories). The 10 categories for all of the FR items were recovered, and although the fit was not very good for the most difficult items (two of them being those that showed DIF), we did not see any clear evidence for collapsing categories (especially after considering the findings from model fit discussed above).
5. We looked at the chi-squared values (the only fit measures provided by the available software); however, given that the samples were very large, these values were not directly relevant. No other fit measures were considered or available.

The findings from above provide some indication that, indeed, the FR items measured the same construct as the MC items (see also the other analyses described above). We concluded that Assumption 2 for the IRT model holds well enough for our analysis.

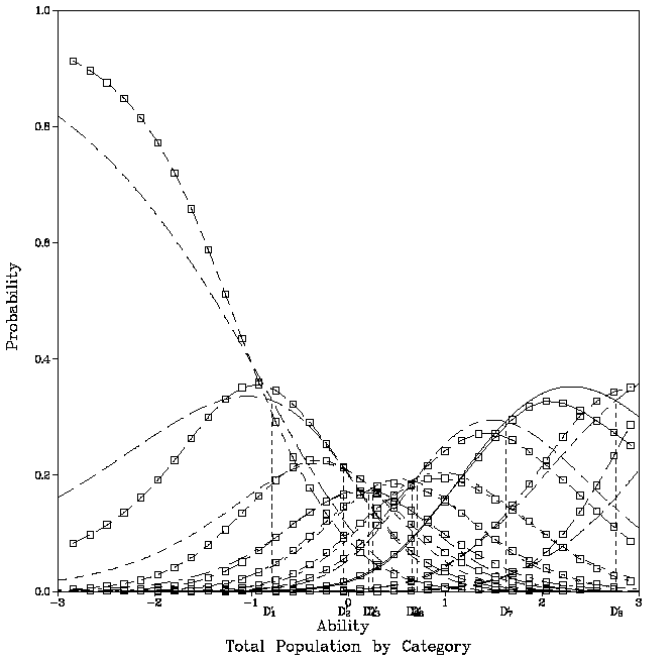

Figure 3. **Example of an item characteristic curve for a free response on the 2001 AP Calculus exam.**

In the following paragraph we describe the steps we took to check Assumptions 3 to 5 and to accomplish the equating.

Analysis of the Assumptions of the IRT True-Score Equating and Process Description

1. The 3PL model from (1) was fitted separately to both administrations. A detailed item analysis was carried out on all items. We found a few difficult items that had a large number of omits (more than 30% of the examinees omitted them), which is not surprising for a formula-scored test, particularly in light of the test instructions that the examinees received. This implies that Assumption 4 might not hold.

2. The item parameters from the two estimations on the common item set were investigated (see Assumption 3). We plotted the item parameters to look for outliers (those items with estimates that do not appear to lie on a straight line). Figures 4 and 5 show the item parameters, the slope (*a*) and the difficulty (*b*-parameters) for the first study (MC only), for the calibration for the total population. Figures 6 and 7 show the respective item parameters for the second study (MC + FR). There were no significant changes among the item parameters for the MC common items in the first study versus the second study (after including the FR section).

3. The item parameters were placed on the same scale by the characteristic curve method of Stocking and Lord (1983). We also looked at the ICCs for the common items, before and after transformation. Only two common items seemed to have a larger difference in the ICCs across administration (Figure 8 and Table 4). However, the differences were not very large and the differences in the two ICCs for these two items had different signs, which implies that the differences were not systematic (i.e., the common items were not systematically more difficult, easy, or omitted). Therefore, we did not consider the evidence to be strong enough to exclude them from the set of the common items. Moreover, the anchor test is relatively short (15 items); hence, excluding items might create problems in the equating process. A very similar pattern for the ICCs for the common items was found when the FR items were included.

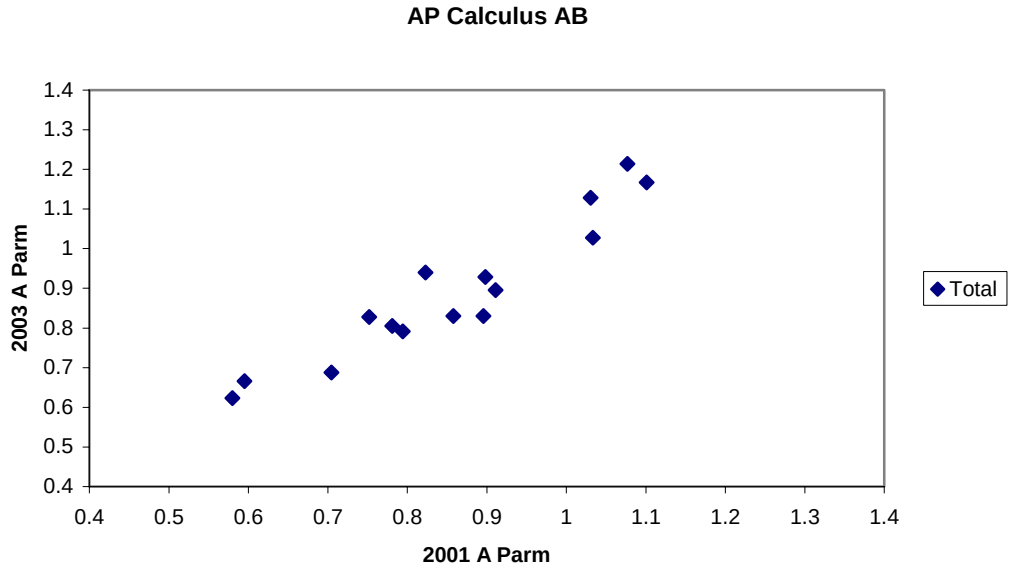

Figure 4. **The *a*-parameters for the anchor items for the two administrations (MC items only).**

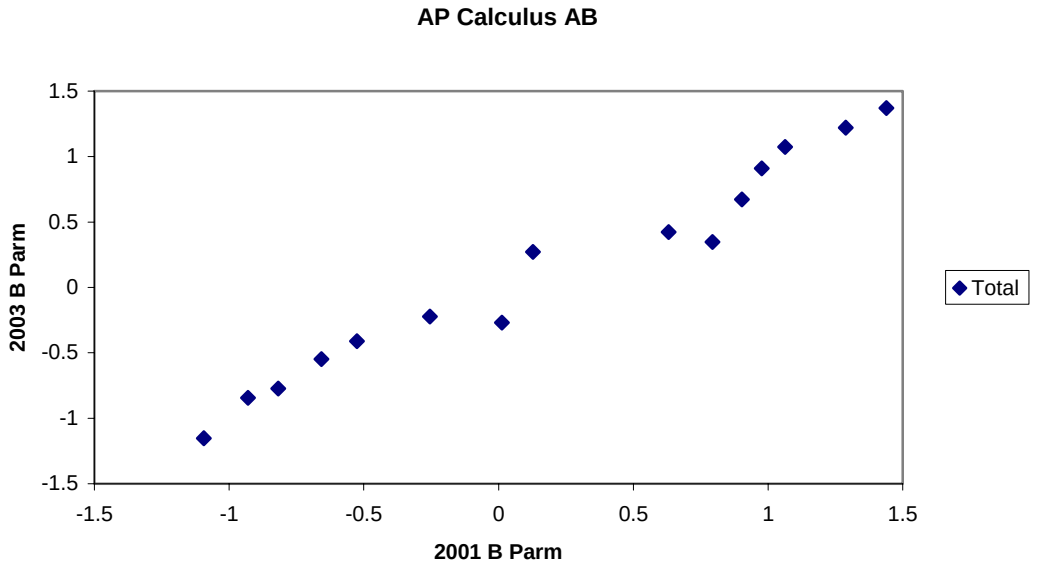

Figure 5. **The *b*-parameters for the anchor items for the two administrations (MC items only).**

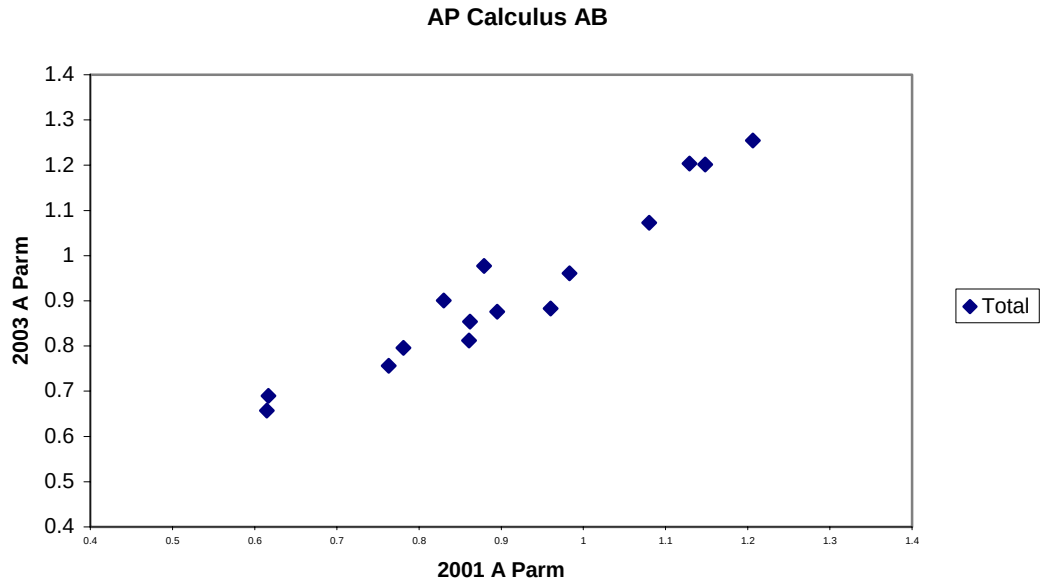

Figure 6. **The *a*-parameters for the anchor items for the two administrations (MC + FR items).**

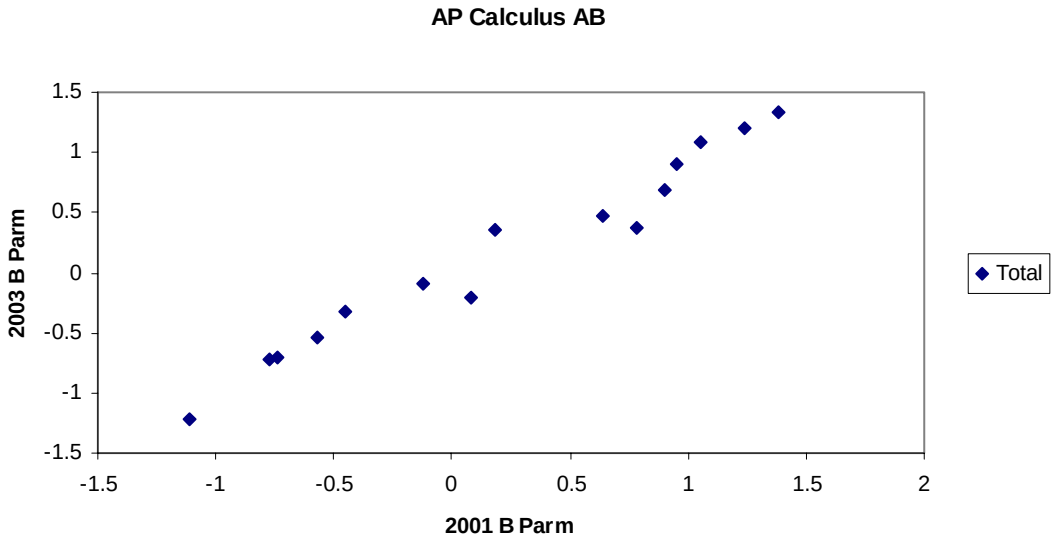

Figure 7. **The *b*-parameters for the anchor items for the two administrations (MC + FR items).**

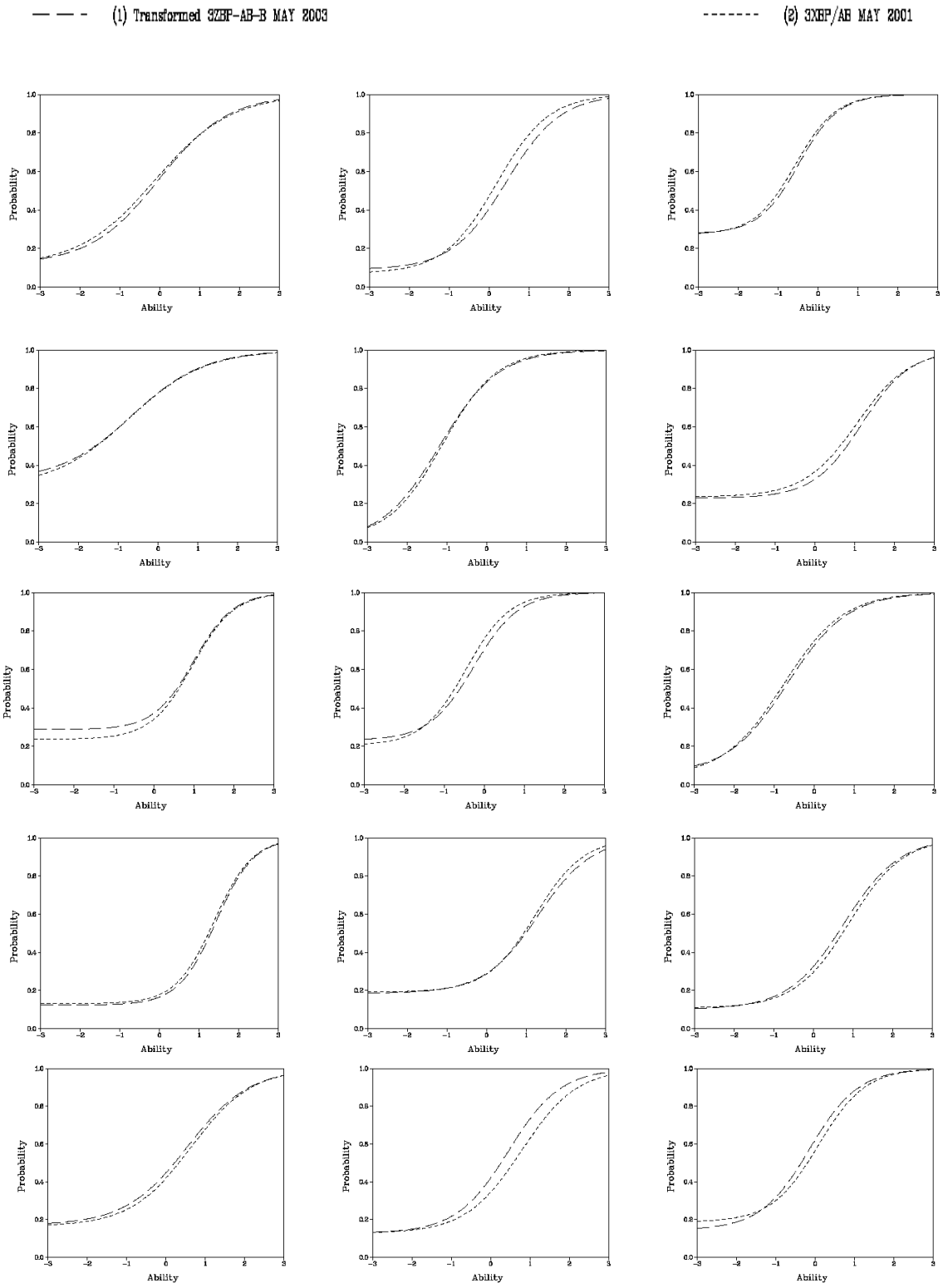

Figure 8. **ICC plots for common items after scaling (calibration with MC items only).**

Table 4

Item Parameters for Common Items After Scaling (Calibration for MC Items Only)

	Reference form parameters (*Y*)			Transformed parameters (*X* after scaling)		
Item no.	*a*	*b*	*c*	*a*	*b*	*c*
5	.595	–.254	.063	.641	–0.160	.068
9	.896	.128	.053	.799	0.355	.062
10	1.077	–.657	.246	1.168	–0.498	.288
12	.580	–.930	.226	.600	–0.807	.279
14	.911	–1.093	.010	.862	–1.127	.007
19	.823	1.063	.230	.904	1.189	.224
20	1.030	.976	.234	1.086	1.019	.285
24	1.033	–.526	.182	.989	–0.355	.200
25	.781	–.817	.023	.774	–0.733	.028
30	1.101	1.440	.129	1.123	1.496	.122
33	.858	1.288	.187	.799	1.342	.181
35	.794	.903	.100	.762	0.772	.093
40	.705	.631	.157	.662	0.512	.154
41	.752	.794	.120	.797	0.432	.118
43	.898	.013	.167	.894	–0.207	.132

4. The test characteristic curves (TCCs) for the two tests (after the parameter estimates for the two tests were placed on the same scale) were almost on top of each other for the MC items only (see Figure 9).

5. In contrast, for the tests that consisted of MC + FR items, the TCCs for the two tests (after the parameter estimates were placed on the same scale) differed from each other for almost the entire ability range, with the TCC for test *Y* (from 2001) being above the TCC for test *X* (from 2003). This suggests that the test given in 2003 was more difficult than the one given in 2001 (see Figure 10). This was an expected result, considering the information given in Tables 2 and 3.

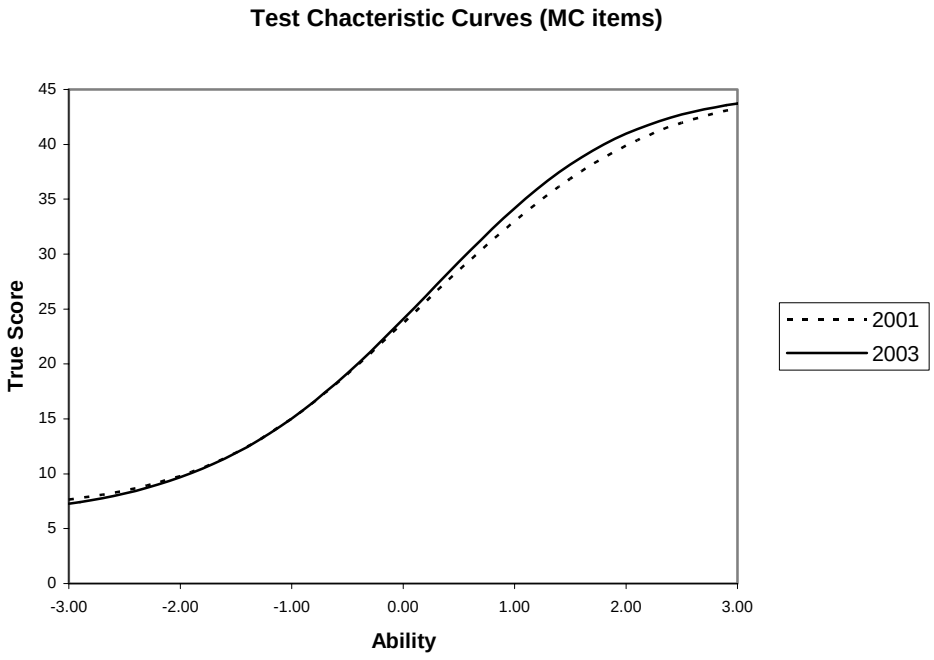

Figure 9. **Test characteristic curves for the 2001 and 2003 AP Calculus exam (MC items only).**

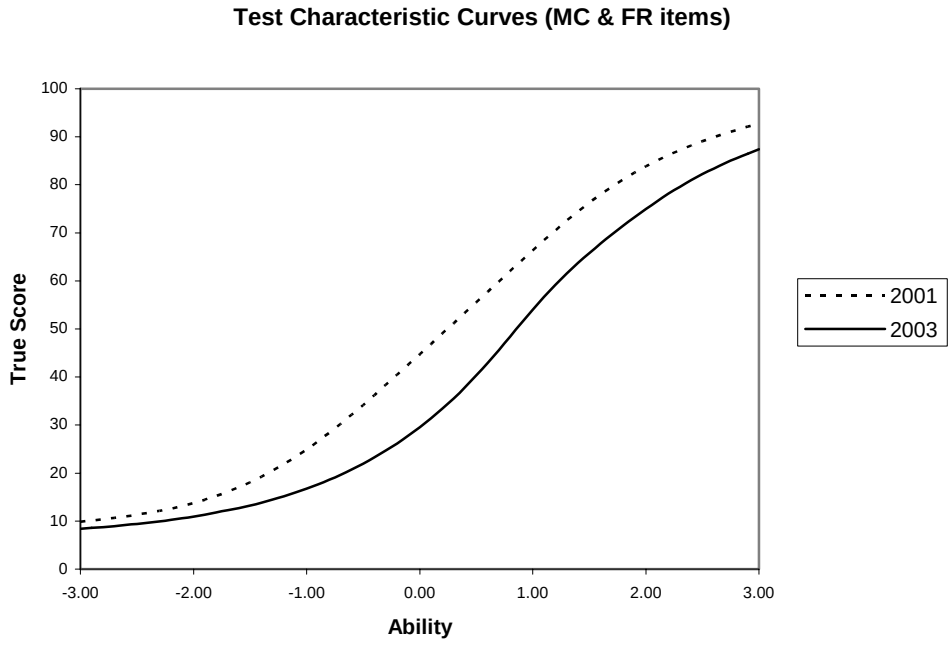

Figure 10. **Test characteristic curves for the 2001 and 2003 AP Calculus exam (MC + FR items).**

6. The IRT true-score equating was performed. This was done by first looking for the corresponding estimated θ value for each estimated true score on the new form, and then finding the estimated true score on the old form for that specific estimated θ value. Hence, a conversion relationship between the true scores was estimated.
7. This estimated conversion function was applied to observed scores, as if the estimated true scores were identical to the observed scores (see Assumption 5). We used the method proposed by Lord (1980) that is implemented in GENASYS (ETS, 2004) to produce equivalents for the number-correct scores below the sum of the *c*-parameters estimates.
8. We used the method proposed by Lord (1980) that is implemented in GENASYS (ETS, 2004) to produce equivalents for the number-correct scores below the sum of the *c*-parameters estimates.
9. The conversion function obtained for the MC items is plotted in Figure 11 and seems to be almost linear. The conversion function obtained for the MC + FR items is shown in Figure 12, and it is clearly nonlinear. It seems that the IRT equating appropriately adjusts for the difference in difficulty in the FR sections across the two administrations.

Figure 12 plots the IRT conversion line for MC + FR items, which is obviously nonlinear, reflecting the differences in the distributions of the two tests (due to the differences in difficulty in the FR items sections).

Comparison of True- and Observed-Score Equating Results

In order to compare the equating results from the IRT approach with the equating results from the traditional methods, we use the difference that matters (DTM) value as a reference.

Dorans and Feigenbaum (1994) and Dorans, Holland, Thayer, and Tateneni (2003) use the notion of a difference that matters in score reporting. The DTM for a particular exam depends on the reporting scale. In AP there are two metrics of interest: the composite score metric and the AP grade scale. The scale that we used in this study for reference was the composite score metric (although we used the same weight, of 1, for the MC items as for the FR items in forming the composite). The unit of this score scale is one point. Hence, a difference

between equating functions larger than a half point on this scale means a change in the reporting score; therefore, it defines the DTM for this particular exam and for our studies. All the equating results were compared to the DTM of a half point in this study.

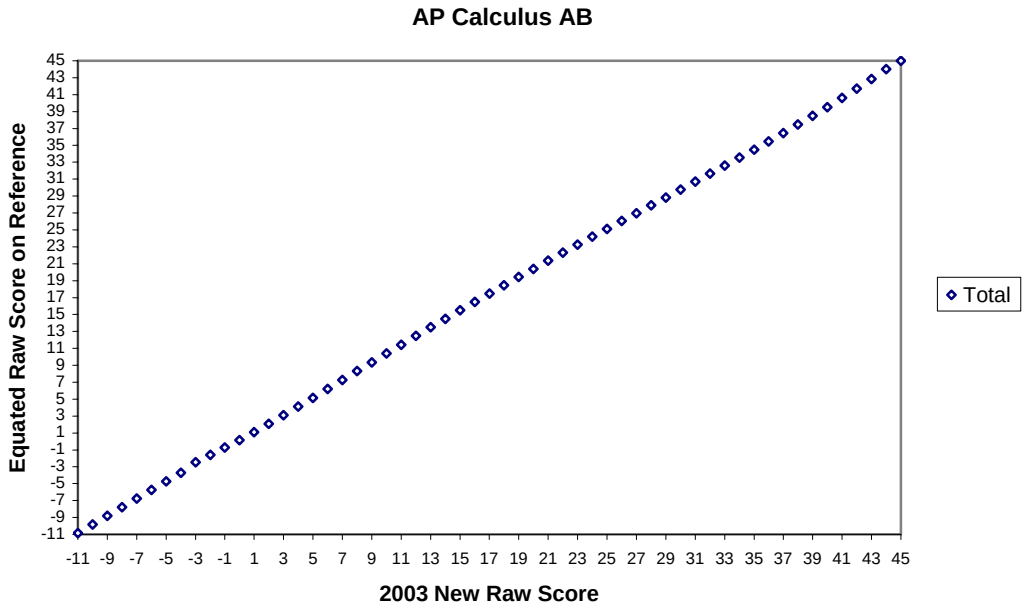

Figure 11. **The IRT conversion (MC items only).**

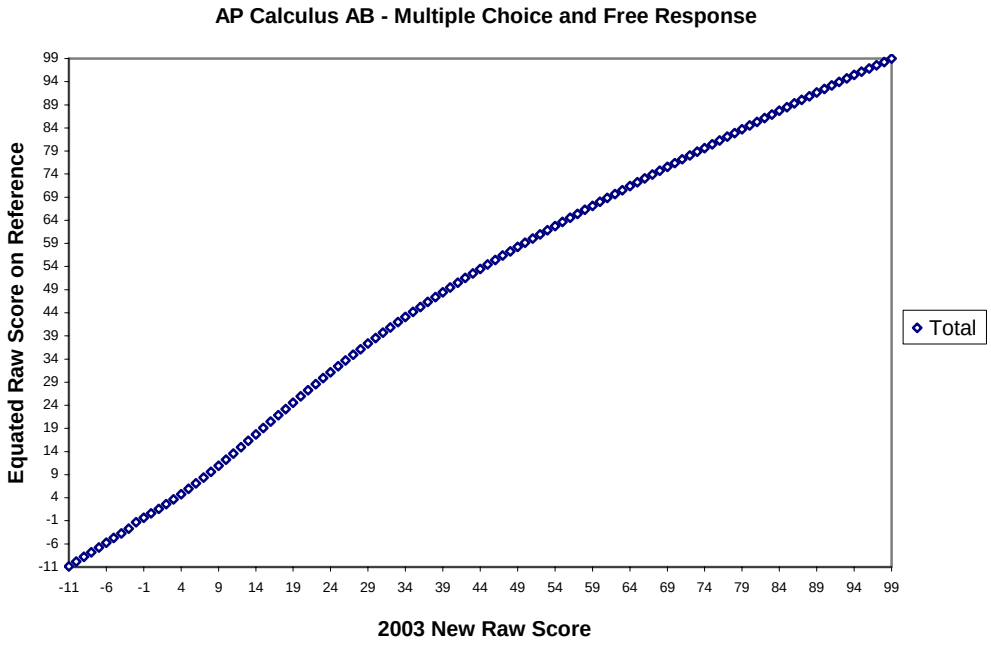

Figure 12. **The IRT conversion (MC + FR items).**

Figure 13 shows the differences between the equating results for Tucker, IRT, and chain equipercentile methods. We see that there are two score points (42 and 43 for the difference between the Tucker and the chain methods, and 44 and 45 for the difference between the Tucker and the IRT methods) where the differences exceed a DTM of 0.5. We also note that the differences between the Tucker and the chain functions seem to be smaller than those between the Tucker and the IRT functions for most of the score points. Given that the two samples from the two populations do not differ too much in ability as measured by the anchor, we expect the observed-score linear functions to agree with each other and the observed-score nonlinear functions to agree with each other; unfortunately, for this case, we only had one linear and one nonlinear observed-score function, each of them requiring different assumptions (Assumptions 6 and 7). The two observed-score equating functions slightly disagreed. This disagreement was due to the differences in shapes of the tests in the two administrations (see Table 2). However, the IRT and the chain equipercentile agreed very well for most of the score range, and displayed large disagreement only at the highest score points, where IRT tends to give higher scores than the chain equipercentile. It appears that all methods disagreed at the higher scores.

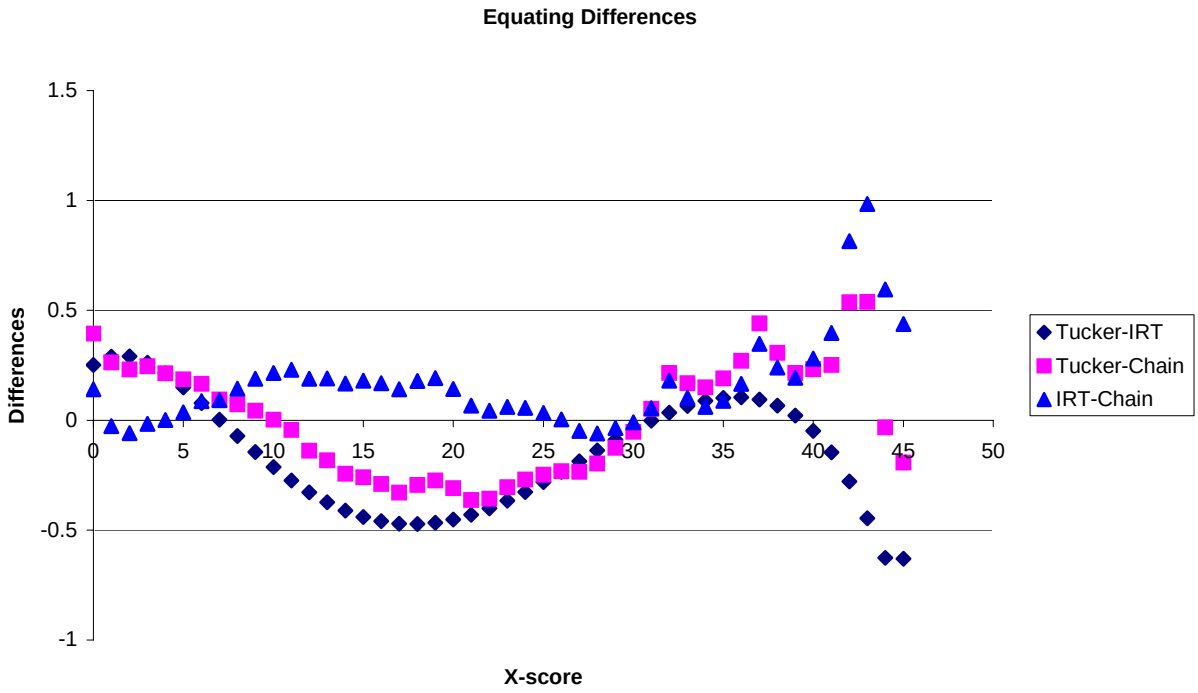

***Figure 13*. Equating differences between functions (Tucker, IRT, and chain equipercentile functions; MC items only).**

Discussion and Conclusions

This study provides a discussion of the assumptions required by the IRT models, the item calibration procedure, and IRT true-score equating method in a NEAT design and a step-by-step check of how well these assumptions are met in the data set at hand. In addition, two sets of analyses were conducted: one where only the MC section was investigated and one where both the MC and the FR sections were considered. In both cases, the anchor was internal and consisted of MC items only. Usually, practitioners construct the set of common items to be a miniature, in content and statistical properties, of the test forms to be equated. In this study, the available anchor set consisted only of MC items and was relatively short. These limitations of the anchor restrict the implications of the results.

This study concludes that, for this particular data set, the IRT true-score equating method might be an appropriate equating method: The IRT models fit the data to an acceptable degree and the IRT function appropriately adjusts for the difference in difficulty in the FR items across administrations. By looking carefully at the Assumptions 3, 6, and 7, we can see that they are all particular types of population invariance assumptions. Hence, both the true-score and the observed-score equating methods make similar types of assumptions.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67–91.

Browne, M. W., Cudeck, R., Tateneni, K., & Mels, G. (1999). CEFA: Comprehensive exploratory factor analysis [Computer software]. Retrieved November 10, 2004, from the Ohio State University Web site: http://quantrm2.psy.ohio-state.edu/browne/software.htm

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1985). *An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating* (ETS RR-85-30). Princeton, NJ: ETS.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*, 37–45.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225–244.

von Davier, A. A. (2003). *Notes on linear equating methods for the non-equivalent groups design* (ETS RR-03-24). Princeton, NJ: ETS.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). The chain and post-stratification methods for observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement*, *41*, 15-32.

von Davier, A. A., Holland, P. W., & Thayer, D. T (2004b). *The kernel method of test equating*. New York: Springer Verlag.

von Davier, A. A., & Wilson, C. (in press). Population invariance of IRT true-score. equating. In A. A. von Davier & M. Liu (Eds.), *Population invariance of test equating and linking: Theory extension and applications across exams*. Princeton, NJ: ETS.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT.* In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.) *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (RM-94-10). Princeton, NJ: ETS.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender groups for three Advanced Placement Program® exams. In N. J. Dorans

(Ed.), *Population invariance of score linking: Theory and applications to Advanced Placement Program examinations* (ETS RR-03-27, pp. 19–36). Princeton, NJ: ETS.

Dorans, N. J., & Schmitt, A.P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (ETS RR-91-47). Princeton, NJ: ETS.

Dorans, N. J., & Tateneni, K. (1993). PolyStand [Computer software.] Princeton, NJ: ETS.

ETS. (2004). GENASYS [Computer software]. Princeton, NJ: Author.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164.

Holland, P. W., & Thayer, D. T. (1988). *Differential item functioning and the Mantel-Haenszel procedure* (ETS RR-86-31). Princeton, NJ: ETS.

Jodoin, M. G., & Davey, T. (2003). *A multidimensional simulation approach to investigate the robustness of IRT common item equating.* Paper presented at the annual meeting of the American Educational Research Association (AERA), Chicago, IL.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 22*, 197–206.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating: methods and practices* (2nd ed.). New York: Springer.

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

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

Muraki, E., & Bock, R. D. (1997). PARSCALE 3.0: IRT item analysis and test scoring for rating scale data [Computer software]. Chicago, IL: Scientific Software International.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137–156.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 221–262). New York: Macmillan.

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

Thissen, D., Wainer, H., & Wang, X.-B. (1994). Are tests comprising both multiple-choice and free responses items necessary less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*(2), 113-123.