Research Report

The Use of Two Anchors in Nonequivalent Groups With Anchor Test (NEAT) Equating

Tim Moses

Weiling Deng

Yu-Li Zhang

November 2010

ETS RR-10-23

The Use of Two Anchors in Nonequivalent Groups With Anchor Test (NEAT) Equating

Tim Moses, Weiling Deng, and Yu-Li Zhang
ETS, Princeton, New Jersey

November 2010

Technical Review Editor: Daniel Eignor

Technical Reviewers: Neil Dorans and Insu Paek

Abstract

In the equating literature, a recurring concern is that equating functions that utilize a single anchor to account for examinee groups' nonequivalence are biased when the groups are extremely different and/or when the anchor only weakly measures what the tests measure. Several proposals have been made to address this equating bias by incorporating more than one anchor into nonequivalent groups with anchor test (NEAT) equating functions. These proposals have not been extensively considered or comparatively evaluated. This study evaluates three methods for incorporating more than one anchor into NEAT equating functions, including poststratification, imputation, and propensity score matching. The three methods are studied and compared in two examples. The implications for using the three equating approaches in practice and for developing alternative strategies to incorporate two anchors are discussed.

Key words: NEAT equating, multiple anchors, psychometrics, standardized tests

Table of Contents

List of Figures

Page

List of Tables

Background

Nonequivalent groups with anchor test (NEAT) equating methods are traditionally based on using a single anchor to account for examinee group differences (Braun & Holland, 1982; Kolen & Brennan, 2004; von Davier, Holland, & Thayer, 2004). These equating methods can be extended so that more than one anchor is incorporated. NEAT equating methods based on multiple anchors are potentially useful when the tests being equated measure such broad content that a single anchor may not reflect them, and/or when the examinee group differences are so large that the use of a single anchor to estimate these differences may produce biased equating results (Angoff, 1984; Livingston, 2004; Lord, 1960).

Suggestions have been made for how to incorporate more than one anchor into NEAT equating (Angoff, 1984; Kolen, 1990; Liou, Cheng & Li, 2001; Livingston, Dorans, & Wright, 1990; Skaggs, 1990). These suggestions are fairly diverse and have included direct extensions of traditional single anchor equating methods and more elaborate propensity score matching and missing data imputation methods. Most of these suggestions have not been extensively researched or compared. The purpose of this paper is to develop and compare three proposed approaches for using two anchors to equate tests taken by nonequivalent examinee groups.

This paper begins by describing the traditional NEAT data collection design where a single anchor is administered to both groups and the extension of this design to the situation where two anchors are administered. The assumptions for equating with one and two anchors are described primarily in terms of the poststratification NEAT equating method (von Davier et al., 2004). Poststratification equating provides a useful basis for understanding the three approaches of interest, including two-anchor poststratification (Angoff, 1984), missing data imputation (Liou et al., 2001), and propensity score matching (Livingston, Dorans, & Wright, 1990). After being described, these three approaches are applied in two equating situations where the use of two anchors is expected to produce improved equating results (i.e., equating when anchors do not perfectly reflect the tests and equating when there are large examinee group differences). The final discussion focuses on the implications for using the three equating approaches in practice and for developing alternative strategies to incorporate two anchors into an equating function.

One- and Two-Anchor Nonequivalent Groups With Anchor Test (NEAT) Data Collection Designs and Equating

Data collection and equating using one anchor. For the traditional NEAT design (Table 1), the data are collected as two samples from nonequivalent populations (*P* and *Q*) that take different tests (*X* or *Y*) and the same anchor (*A*). The goal of equating is to produce a conversion from the scores of *X* to the scores of *Y* that eliminates the test forms' difficulty differences. The equating conversion must account for how examinee group differences influence the test scores. One way to address examinee group differences is to use the groups' *A* scores to estimate the *X* and *Y* distributions for a hypothetical single population, *T*, that is a synthetic mixture of *P* and *Q*,

$$T = wP + (1-w)Q,\ \ 0 \le w \le 1. \tag{1}$$

When *X* and *Y* data are available for population *T*, the *X*-to-*Y* equating function can be computed. This equating approach is poststratification equating using a single anchor, *A*.

Table 1

The One-Anchor Nonequivalent Groups With Anchor Test (NEAT) Design

	New test (*X*)	Anchor (*A*)	Old test (*Y*)
New group (*P*)	√	√	
Old group (*Q*)		√	√

The *X* and *Y* distributions in population *T* can be obtained through estimating *T's* bivariate (*X*, *A*) and (*Y*, *A*) distributions using the observed data (Table 1) and making assumptions about the unobserved data. For poststratification equating, the (*X*, *A*) probability distribution in *T*, $\text{Prob}_{\text{T}}(X, A)$, can be estimated as

$$\text{Prob}_{\text{T}}(X, A) = w\text{Prob}_{\text{P}}(X, A) + (1-w)\text{Prob}_{\text{Q}}(X, A). \tag{2}$$

In Equation 2, $\text{Prob}_{\text{P}}(X, A)$ is the joint (*X, A*) probability distribution observed in examinee group *P*. $\text{Prob}_{\text{Q}}(X, A)$ is the joint (*X, A*) probability distribution estimated for examinee group *Q* by assuming that the "*X*-given-*A*" conditional probabilities observed in examinee group *P*,

$\mathrm{Prob_P}(X \mid A)$, are group invariant and can be used to predict *Q*’s joint (*X, A*) probability distribution based on *Q*’s *A* distribution,

$$\begin{aligned}\mathrm{Prob_Q}(X, A) &= \mathrm{Prob_P}(X \mid A)\mathrm{Prob_Q}(A)\\ &= \frac{\mathrm{Prob_P}(X, A)}{\mathrm{Prob_P}(A)}\mathrm{Prob_Q}(A).\end{aligned} \tag{3}$$

Data collection and equating using two anchors. The interest of this study is in extending and comparing approaches such as Equations 2 and 3 to the situation where the *P* and *Q* groups’ data are collected for two anchors, *A1* and *A2* (Table 2). The goal of equating in this situation is the same as for the one-anchor situation: to produce a conversion from the scores of *X* to the scores of *Y* that eliminates the test forms’ difficulty differences. To use two anchors to account for examinee groups’ influences on test scores, population *T’*s trivariate distributions can be estimated by extending Equation 2,

$$\mathrm{Prob_T}(X, A1, A2) = w\mathrm{Prob_P}(X, A1, A2) + (1-w)\mathrm{Prob_Q}(X, A1, A2), \tag{4}$$

by using the observed $\mathrm{Prob_P}(X, A1, A2)$ and making group invariance assumptions that extend Equation 3,

$$\begin{aligned}\mathrm{Prob_Q}(X, A1, A2) &= \mathrm{Prob_P}(X \mid A1, A2)\mathrm{Prob_Q}(A1, A2)\\ &= \frac{\mathrm{Prob_P}(X, A1, A2)}{\mathrm{Prob_P}(A1, A2)}\mathrm{Prob_Q}(A1, A2)\cdot\end{aligned} \tag{5}$$

Table 2

The Two-Anchor Nonequivalent Groups With Anchor Test (NEAT) Design

	New test (*X*)	Anchor (*A1*)	Anchor (*A2*)	Old test (*Y*)
New group (*P*)	√	√	√	
Old group (*Q*)		√	√	√

Two-anchor equating approaches based on Equations 4 and 5 may result in improved equating results relative to one-anchor poststratification equating based on Equations 2 and 3. One-anchor equating results based on Equations 2 and 3 are known to be inaccurate when test–anchor correlations are weak (Livingston, 2004). The two anchors are likely to be more highly correlated with the test scores than one anchor, meaning that the use of two anchors should give a more accurate account of how examinee group differences influence test scores than one anchor. Equating research has shown that equating functions based on the two-anchor group invariance assumption in Equation 5 can be more accurate than equating functions based on the one-anchor group invariance assumption in Equation 3 (Dorans, Liu, & Hammond, 2008). The next section describes three approaches for incorporating two anchors as in Equations 4 and 5.

Three Equating Approaches Involving Two Anchors in the Nonequivalent Groups With Anchor Test (NEAT) Design

The three equating approaches proposed for utilizing two anchors in NEAT equating are poststratification (Angoff, 1984), imputation (Liou et al., 2001), and propensity score matching (Livingston et al., 1990). All three approaches are based on similar assumptions–that the *X* and *Y* distributions that are not directly observed in *P* or *Q* can be estimated using *P* and *Q*'s *A1* and *A2* scores and the conditional relationships observed in *Q* and *P*. All three approaches can be used to implement the major steps of observed score equating (von Davier et al., 2004), including presmoothing, estimating the *X* and *Y* distributions for synthetic group *T*, continuizing the *X* and *Y* distributions, computing linear and curvilinear equating functions, and assessing the equating functions with respect to their standard errors. The general characteristics of the poststratification, imputation, and propensity score matching approaches are described in the following section. Additional details of the approaches are described in the Appendix and in von Davier et al.

Poststratification. The two-anchor poststratification equating method builds directly on the one-anchor poststratification method (Angoff, 1984; Lord, 1975). This approach extends the one-anchor poststratification method (von Davier et al., 2004) by applying loglinear models to presmooth P's trivariate (*X*, *A1*, *A2*) distribution and *Q*'s trivariate (*Y*, *A1*, *A2*) distribution, computing *T*'s *X* and *Y* distributions from the presmoothed distributions using Equations 4 and 5, and computing linear and curvilinear *X*-to-*Y* equating functions and their standard errors based on *T*'s *X* and *Y* distributions. Standard errors can be estimated for the differences between linear and curvilinear equating functions. In contrast to the imputation and propensity score matching approaches, standard errors can also be estimated for the differences between two-anchor and one-anchor equating functions (see Appendix).

Imputation. The application of missing data imputation (Little & Rubin, 1987) to the computation of synthetic population equating functions was considered by Liou and Cheng (1995) and Liou et al. (2001). In this imputation approach, population *T*'s *X* and *Y* distributions are estimated by the imputation of population *Q*'s missing *X* data and population P's missing *Y* data. The assumption of the imputation is that, given the anchor scores, the missing *X* data in *Q* and the missing *Y* data in *P* are missing at random and therefore imputable, based on the anchor scores' observed relationships with the tests (Equations 3 and 5). To impute the missing data, Liou et al. modified Holland and Thayer's (1987, 2000) loglinear presmoothing algorithm so that Equations 4 and 5 are used to repeatedly compute expectations of the missing data in an iterative expectation maximization (EM) algorithm. When the EM algorithm converges, population *T*'s *X* and *Y* distributions can be computed from *T*'s imputed and loglinear presmoothed (*X*, *A1*, *A2*) and (Y, *A1*, *A2*) distributions. The imputed *X* and *Y* distributions for population *T* imply a single group design, meaning that *X*-to-*Y* curvilinear and linear equating functions and their standard errors can be computed as single group equating functions (von Davier et al., 2004).

Propensity score matching. The application of propensity score matching to equating was suggested in Livingston et al. (1990). Rather than use the two anchors in their original form, a single variable (i.e., propensity score) is constructed as the weighted combination of *A1* and *A2* that maximally predicts membership in the examinee administration groups. For example, a logistic regression that predicts membership in *P* can be estimated for all *P* and *Q* examinees' data based on examinees' anchor scores,

$$\text{Propensity}(P \mid A1, A2) = \frac{1}{1+e^{-\beta_0-\beta_1 A1-\beta_2 A2-\beta_3 A1A2}}. \qquad (6)$$

Examinees from *P* and *Q* who have the same $\text{Propensity}(P \mid A1, A2)$ scores are considered equivalent (i.e., matched). Alternative parameterizations of Equation 6 could be used, and Equation 6 can be extended to a large number of anchors and matching variables.

To apply propensity score matching to this study's equating context, the recommended propensity score matching approach from Rosenbaum and Rubin (1984) is followed. In Rosenbaum and Rubin's (1984) proposal, categories of *P* and *Q*'s $\text{Propensity}(P \mid A1, A2)$ scores are formed based on the percentiles of the $\text{Propensity}(P \mid A1, A2)$ scores and the *P* and *Q* examinees who fall into the same category are considered equivalent. In the current study, the categorized propensity scores are used as a single anchor for estimating T's X and Y distributions and equating X-to-Y as in traditional one-anchor poststratification equating, that is, substituting the categorized $\text{Propensity}(P \mid A1, A2)$ scores for A in Equations 2 and 3. Other propensity score matching approaches developed for nonrandomized medical studies propose the use of the uncategorized propensity scores for drawing a small number of individuals from a large control group to match each individual from a small treatment group (Rosenbaum & Rubin, 1985; Rubin & Thomas, 1996; Rubin & Thomas, 2000). The use of categorized propensity scores was followed rather than other propensity score matching approaches because the categorized propensity scores allow for using all available examinee data (not drawing samples from either P or Q) and for defining the equating group of interest as a weighted, synthetic mixture of P and Q's data (i.e., Equations 1, 2, and 4 where w does not have to be set to 0 or 1).

This Study

The discussion from the previous section shows that the poststratification, imputation, and propensity score matching approaches can all be used to incorporate two anchors to estimate a synthetic population's equating function. Perhaps the two-anchor results of the three approaches will be similar, but this has not been extensively considered in prior work. Some research has shown that for situations involving one anchor, poststratification and imputation can produce similar results (Liou & Cheng, 1995). Other work has shown that imputation based on one anchor and one demographic variable can produce results that are similar to those of unsmoothed poststratification equating based on one anchor (Liou et al., 2001). Applying the poststratification and imputation approaches to situations involving two anchors should be useful for determining if these approaches' similarities hold when they are based on the same presmoothing models and when the approaches are used to compare two-anchor curvilinear and linear functions.

The evaluation of the application of propensity score matching to two-anchor equating functions has been less researched and is a more exploratory approach at this point than the poststratification and imputation approaches. It would seem that the estimation and categorization of the propensity scores would introduce inaccuracy into the results. These potential inaccuracies were not described in one study that assessed the potential of propensity score matching for including demographic variables in equating applications (Paek, Liu, & Oh, 2006). By comparing results based on propensity score matching to those obtained from the poststratification and imputation approaches, the current study can provide bases for evaluating the accuracy of equating results based on propensity score matching.

The next two sections apply the poststratification, imputation, and propensity score matching approaches in two examples involving two possible anchors. In both situations, the use of two anchors is expected to improve equating. In the first example, tests are to be equated across extremely different examinee groups. In the second example, composite tests that include multiple-choice and constructed response items and anchors are equated.

First Example: Equating Across Very Different Groups With Internal and External Anchors[1]

In the following example, the two-anchor poststratification, imputation, and propensity score matching approaches are used to produce a conversion for the scores of two forms of a formula-scored, multiple-choice mathematics test. The descriptive statistics for the *P* group's (*X, A1, A2)* scores and the *Q* group's (*Y, A1, A2)* scores are shown in Tables 3 and 4. *A1* is a 16-item anchor that is internal to test forms *X* and *Y* and is the anchor that was intended to be used in the actual *X*-to-*Y* equating. The importance of using two anchors (*A1, A2)* is apparent when the implications of using only anchor *A1* are described. Specifically, *A1*'s correlations with *X* and *Y* (0.90) can be interpreted as not quite as large as would be desired to address the fairly large standardized mean differences between *P* and *Q* (-0.57). Test–anchor correlations that are not as large as desired and large standardized mean differences on the anchor suggest that equating results based only on the use of *A1* could be inaccurate (Livingston, 2004).

Table 3

First Example: Statistics for Test X and Anchors A1 and A2 in P (N_P = 13,639)

	Min. observed & (possible)	Max observed & (possible)	Mean	SD	Skew	Kurtosis	Correlations		
							X	*A1*	*A2*
X	-5 & (-12)	50 & (50)	20.89	10.48	0.09	-0.65	1.00		
A1	-4 & (-4)	16 & (16)	7.47	3.75	-0.06	-0.51	0.90	1.00	
A2	200 & (200)	800 & (800)	609.38	101.27	-0.55	0.05	0.84	0.76	1.00

Table 4

First Example: Statistics for Test Y and Anchors A1 and A2 in Q (N_Q = 11,389)

	Min. observed & (possible)	Max observed & (possible)	Mean	SD	Skew	Kurtosis	Correlations		
							Y	*A1*	*A2*
Y	-8 & (-12)	50 & (50)	28.64	9.72	-0.42	-0.15	1.00		
A1	-4 & (-4)	16 & (16)	9.52	3.45	-0.36	-0.24	0.90	1.00	
A2	200 & (200)	800 & (800)	662.91	83.14	-0.75	1.01	0.80	0.71	1.00

A2 is a second external anchor, an equated and scaled score on a mathematics test from a different testing program. Similar to *A1*, the *P* group is of lower ability than the *Q* group on *A2* (i.e., standardized mean difference = -0.58; Tables 3 and 4). The correlations of *A2* with *X* and *Y* are moderately high (0.84 and 0.80, respectively). The correlations of *A2* with *A1* are also moderately high (0.76 and 0.71), but perhaps not so large as to indicate that the anchors provide redundant information about examinee abilities.

Multiple regression analyses show that the predictions of test scores *X* and *Y* can be improved from squared correlations of about 0.82 with only *A1* to squared correlations of about 0.87 with both *A1* and *A2*. These improved squared correlations with the test scores are descriptive evidence that using both *A1* and *A2* will provide a more accurate account of how examinee differences affect the *X* and *Y* test score differences and will enhance the accuracy of the *X*-to-*Y* results. The statistical implications of using both *A1* and *A2* are assessed directly on the *X*-to-*Y* results.

The Use of *A1* and *A2* in the Equating Process

The following paragraphs describe the results of using poststratification, imputation, and propensity score matching to compute the *X*-to-*Y* scaling function using anchors *A1* and *A2*. The major steps of equating are presented, including presmoothing, the estimation of the *X* and *Y* distributions in synthetic population *T*, and the comparison of linear and curvilinear two-anchor functions in terms of scaled score differences and standard errors. The major interest is using the poststratification, imputation, and propensity score matching approaches to gauge the importance of *A2* for the actual scaling results. For this interest, the three approaches' results based on using both *A1* and *A2* will be compared to those based on using only *A1*.

Presmoothing. Loglinear models were used to presmooth the (*X*, *A1*, *A2*) and (*Y*, *A1*, *A2*) trivariate distributions. For the poststratification method, the loglinear presmoothing is applied to

P's (*X*, *A1*, *A2*) distribution and *Q*'s (*Y*, *A1*, *A2*) distribution (e.g., von Davier et al., 2004). For missing data imputation, the loglinear presmoothing is applied to population *T'*s trivariate (*X*, *A1*, *A2*) and (*Y*, *A1*, *A2*) distributions (Liou & Cheng, 1995; Liou et al., 2001). The loglinear models used to presmooth these four trivariate distributions were based on the same parameterization, fitting five moments in the marginal test distributions (*X* or *Y*), five moments in the *A1* distributions, six moments in the *A2* distributions, and the first and second cross-moments of the joint (test, *A1*), (test, *A2*), and (*A1*, *A2*) distributions and the (*X*, *A1*, *A2*) and (Y,*A1*, *A2*) distributions. These models were selected because they resembled the models actually used to equate these data in practice, and also because evaluations of residuals and model fit indices did not reveal obvious model misspecifications.

For the propensity score matching approach, propensity scores were estimated by predicting *P* and *Q* group membership for all of the *P* and *Q* data, using the logistic regression model in Equation 6. These propensity scores were divided into 10 categories, based on the predicted probabilities' deciles. Sensitivity analyses were conducted to compare these categorized propensity scores to those produced from alternative logistic regression models and categorization schemes. The categorized propensity scores based on model Equation 6 and categories defined in terms of deciles were used because they had high correlations with tests *X* and *Y* (0.92) and because the standardized mean differences between *P* and *Q* for *A1* and *A2* within each of the 10 categories were smaller than with alternative models and categorizations. Bivariate loglinear presmoothing models were used to presmooth *P*'s *(X, CategorizedPropensity)* and *Q*'s *(Y, CategorizedPropensity)* bivariate distributions, fitting five moments in the test distributions, five moments in the categorized propensity score distribution, and the first cross-moment between the test and categorized propensity scores.

Test score distribution estimation in synthetic population T. All of the presmoothing results from the presmoothing step were used to estimate the *X* and *Y* score distributions in the synthetic population, $T = wP + (1 - w)Q,\ w = 0.5$. For the poststratification method, this estimation was done using Equations 4 and 5 and *P* and *Q*'s presmoothed trivariate distributions. For missing data imputation, this estimation was done using the trivariate distributions imputed for population *T* in the presmoothing step. For propensity score matching, this estimation was done using Equations 2 and 3 and *P* and *Q*'s presmoothed bivariate distributions of the tests and the categorized propensity scores.

The descriptive statistics of population *T*'s *X* and *Y* score distributions are shown in Table 5. The *X* and *Y* score distributions are plotted in Figures 1 and 2. The score distributions are essentially identical for the two-anchor poststratification and imputation approaches and are somewhat different for the propensity score matching approach. The *X* and *Y* means in Table 5 indicate that *X* is more difficult than *Y* by 1.9 or 2.1 points.

Table 5

First Example: Synthetic Population Distributions for X and Y, Two-Anchor Matching

	X_{P+Q} PSE	X_{P+Q} Imputation	X_{P+Q} Propensity score matching	Y_{P+Q} PSE	Y_{P+Q} Imputation	Y_{P+Q} Propensity score matching
Mean	23.76	23.76	23.69	25.69	25.69	25.81
SD	10.52	10.51	10.51	10.71	10.71	10.58
Skew	-0.10	-0.10	-0.13	-0.30	-0.30	-0.26
Kurtosis	-0.60	-0.60	-0.61	-0.47	-0.47	-0.47

Note. PSE = poststratification equating.

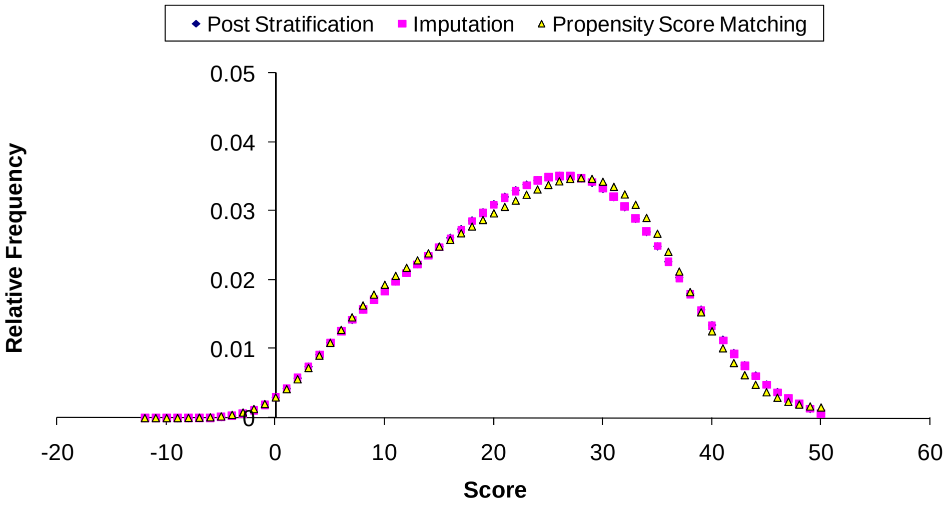

***Figure 1*. First example. Relative frequency distributions of New Form X in Synthetic Population *T* based on the internal and external anchors.**

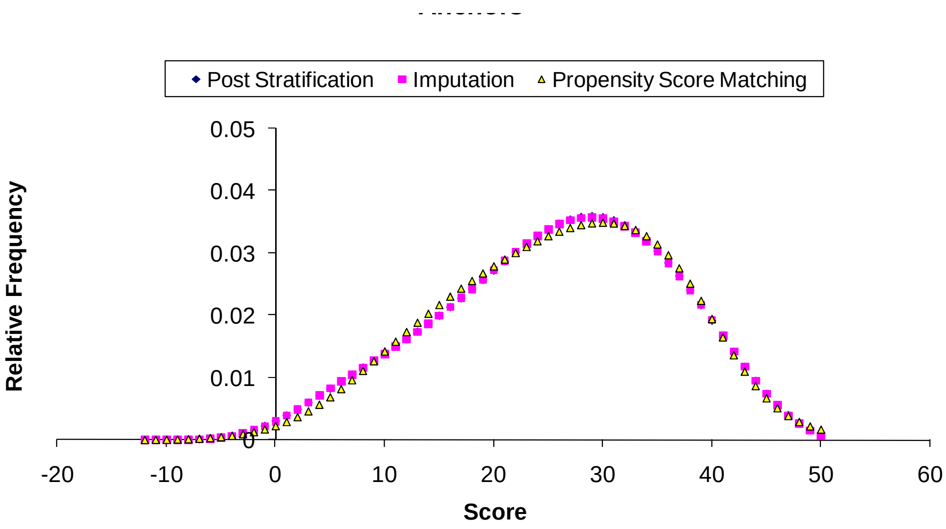

***Figure 2*. First example. Relative frequency distributions of Reference Form Y in Synthetic Population *T* based on the internal and external anchors.**

Test score conversion functions and their evaluation. Several test score conversions based on the poststratification, imputation, and propensity score matching approaches were evaluated. To assess the extent of curvilinearity in the two-anchor conversions, linear and curvilinear *X*-to-*Y* kernel functions were computed from the three approaches' *X* and *Y* distributions estimated in population *T*. The curvilinear versus linear score differences and the +/- 2 *standard errors of these equated differences* (SEEDs) are plotted in Figures 3 (poststratification), 4 (imputation), and 5 (propensity score matching). For all three approaches, the differences between the curvilinear and linear functions exceed two standard errors throughout most of the score range. The largest differences between the curvilinear and linear functions occur at the minimum and maximum scores. These differences based on the propensity score matching approach are somewhat different from those based on the poststratification and imputation approaches. The overall results of Figures 3 through 5 show that, based on all approaches, the curvilinear function should be selected rather than the linear function.

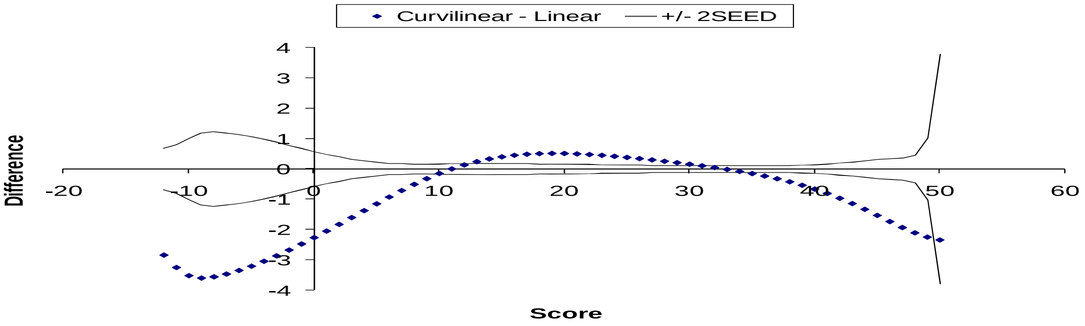

***Figure 3*. First example. Curvilinear vs. linear scaled score differences based on two-anchor poststratification.**

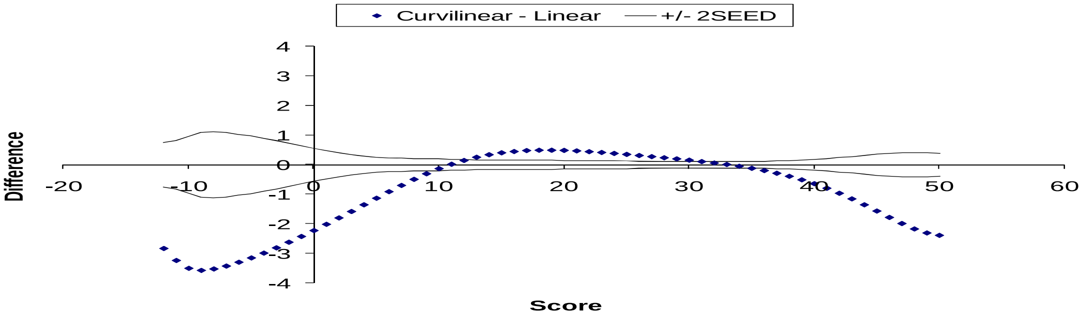

***Figure 4*. First example. Curvilinear vs. linear scaled score differences based on two-anchor imputation.**

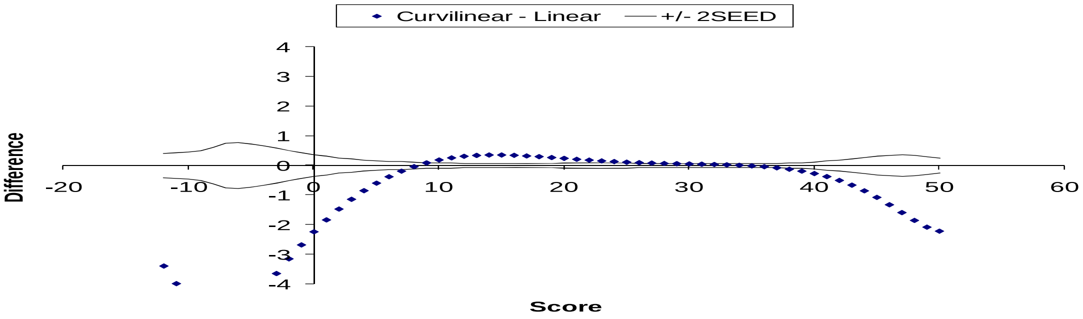

***Figure 5*. First example. Curvilinear vs. linear scaled score differences based on two-anchor propensity score matching.**

A final interest was assessing the implications of using both *A1* and *A2* relative to using only *A1* in the *X*-to-*Y* conversions. To repeat a concern made when introducing this example, the 0.90

correlations between *A1* and tests *X* and *Y* may not be large enough to completely account for the large differences between *P* and *Q*. While the use of *A2* would appear to improve the *X*-to-*Y* conversion because it improves the correlations between the anchors and the tests, the question is what the impact is on the actual *X*-to-*Y* conversion. For this assessment, the three approaches' curvilinear *X*-to-*Y* functions were computed with only *A1* using the previously described presmoothing, test score distribution estimation and equating steps. The two-anchor (*A1* and *A2*) versus one-anchor (*A1* only*)* differences for the approaches are plotted in Figures 6 (poststratification), 7 (imputation), and 8 (propensity score matching). For the poststratification approach, it was possible to compute +/- 2SEED lines for the two-anchor versus one-anchor differences.

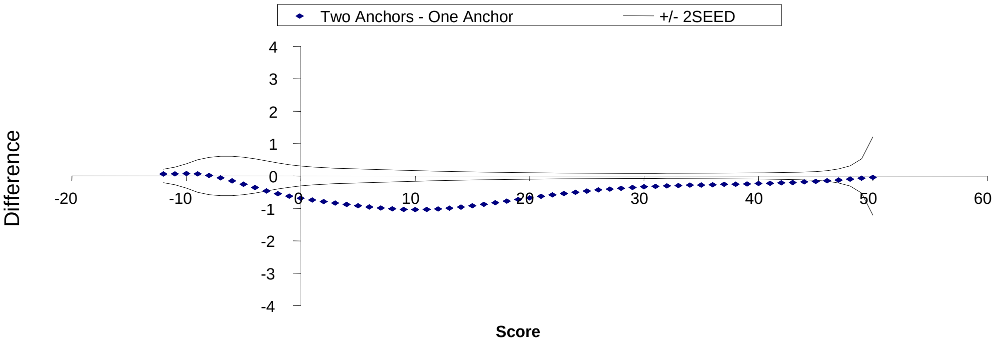

***Figure 6*. First example. Curvilinear two-anchor poststratification vs. curvilinear one-anchor poststratification.**

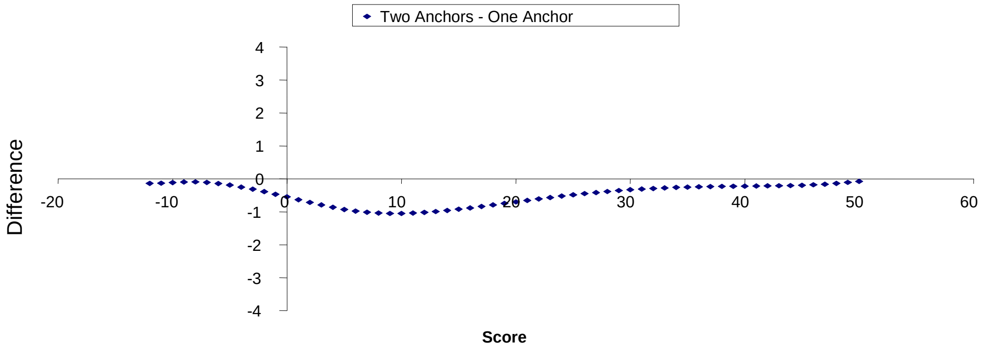

***Figure 7*. First example. Curvilinear two-anchor imputation vs. curvilinear one-anchor imputation.**

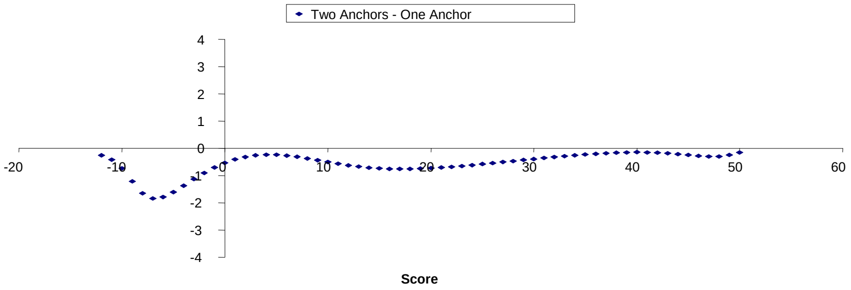

***Figure 8*. First example. Curvilinear two-anchor propensity score matching vs. curvilinear one-anchor propensity score matching.**

The results show that the two-anchor function is lower than the one-anchor equating function for most of the *X* scores. These differences are about 1 score point at their largest and many exceed the +/- 2 SEED lines (Figure 6). The poststratification, imputation, and propensity score matching approaches produce similar results in terms of the magnitude of the two-anchor versus one-anchor score differences. The score differences based on propensity score matching (Figure 8) are visibly different from those of the poststratification (Figure 6) and imputation (Figure 7) approaches.

Second Example: Equating Composite Test Forms With Multiple-Choice and Constructed Response Anchors[2]

The second considered example involves the equating of the composite (multiple-choice plus constructed response) forms of a teacher certification exam. There are two anchors available, where *A1* denotes a 12-item multiple-choice anchor and *A2* denotes a sum of six 2-point constructed response items that is multiplied by 2 when included in the composite. Composite form *X* has a total of 70 points and composite form *Y* has a total of 72 points. To account for human rater drift in the scoring of *A2*, only a small sample of the available examinee data for group *Q* are used ($N_Q = 403$) in the equatings; the data from examinees whose *A2* responses were re-scored when group *P*'s *A2* responses were scored. The result is that in *P*'s trivariate *(X, A1, A2)* distribution *A2* is internal to and contributes to the *X* score while in *Q*'s trivariate *(Y, A1, A2)* distribution, *A2* is external to and does not contribute to the *Y* score[3].

Comparisons of the *P* and *Q* groups' test and anchor scores provide somewhat ambiguous results (Tables 6 and 7), suggesting that the large group of *P* examinees is essentially equivalent to *Q* on *A1* (i.e., the mean differences between *P* and *Q* are 0.02 standardized units) but considerably less able than *Q* on *A2* (i.e., the mean differences between *P* and *Q* are -0.20 standardized units). For both groups, *A2* is more highly correlated with composite forms *X* and *Y* than *A1* (0.79 and 0.69 vs. 0.59 and 0.60). Tables 6 and 7 show that *A1* and *A2* are weakly correlated with each other (0.28 and 0.30), an expected result for composite forms where multiple-choice and constructed response questions likely measure different skills and abilities.

Table 6

Second Example: Statistics for Test X and Anchors A1 and A2 in P (N_P = 2,875)

	Min. observed & (possible)	Max observed & (possible)	Mean	SD	Skew	Kurtosis	Correlations		
							X	*A1*	*A2*
X	13 & (0)	64 & (70)	43.10	8.06	-0.42	0.22	1.00		
A1	1 & (0)	12 & (12)	8.28	2.12	-0.42	-0.08	0.59	1.00	
A2	0 & (0)	24 & (24)	12.39	4.47	-0.32	-0.10	0.79	0.28	1.00

Table 7

Second Example: Statistics for Test Y and Anchors A1 and A2 in Q (N_Q = 403)

	Min. observed & (possible)	Max observed & (possible)	Mean	SD	Skew	Kurtosis	Correlations		
							Y	*A1*	*A2*
Y	16 & (0)	64 & (72)	42.66	9.47	-0.45	0.08	1.00		
A1	2 & (0)	12 & (12)	8.23	2.18	-0.30	-0.53	0.60	1.00	
A2	0 & (0)	24 & (24)	13.35	5.26	-0.56	-0.08	0.69	0.30	1.00

Two questions are particularly important for evaluating the use of two anchors in this situation. The first is whether the multiple-choice anchor, *A1*, can adequately account for examinee group differences on the *X* and *Y* composite forms. The use of multiple-choice anchors to equate composite forms is not the most recommended practice in recent research (Kim, Walker, & McHale, 2008) and is not likely to produce strong equatings (see Note 2, p. 31). However, in several testing programs, there is interest in using multiple-choice anchors due partly to the high costs associated with the use of constructed response anchors. The squared correlations from predicting the *X* and *Y* scores with *A1* are 0.35 (*P*) and 0.36 (*Q*). The squared correlations from predicting the *X* and *Y* scores with both *A1* and *A2* are 0.78 (*P*) and 0.64 (*Q*).

The substantial increases in squared correlations suggest that using both anchors rather than *A1* will result in a significant improvement in the results. The differences in *A1* and *A2'*s indications of *P* and *Q* differences (i.e., *P* and *Q* are nearly equivalent on *A1* but different on *A2*) may also contribute to results that differ when using only *A1* rather than using both *A1* and *A2*. The question of how the results based on *A1* differ from those based on *A1* and *A2* for the poststratification, imputation, and propensity score matching approaches requires evaluation.

A second question that arises when both multiple-choice and constructed response anchors are available is how to make the best use of the two anchors. While the approaches described in this study utilize each anchor to the extent that they jointly correlate with test scores, in practice the two anchors are usually used as a single summed score. The squared correlations from predicting the *X* and *Y* scores with the summed anchor are 0.77 (*P*) and 0.62 (*Q)*. The squared correlations from predicting the *X* and *Y* scores with the separate anchors are 0.78 (*P*) and 0.64 (*Q*). From the perspective of correlations and prediction accuracy, there is potential for slight improvements in equating from using the two anchors separately rather than in a summed form. The implications of the slight improvements in the squared correlations are evaluated in direct comparisons of the equating results based on the two anchors and on the single summed anchor.

The Use of *A1* and A2 in the Equating Process

The following paragraphs describe the results of using poststratification, imputation, and propensity score matching to compute the *X*-to-*Y* equating function using anchors *A1* and *A2*. The discussion focuses on the major steps of equating, including presmoothing, the estimation of the *X* and *Y* distributions in synthetic population *T*, and the comparison of linear and curvilinear two-anchor equating functions with respect to equated score differences and standard errors. Two additional interests are what results based on the poststratification, imputation, and propensity score matching approaches suggest about the importance of *A2* for the actual equating results, and what the approaches suggest about using a single summed anchor rather than the separate use of *A1* and *A2*.

Presmoothing. Loglinear models were used to presmooth the *(X, A1, A2)* and *(Y, A1, A2)* trivariate distributions. For the poststratification method, the loglinear presmoothing is applied to *P'*s *(X, A1, A2)* distribution and *Q'*s *(Y, A1, A2)* distribution (e.g., von Davier et al., 2004). For missing data imputation, the loglinear presmoothing is applied to population *T'*s trivariate *(X,* *A1, A2)* and *(Y, A1, A2)* distributions (Liou & Cheng, 1995; Liou et al., 2001). The loglinear models used to presmooth these four trivariate distributions were based on the same parameterization, fitting five moments in the marginal test distributions (*X* or *Y*), five moments in the *A1* distributions, five moments in the *A2* distributions, and the first cross-moments of the joint (test, *A1)*, (test, *A2),* and (*A1*, *A2)* distributions and the (*X, A1, A2)* and (*Y, A1, A2)* distributions. The models for *P* treated *A2* as an internal anchor and the models for *Q* treated *A2* as an external anchor. These models were selected because they resembled the models actually used to equate these data in practice, and also because evaluations of residuals and model fit indices did not reveal obvious model misspecifications.

For the propensity score matching approach, propensity scores were estimated by predicting *P* and *Q* group membership for all of the *P* and *Q* data, using logistic regression. Several logistic regression models similar to Equation 6 were considered, including those that used linear and quadratic functions of *A1* and *A2* and *A1A2*. Several categorization schemes were considered for the propensity scores produced from the logistic regression models. The logistic regression model and propensity score categorization that produced categorized propensity scores with the highest correlations with the tests (0.67) and also produced the smallest standardized mean differences between *P* and *Q* for *A1* and *A2* within the categories were selected. The propensity scores were obtained by predicting membership in *P* and *Q* for all of the *P* and *Q* group data with the following logistic regression model:

$$\text{Propensity}(P \mid A1, A2) = \frac{1}{1+e^{-\beta_0-\beta_1 A2-\beta_2 A1A2}}. \qquad (7)$$

These propensity scores were divided into 10 categories based on the predicted probabilities' deciles. Bivariate loglinear presmoothing models were used to presmooth *P'*s *(X, CategorizedPropensity)* and *Q'*s *(Y, CategorizedPropensity)* bivariate distributions, fitting five moments in the test distributions, five moments in the categorized propensity score distribution, and the first cross-moment between the test and categorized propensity scores.

Test score distribution estimation in synthetic population *T*. All of the presmoothing results from the presmoothing step were used to estimate the *X* and *Y* score distributions in synthetic population, $T = wP + (1-w)Q,\ w = 0.5$. For the poststratification method, this estimation was done using Equations 4 and 5 and *P* and *Q'*s presmoothed trivariate distributions.

For missing data imputation, this estimation was done using the trivariate distributions imputed for population *T* in the presmoothing step. For propensity score matching, this estimation was done using Equations 2 and 3 and *P* and *Q*'s presmoothed bivariate distributions of the tests and categorized propensity scores.

The descriptive statistics of population *T*'s *X* and *Y* score distributions are shown in Table 8. The score distributions are plotted in Figures 9 and 10. The score distributions are very similar for the two-anchor poststratification and imputation approaches. The *X* and *Y* score distributions based on propensity score matching differ from those of the poststratification and imputation approaches, particularly in their standard deviations. The *X* and *Y* means in Table 5 indicate that *X* is easier than *Y*.

Table 8

Second Example: Synthetic Population Distributions for X and Y, Two-Anchor Matching

	X_{P+Q} PSE	X_{P+Q} Imputation	X_{P+Q} Propensity Score Matching	Y_{P+Q} PSE	Y_{P+Q} Imputation	Y_{P+Q} Propensity score matching
Mean	43.61	43.61	43.72	42.28	42.28	42.07
SD	8.51	8.54	8.16	9.18	9.18	9.46
Skew	-0.44	-0.47	-0.46	-0.43	-0.43	-0.41
Kurtosis	0.15	0.18	0.25	0.10	0.12	0.00

Note. PSE = poststratification equating

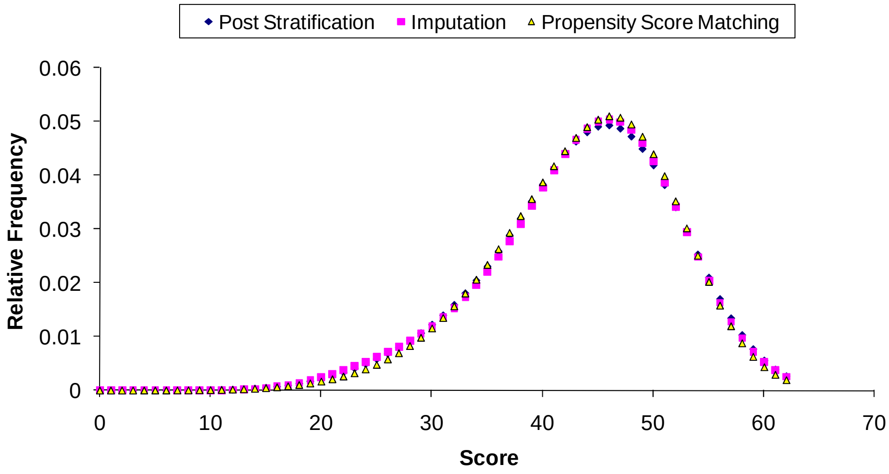

***Figure 9*. Second example. Relative frequency distributions of New Form X in Synthetic Population *T* based on the multiple-choice (MC) and constructed-response (CR) anchor scores.**

Conversion functions and their evaluation. In evaluating the equating results from the poststratification, imputation, and propensity score matching approaches, three comparisons were of interest. The first comparison assesses the curvilinearity of the approaches' two-anchor equating functions. The second comparison addresses the question of whether the multiple-choice anchor, *A1*, can adequately account for examinee group differences on the composite forms with both multiple-choice and constructed response items. The third comparison considers whether the approaches' two-anchor results differ from their results when a summed anchor is used, *A1*+ *A2*.

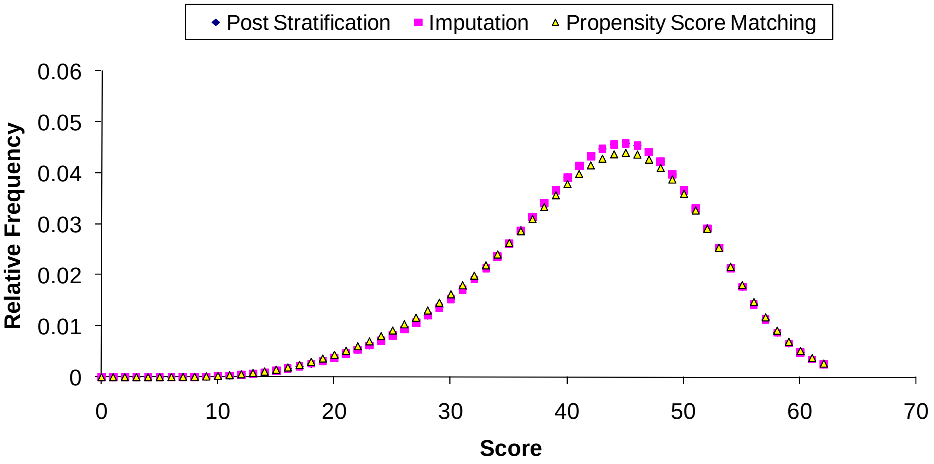

Figure 10**. Second example. Relative frequency distributions of Reference Form Y in Synthetic Population *T* based on the multiple-choice (MC) and constructed-response (CR) anchor scores.**

To evaluate the curvilinearity of the approaches' two-anchor equating functions, linear and curvilinear two-anchor equating functions were computed using the poststratification, imputation, and propensity score matching approaches. The differences between these equating functions are plotted in Figures 11 (poststratification), 12 (imputation), and 13 (propensity score matching). For poststratification and imputation, the curvilinear equating function's differences from the linear equating function are within two SEEDs throughout the score range. For propensity score matching, the differences between the curvilinear and linear equating functions are small and within two SEEDs for all but the lowest scores. The overall results of Figures 11 through 13 show that, based on all three approaches, the linear equating function should be selected rather than the curvilinear equating function.

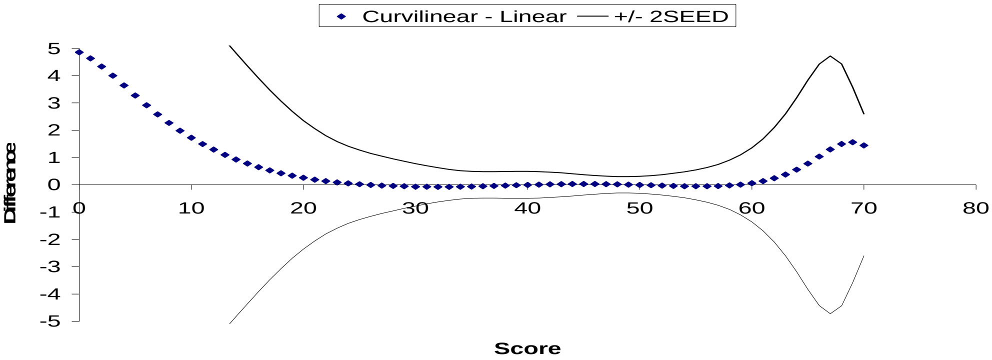

***Figure 11*. Second example. Curvilinear versus linear equated score differences based on two-anchor poststratification.**

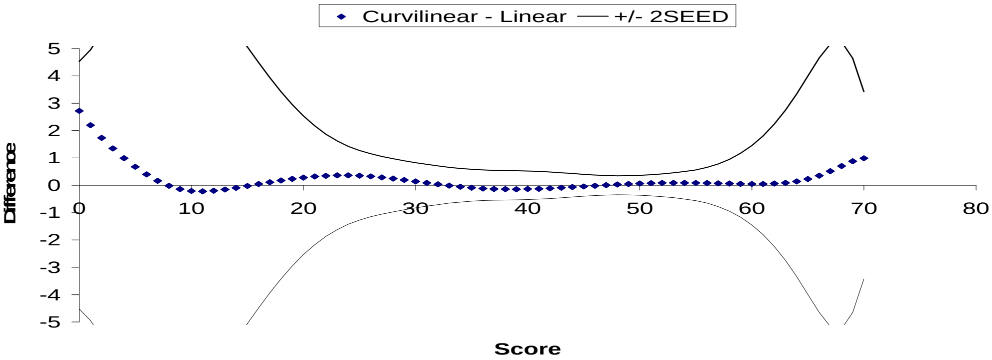

***Figure 12*. Second example. Curvilinear versus linear equated score differences based on two-anchor imputation.**

For the question of whether the multiple-choice anchor *A1* can adequately account for examinee group differences on the *X* and *Y* composite forms, linear equating functions based on using the two anchors, *A1* and *A2*, were computed and compared to linear scaling functions based on only *A1* for the poststratification, imputation, and propensity score matching approaches. The differences between these functions are plotted in Figures 14 (poststratification), 15 (imputation), and 16 (propensity score matching). The results are very similar for the poststratification and imputation approaches, showing that the two-anchor equating function is higher than the one-anchor scaling function for *X* scores below 30, and lower than the one-anchor scaling function for scores between 30 and 70. These differences are about 3 score points, at their largest, and exceed the +/- 2 SEED lines for scores above 40 (Figure 14). For the propensity score matching approach, the two-anchor equating function is lower than the one-anchor scaling function for the whole score range by about 1.5 points (Figure 16). The somewhat different results produced from the propensity score matching approach correspond to differences in the estimates of population *T*'s *X* and *Y* standard deviations when based on propensity score matching rather than on the poststratification and imputation approaches (Table 8).

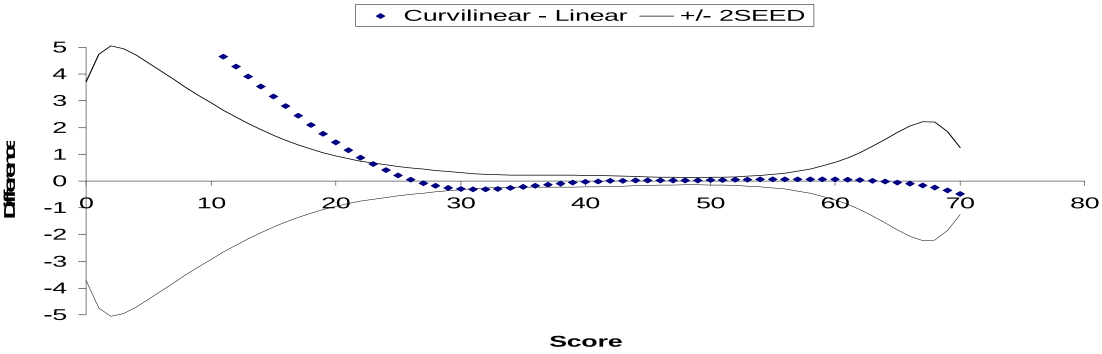

***Figure 13*. Second example. Curvilinear versus linear equated score differences based on two-anchor propensity score matching.**

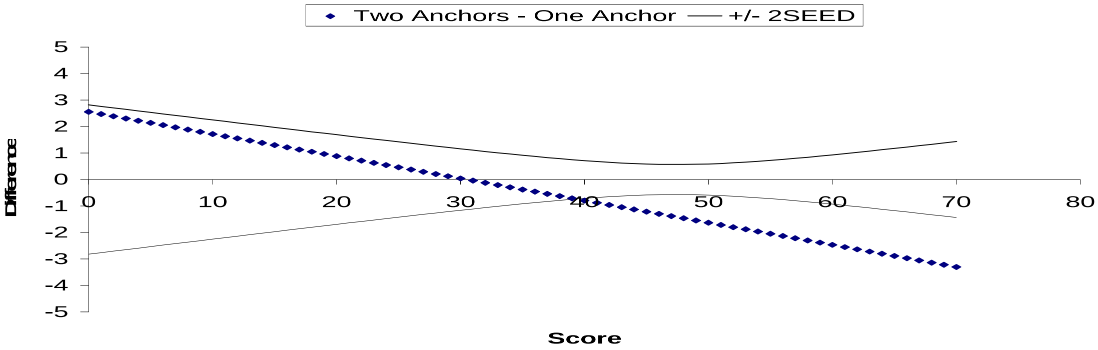

***Figure 14*. Second example. Linear two-anchor poststratification versus linear one-anchor poststratification.**

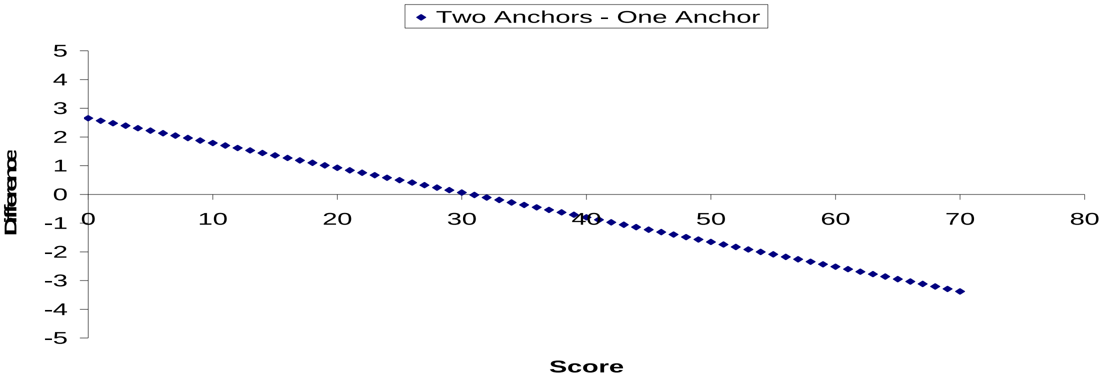

***Figure 15*. Second example. Linear two-anchor imputation versus linear one-anchor imputation.**

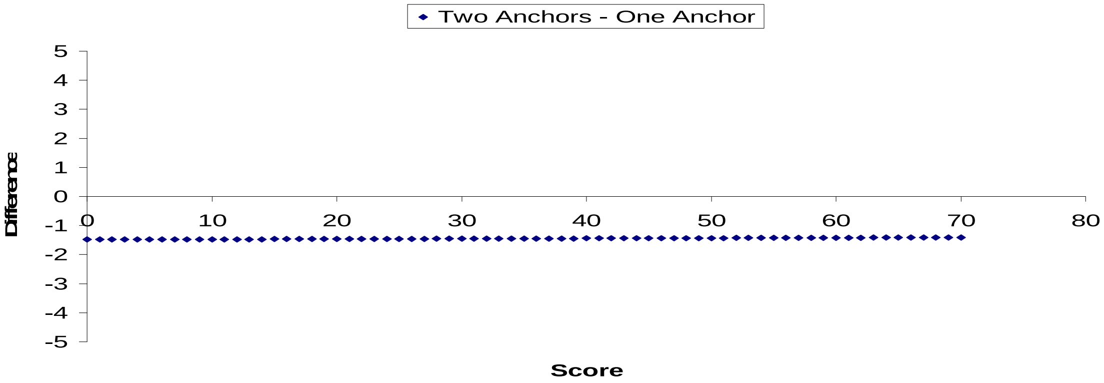

***Figure 16*. Second example. Linear two-anchor propensity score matching versus linear one-anchor propensity score matching.**

For the question of whether the use of *A1* and *A2* in a summed form produces equating results that differ from the results obtained from using *A1* and *A2* jointly, linear equating functions were computed using the sum of *A1* and *A2* as an anchor, and these equating functions were compared to linear equating functions computed using the *A1* and *A2* jointly. The differences between these equating functions are plotted in Figures 17 (poststratification), 18 (imputation), and 19 (propensity score matching). The results show that, for the poststratification and imputation approaches, the two-anchor equating function is almost identical to the summed anchor equating function for the lower range of the *X* scores and is slightly higher than the summed anchor equating function for the upper range of the *X* scores. Although the differences are small (between one third to one half of 1 score point), many exceed the +/- 2 SEED lines (Figure 17).

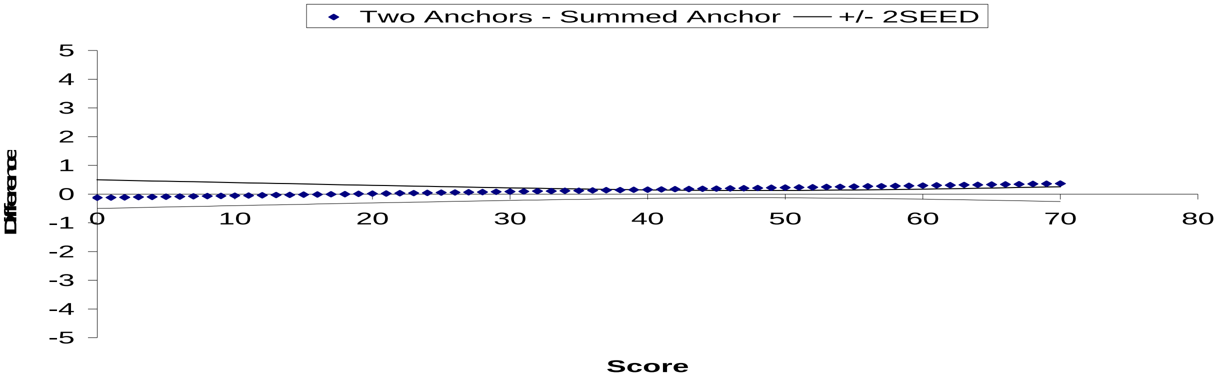

***Figure 17*. Second example. Linear two-anchor poststratification versus linear summed anchor poststratification.**

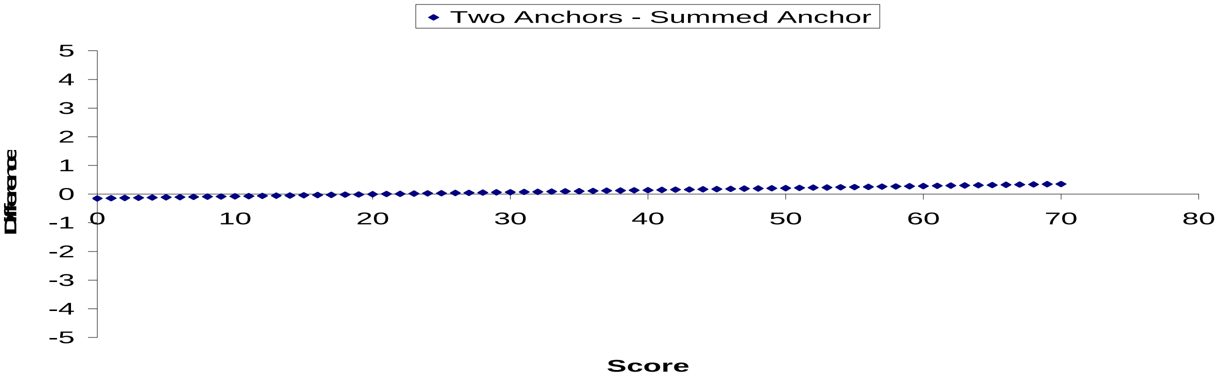

***Figure 18*. Second example. Linear two-anchor imputation versus linear summed anchor imputation.**

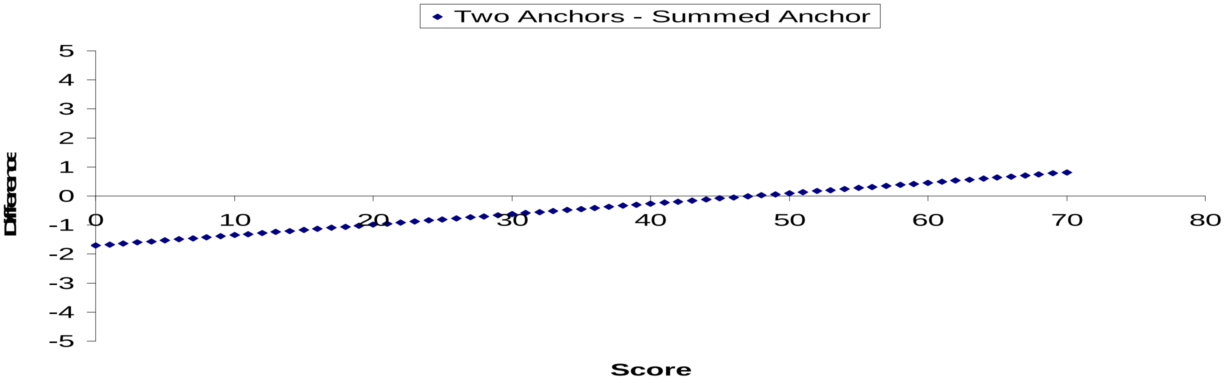

***Figure 19*. Second example. Linear two-anchor propensity score matching versus linear summed anchor propensity score matching.**

For the propensity score matching approach (Figure 19), the two-anchor equating function is lower than the summed anchor equating function for *X* scores below 48 and is higher than the summed anchor equating function for *X* scores above 48. As in the results of comparing equating functions based on only *A1* to those based on *A1* and *A2*, the difference between the propensity score matching results and the poststratification and imputation results correspond to the differences in the approaches' estimates of *T*'s *X* and *Y* standard deviations (Table 8). The slightly different results between the two-anchor and summed anchor equating functions based on the poststratification and imputation approaches are more consistent with the slightly different squared correlations of the summed anchors and tests (0.77 and 0.62) and the joint anchors and tests (0.78 and 0.64) than the larger differences indicated in propensity score matching.

Discussion

The traditional NEAT design's incorporation of a single anchor to address examinee group differences on test scores has been a long-standing concern in equating research (Angoff, 1984; Kolen & Brennan, 2004; Livingston et al., 1990). This concern appears to be most serious when NEAT equating is conducted across examinee groups that are extremely different in ability and/or when the tests measure content that is likely to be broader than the content measured by the anchor. Various proposals have been made for using more than one anchor to account for examinee group differences. Three approaches to incorporating two anchors in equating have been proposed but not extensively studied or compared: poststratification, missing data imputation, and propensity score matching (Angoff, 1984; Liou & Cheng, 1995; Liou et al., 2001; Livingston et al., 1990). This paper described how the approaches could be used to implement the assumptions and equating of the poststratification method. The three approaches were demonstrated in two situations where the use of two anchors would appear to be warranted.

The results of this study's applications showed that the poststratification, imputation, and propensity score matching approaches could all be used in similar ways to incorporate two anchors and compute equating and scaling functions. The poststratification and imputation methods produced results that were essentially identical for both examples of this study, a finding that was not emphasized in prior evaluations of imputation applications in equating (Liou & Cheng, 1995; Liou et al., 2001). Propensity score matching produced results that were similar to the results of the other approaches for the first example, but somewhat different results for the second example. The next section describes the issues involved in the three approaches in more detail.

Two-Anchor Equating Methods

As shown in this paper, the missing data imputation approach is a more limited version of the poststratification approach. Imputation uses the same loglinear presmoothing methods as used in poststratification equating, but incorporates the poststratification assumptions directly into the presmoothing to impute test and anchor distributions for synthetic population *T*. The result is a more complex equating process that produces results that are similar to those of poststratification equating. One difficulty with imputation is that the standard errors and SEEDs tend to be inaccurate due to difficulties with incorporating sample sizes into standard error formulas that reflect both the complete and the imputed data. The unresolved question is how to simultaneously account for complete data on *A1* and *A2,* but incomplete data on *X* and *Y*.

The application of propensity score matching to two-anchor equating requires logistic regression models and categorizations that add ambiguity and probably inaccuracy to its results. For this study's second example, where these modeling and categorization decisions were complex due to each anchor giving inconsistent information about the examinee groups, the propensity score matching produced *X* and *Y* distributions on *T* with standard deviations that were different from those of the poststratification and imputation approaches (Table 8). The resulting linear functions based on propensity score matching had considerably different slopes from those of the other methods, and the differences between equating functions reflected these different slopes (Figures 16 and 19). Some follow-up simulations showed that the standard errors and SEEDs based on propensity score matching were inaccurate due to uncertainty in how to incorporate the influences of the categorization decisions in the equating variability estimates. Approaches other than this study's use of categorized propensity scores would have likely produced more closely matching examinees from the *P* and *Q* groups (Rosenbaum & Rubin, 1985; Rubin & Thomas, 1996; Rubin & Thomas, 2000), but these approaches involve discarding data and matching the individuals of one group to those in the other group rather than estimating score distributions for a synthetic mixture of the *P* and *Q* groups' data.

The problems with propensity score matching prompt the question of whether it should be used at all in equating and scaling situations. While the results of this study discourage the use of propensity score matching for situations involving one or two anchors, propensity score matching could still be valuable when there are more than two possible anchors. For example, a situation could be encountered where there are several available anchors, including examinees' pass/fail decisions, grade point averages, scaled item response theory (IRT) thetas, and more than one internal anchor. For this situation, the loglinear presmoothing models used by the poststratification and imputation approaches would be exceedingly large and much more unwieldy than the logistic regression modeling used by propensity score matching.

The overall results of this study suggest that when using two anchors, the poststratification approach works better than the imputation and propensity score matching approaches. Poststratification is the most flexible approach in terms of the SEEDs that can be produced for evaluating competing equating and scaling functions. Some follow-up simulations have shown that the standard errors and SEEDs for poststratification functions are more accurate than those of the imputation and propensity score matching approaches. To the extent one-anchor poststratification equating functions are biased due to test–anchor correlations that are too small (Livingston, 2004), two-anchor poststratification improves accuracy because the two anchors likely have larger correlations with the tests than one anchor. A question for further study is whether approaches can be developed for incorporating multiple anchors outside of the poststratification framework.

Other Two-Anchor Possibilities

In the situation of NEAT equating with one anchor, the chained equating approach may be more accurate than poststratification because it ignores the test–anchor correlations (Livingston, 2004; Livingston et al., 1990). The incorporation of two anchors into chained equating is less straightforward than for poststratification equating. Chained equating is based on the marginal distributions of the tests and anchors involved, so that there are several possible chained equating functions based on each anchor that can potentially be used. If multiple anchors are used in chained equating, there is a different chained equating function corresponding to each possible order of the anchors in the equating chain. Reasonable ways to convert two anchors into a summed, single anchor for chained equating exist in some situations (this study's second example) but not in others (this study's first example). Perhaps the way to use multiple anchors that have no obvious way of being summed in chained equating is through converting them into a single propensity score.

Another alternative to poststratification equating might be to poststratify on the joint distribution of one or both of the anchors' expected true scores rather than on the joint distribution of their observed scores. This strategy would be analogous to the Levine observed equating approach. Some work has been done to conduct loglinear presmoothing on observed and expected true score distributions (Chen & Holland, in preparation). Such an approach could expand on the flexibility of this study's two-anchor poststratification approach. In particular, some potential anchors may be considered as directly measuring what the tests measure, while others are predictors of examinee group membership but not necessarily congeneric with the tests (Wright & Dorans, 1993). The ability to presmooth the expected true score distributions of the first type of anchors and presmooth the observed score distributions of the second type of anchors could result in the most appropriate treatment of multiple anchors in NEAT equating.

Terminology Implications

An important question in this paper's discussion of the first and second examples was whether the two-anchor results produced for the first example and the one-anchor results produced for the second example constituted equatings or whether they were more appropriately referred to as *scalings*. Although interchangeable scores was an equating goal that was common to both examples, *scaling* and *equating* labels were used to distinguish results (see Notes 1 and 2, p. 30). These labels were used primarily based on equating practitioners' judgments about how representative the anchors were of the tests involved. The equating literature's syntheses and studies are not completely clear on what label is most appropriate for a score conversion that utilizes one or more anchors that represent the tests to varying degrees (Holland & Dorans, 2006; Dorans, Liu, & Hammond, 2008; Wright & Dorans, 1993). A need for future discussions of equating and scaling terminology is clarification and criteria for what types of anchors are likely to produce scalings and equatings.

References

Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Princeton, NJ: ETS.

Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York, NY: Academic Press.

Chen, H., & Holland, P. W. (2009). *Nonlinear Levine observed score equating. Or is it?* Manuscript in preparation.

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*(1), 81–97.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger Publishers.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (ETS Research Rep. No. RR-87-31). Princeton, NJ: ETS.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133–183.

Kim, S., Walker, M. E., & McHale, F. (2008). *Equating of mixed-format tests in large-scale assessments* (ETS Research Rep. No. RR-08-26). Princeton, NJ: ETS.

Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education, 3*(1), 97–104.

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

Liou, M., & Cheng, P. E. (1995). Equipercentile equating via data-imputation techniques. *Psychometrika, 60*(1), 119–136.

Liou, M., Cheng, P. E., & Li, M. (2001). Estimating comparable scores using surrogate variables. *Applied Measurement in Education, 25*(2), 197–207.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.

Livingston, S. A. (2004). *Equating test scores (without IRT*). Princeton, NJ: ETS.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*(1), 73–95.

Lord, F. M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55*, 307–321.

Lord, F. M. (1975). Automated hypothesis tests and standard errors for nonstandard problems. *The American Statistician, 29*(1), 56–59.

Paek, I., Liu, J., & Oh, H. J. (2006). *Investigation of propensity score matching on linear/nonlinear equating method for the P/N/NMSQT* (Statistical Rep. No. SR-2006-55). Princeton, NJ: ETS.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*, 249–264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95*(450), 573–585.

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education, 3*(1), 105–113.

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

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (ETS Research Rep. No. RR-93-4). Princeton, NJ: ETS.

Notes

[1] Practitioners from the testing program where this first example's data came from have used internal anchors for some test score conversions and external anchors for others. These practitioners refer to the test score conversions produced from internal anchors as equatings. Conversions based on external anchors are referred to as scalings rather than equatings because the external anchor is not completely representative of the tests. Though the general goal of this first example is to produce an equating function for two tests, the two anchor results produced from using both the internal and external anchors will be referred to as scalings rather than as equatings.

[2] Equating practitioners often regard conversions of composite scores using anchors composed of multiple-choice and constructed response items as equatings and conversions using anchors composed of only multiple-choice items as scalings. Though the general goal of this second example is to produce an equating for two composite forms, the results produced from using both multiple-choice and constructed response items in the anchor will be referred to as equatings and the results produced from using only multiple-choice items in the anchor will be referred to as scalings.

[3] In the equating sample, *Q*'s *A2* responses on the reference form were rescored by the same pool of readers who scored *P*'s *A2* responses on the new form. Therefore, for *Q*, *A2* is external to total score *Y* because *Y* score came from the operational scoring at the time when reference form was administered; the part score on *A2* that contributed to *Y* was given by another set of readers.

Appendix
SEEDs for Two-Anchor Equating and Scaling Functions

This appendix provides an overview of the standard errors of equating (or scaling) differences (SEEDs) that are used to evaluate the two anchor functions. The SEED has the general form (von Davier, Holland, & Thayer, 2004)

$$\mathrm{SEED}_{\mathrm{y}}(\mathrm{x}) = \left\| \mathbf{J}_{\mathbf{e1}}\mathbf{J}_{\mathbf{DF1}}\mathbf{C} - \mathbf{J}_{\mathbf{e2}}\mathbf{J}_{\mathbf{DF2}}\mathbf{C} \right\|, \qquad \text{(A1)}$$

where the $\mathbf{J}_{\mathbf{DF}}\mathbf{C}$ terms denote the transformation of the loglinear presmoothed distributions into the *X* and *Y* score probabilities in synthetic population *T,* and the $\mathbf{J}_{\mathbf{e}}$ terms denote vectors of the derivatives of the equating (or scaling) functions with respect to the *X* and *Y* score probabilities in synthetic population *T*. Different SEEDs can be calculated using Equation *A1*.

Approaches' Loglinear Presmoothing Models and Their C terms

The poststratification, imputation, and propensity score matching approaches all rely on loglinear models for their presmoothing. For the poststratification and imputation approaches, the loglinear models used are trivariate models of the (*X, A1, A2*) and (*Y, A1, A2*) distributions in the data provided by *P* and *Q* (poststratification) or directly in the synthetic *T* distribution (imputation). To illustrate, the loglinear presmoothing model of the (*X, A1, A2*) distribution has the following form,

$$\log_e(p_{jlm}) = \alpha + \sum_{i=1}^{I}\beta_{xi}(x_j)^i + \sum_{h=1}^{H}\beta_{a1h}(a1_l)^h + \sum_{g=1}^{G}\beta_{a2g}(a2_m)^g$$
$$+ \sum_{f=0}^{F}\sum_{e=0}^{E}\sum_{d=0}^{D}\beta_{fed}(x_j)^f(a1_l)^e(a2_m)^d \qquad , \qquad \text{(A2)}$$

where p_{jlm} is the probability of the score x_j, $A1_l$, $a2_m$ (i.e., score x_j on test *X*, score $A1_l$ on test *A1,* and score $a2_m$ on test *A2*) and the α and β 's are free parameters estimated in the model-fitting process. The propensity score matching approach also uses a loglinear model for its presmoothing, but on the bivariate distribution of the tests and the categorized propensity score.

To illustrate, the loglinear presmoothing model of the $(X, \text{CAT Propensity}(P \mid A1, A2))$ distribution has the following form,

$$\log_e(p_{jl}) = \alpha + \sum_{i=1}^{I} \beta_{xi}(x_j)^i + \sum_{h=1}^{H} \beta_{ah}(Cat\text{Propensity}(P \mid A1, A2)_l)^h$$
$$+\sum_{g=1}^{G} \sum_{f=1}^{F} \beta_{gf}(x_j)^g (Cat\text{Propensity}(P \mid A1, A2)_l)^f \quad . \qquad \text{(A3)}$$

The $\mathbf{C}$ matrices for loglinear models such as Equations *A2* and A3 have the same number of columns as the parameters in the loglinear models and can be computed as described in Holland and Thayer (2000).

Approaches' $\mathbf{J_{DF}}$ Terms

The $\mathbf{J_{DF}}$ matrices are based on how the $\mathbf{C}$ matrices from the poststratification, imputation, and propensity score matching approaches are transformed into *X* and *Y* probability distributions for synthetic population *T*. The transformations are conceptually described in Equations 2 through 4 in this paper and are described in more detail in von Davier et al. (2004). The one-anchor $\mathbf{J_{DF}}$ that is directly used in propensity score matching is specifically described in von Davier et al. The two-anchor $\mathbf{J_{DF}}$ that is used in two-anchor poststratification extends the computations of the one-anchor $\mathbf{J_{DF}}$ by using the joint probabilities of the two anchors, *A1* and *A2*, rather than the univariate probabilities of the single anchor, *A*. The imputation $\mathbf{J_{DF}}$ is based on the single group design described in von Davier et al. because imputation produces *X* and *Y* distributions for a single group, *T*, in its loglinear presmoothing step.

$\mathbf{J_e}$ Terms

The $\mathbf{J_e}$ terms are the derivatives of the equating (or scaling) functions with respect to the *X* and *Y* score probabilities. Throughout this paper, the $\mathbf{J_e}$ terms pertain to linear and curvilinear kernel functions and are similarly computed for the poststratification, imputation, and propensity score matching approaches. The details for computing $\mathbf{J_e}$ are in von Davier et al. (2004).

SEEDs and Their $\mathbf{J_e J_{DF} C}$ Terms

SEEDs can be calculated as in Equation A1 based on the two functions' $\mathbf{J_e J_{DF} C}$ terms. The poststratification, imputation, and propensity score matching approaches can all be used to compute SEEDs to compare linear and curvilinear functions that differ only in their $\mathbf{J_e}$ values.

The poststratification approach is somewhat more general than imputation and propensity score matching, so that SEEDs for additional function comparisons can be computed. Because the loglinear presmoothing models and the estimation of T's distributions are done in separate steps in poststratification equating, it is possible to add additional conversions of the loglinear presmoothing results. Two such conversions of interest involve transforming the trivariate loglinear presmoothed distributions into bivariate distributions, by either aggregating the presmoothed results over one anchor or over the sum of the scores on the two anchors. When these conversions are applied to the loglinear presmoothed results and its $\mathbf{C}$ matrices and are used to compute one-anchor results, SEEDs for evaluating two-anchor versus one-anchor or two-anchor versus a summed anchor functions can be computed.