Research Report

A Study of Frequency Estimation Equipercentile Equating When There Are Large Ability Differences

Hongwen Guo

Hyeonjoo J. Oh

December 2009

ETS RR-09-45

A Study of Frequency Estimation Equipercentile Equating When There Are Large Ability Differences

Hongwen Guo and Hyeonjoo J. Oh
ETS, Princeton, New Jersey

December 2009

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

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

Abstract

In operational equating, frequency estimation (FE) equipercentile equating is often excluded from consideration when the old and new groups have a large ability difference. This convention may, in some instances, cause the exclusion of one competitive equating method from the set of methods under consideration. In this report, we study the possibility of using the FE equating method when the group ability difference is large. Three situations are identified: (a) a situation in which neither the two forms nor the observed conditional distributions are very different so that the FE equating assumptions are likely to hold, and FE equating is recommended; (b) a situation in which forms are not very different, but the observed conditional distributions are different, so that FE equating is not recommended; and (c) a situation in which forms are very different, but the observed conditional distributions are not different, so that FE equating is not recommended. Statistical analysis procedures for comparing distributions are provided. An application of equating to a large-scale admission test is discussed to illustrate the proposed methodology.

Key words: FE equipercentile equating, conditional distribution, form distribution, comparison

Acknowledgments

The authors would like to thank Drs. Dan Eignor, Skip Livingston, and Jinghua Liu for their helpful comments and suggestions. They are also thankful to Ms. Miriam Feigenbaum for collecting data and to Dr. Eignor and Ms. Linda DeLauro for their editorial help.

1 Introduction

The frequency estimation (FE) equipercentile equating method, described by Angoff (1971) and Braun and Holland (1982), is one of a small number of observed equating methods that can be used with the common-item nonequivalent group design or the nonequivalent groups with anchor test (NEAT) design when the new and old test score distributions are found to have a curvilinear relationship. The FE equipercentile equating method assumes the invariance of conditional score distributions of the focus, or new group (P), and the reference, or old group (Q), on the anchor (the common items). Specifically, let X and Y be the new and old test forms, and let x and y be the test scores on X and Y, respectively. Let V and v be the anchor test and score, and let $h_P(v)$ and $h_Q(v)$ be the anchor score distributions of P and Q, respectively. In the NEAT design, population P takes form X and anchor V; population Q takes form Y and anchor V. The observable $f_P(x|v)$ stands for the conditional distribution of x given an anchor score v for population P, and $g_Q(y|v)$ stands for the conditional distribution of y given an anchor score v for population Q. The FE equipercentile equating makes the assumptions that, for all v,

$$f_P(x|v) = f_Q(x|v) \quad \text{and} \quad g_P(y|v) = g_Q(y|v), \tag{1}$$

where $f_Q(x|v)$ and $g_P(y|v)$ are unobservable.

Because the preceding FE equipercentile equating assumptions are untestable, it can be difficult to decide whether FE equating should be considered in a particular equating situation. Many researchers have been exploring the criteria used in making equating decisions by using comparison methods under the expectation that they can guide practitioners in their choice of equating method (a survey can be found in Harris & Crouse, 1993). Other researchers have tried to investigate equating assumptions from a theoretical point of view or through special data designs. Von Davier, Holland, and Thayer (2004) showed, on a theoretical level, that FE equating and chained equipercentile (CE) equating are identical under certain ideal conditions. One ideal condition is when the two groups are of the same ability, that is, when $h_P(v) = h_Q(v)$. In this case, the assumptions of (1) are true. Both FE equipercentile equating and CE equating yield the same equating function,

as discussed in Theorem 2 of von Davier et al. (2004). FE equipercentile equating will work perfectly to adjust for form difficulty. As might be expected, FE equipercentile equating and CE equating work in the same way for equipercentile equating under the equivalent groups (EG) design. Another ideal condition is when $X = Y = V$ with perfect correlation, as discussed in Theorem 3 of von Davier et al. (2004). In this case, the equating functions of both FE and CE are the identity function. Holland, von Davier, Sinharay, and Han (2008) studied the untestable assumptions of FE equating and CE equating. The authors compared predictions from both FE equipercentile equating and CE equating assumptions based on a special data set and an elaborate design with a true criterion. Their results indicated that both FE equating and CE equating make very similar predictions. Overall, as mentioned by Kolen and Brennan (2004, p. 298), FE equipercentile equating might be preferred when groups are similar because the FE equipercentile equating assumptions are most likely to be true in that situation.

Is it possible to use FE equipercentile equating even when group abilities are significantly different? In this study, we investigate the assumptions of FE equipercentile equating based on observed score distributions and item difficulty distributions for the forms, and we try to verify indirectly when the FE equating assumptions are true and when they are incorrect. This verification will then guide practitioners in deciding whether FE equating is applicable when the ability difference between groups is significant.

In section 2, we generalize the results of Theorem 3 of von Davier et al. (2004) for obtaining an identity equating function; we then discuss theoretically whether FE equipercentile equating assumptions are true under different situations and their practical implications. While our focus is on the FE equipercentile equating method, the CE equating method is discussed in this section for the section to be comparable to the results of von Davier et al. (2004). In section 3, we suggest statistical methods to test for homogeneity of probability distributions, which can help verify FE equipercentile equating assumptions indirectly in practice. The difficulty in choosing an equating method lies in the fact that it is impossible to test some of the crucial assumptions of each equating method. When the two groups' abilities are similar, that is, $h_P(v) \sim h_Q(v)$, the assumptions in (1) are most

likely to be true, and practitioners feel safe in choosing the FE equating method. On the other hand, when $h_P(v)$ and $h_Q(v)$ are significantly different, practitioners are reluctant to use FE equating in their operational work, even though the FE equating assumptions might be true under some circumstances. If so, a competitive equating method (i.e., FE equating) might easily be excluded from consideration, as is shown in our illustrative example in section 4. Section 5 draws conclusions. Note again that we assume throughout the report that the group ability difference is large.

2 When Are Frequency Estimation Assumptions True, or Likely to Be True?

We first generalize the results of Theorem 3 of von Davier et al. (2004), in which it is assumed that $X = Y = V$, so that an examinee has the same score on the three tests. It is obvious that the FE equipercentile equating assumptions are true and that the equating function is the identity function no matter how different P and Q are. The assumption $X = Y = V$ is very stringent. We argue that to obtain the same conclusion, the anchor test need not be the same as the test forms. Denote the synthetic group as

$$T = wP + (1 - w)Q, \qquad w \in [0, 1].$$

Let f_T, g_T, and h_T be the corresponding score density functions for the synthetic group and F_T, G_T, and H_T be the corresponding cumulative distribution functions on tests X, Y, and V, respectively. We assume that the cumulative distributions have been made continuous, or *continuized*, so that the inverse functions exist.

Theorem 1. If test forms $X = Y$, and the test score and anchor score satisfy $v = l(x)$ for an examinee, where $l(\cdot)$ is a monotonic function, then the equating functions of FE and CE are both the identity function.

Proof. When $X = Y$, $f_P(x|v) = f_Q(x|v)$ because they are 1 when $v = l(x)$, or 0 otherwise. So are $g_P(y|v) = g_Q(y|v)$. Therefore, as in Theorem 3 of von Davier et al. (2004), for $v = l(x)$,

$$f_T(x) = \sum_v f_P(x|v)h_T(v) = h_T(v)$$

$$g_T(y) = \sum_v g_Q(x|v)h_T(v) = h_T(v),$$

where $h_T(v) = wh_P(v) + (1-w)h_Q(v)$. Thus the equating function of FE is the identity function.

Note that $v = l(x)$. Then $H_P(v) = H_P \circ l(x) = F_P(x)$ and $H_Q(v) = H_Q \circ l(y)$ $= G_Q(y)$. Therefore the equated score $e_y(x)$ on form Y for the score x on form X by CE equating is

$$\begin{aligned} e_Y(x) = G_Q^{-1} \circ H_Q \circ H_P^{-1} \circ F_P(x) &= G_Q^{-1} \circ H_Q \circ H_P^{-1} \circ H_P(v) \\ &= G_Q^{-1} \circ H_Q(v) = y = x, \end{aligned}$$

where $v = l(x) = l(y)$. The theorem is obtained.

Note that when $l(x) = x$, one obtains Theorem 3 of von Davier et al. (2004) from Theorem 1 of this report. The function l can be either a linear or a nonlinear function, but the crucial property of the conditions in Theorem 1 is that the anchor score and the form score can be uniquely determined by each other, which, in reality, is usually impossible.

Also note that $v = l(x)$ is a sufficient condition for the FE equating assumptions. The following result applies to more general settings.

Theorem 2. If $X = Y$, and the FE equating assumptions (1) are true, then the equating function of FE is the identity function.

Proof. By FE assumptions, $f_P(x|v) = f_Q(x|v)$ and $g_P(y|v) = g_Q(y|v)$ so that the observed conditional density functions $f_P(x|v)$ and $g_Q(y|v)$ are the same because $X = Y$. Note that the score distributions of the synthetic group on the two forms are

$$f_T(x) = \sum_v f_P(x|v)h_T(v)$$

$$g_T(y) = \sum_v g_Q(y|v)h_T(v),$$

which are identical functions. Thus we obtain $e_Y(x) = x$.

For theoretical purposes, we list several situations in which we can tell whether the FE equating assumptions (1) are true by using observed data when the ability difference between P and Q is large:

1. When we observe that the two forms $X = Y$,
   (a) if we further observe that $f_P(x|v)$ and $g_Q(y|v)$ are the same, then the assumptions (1) hold because $f_P(x|v) = g_P(y|v)$ and $g_Q(y|v) = f_Q(x|v)$ when $X = Y$. Therefore $f_P(x|v) = f_Q(x|v)$ and $g_Q(y|v) = g_P(y|v)$.
   (b) if we further observe that $f_P(x|v)$ and $g_Q(y|v)$ are different, then the assumptions are wrong for the same reason as given in (1a).
2. When we observe that the two forms X and Y are different,
   (a) if we further observe that $f_P(x|v)$ and $g_Q(y|v)$ are the same, then the assumptions (1) are wrong. Otherwise, if the assumptions are true, then $f_P(x|v)$ and $g_P(y|v)$ are the same; that is, the score distributions on X and Y are the same for P (and for Q), which is contradictory to the fact that $X \neq Y$.
   (b) if we further observe that $f_P(x|v)$ and $g_Q(y|v)$ are different, we cannot determine whether the assumptions are valid. Other evidence has to be collected.

We can generalize Theorem 2 under the following conventions: If two test forms X and Y are statistically the same, that is, the two test forms have the same distributions of item difficulty statistics, then a population will have the same score distributions on the two test forms. Therefore, when two test forms are statistically the same, Theorem 2 holds.

Embedding the preceding discussion in a practical setting, we conclude that when the ability difference is large, FE equating is recommended only when we observe that the two forms are very close in their distributions of item difficulties, and $f_P(x|v)$ and $g_Q(y|v)$ are the same, because the FE equating assumptions are likely to be true. In this case, the FE equating function is expected to be close to the identity function from Theorem 2. FE equating is not appropriate when we observe that the two forms are very different but $f_P(x|v)$ and $g_Q(y|v)$ are the same, or when the two forms are very close but $f_P(x|v)$ and $g_Q(y|v)$ are different.

Of course, there are also issues related to whether equating should be done when the two forms are really close in form difficulty (see, e.g., Dorans & Lawrence, 1990; Hanson, 1996). However, as Kolen and Brennan (2004, p. 296) pointed out, Hanson's (1996)

approach only considered random sampling error and not systematic error. When the two forms are similar and the observed conditional distributions are the same under the NEAT design, the FE equipercentile equating assumptions are likely to be true. Then FE equating is recommended, instead of no equating, in the hope that FE equating can adjust for some systematic error.

Is it possible that the two groups are different, but the observed conditional distributions are the same? We illustrate this possibility using an ideal example. Let $f_P(x, v)$ be the observed bivariate score distributions of population P on the test form X and the anchor V, and let $g_Q(y, v)$ be the observed bivariate score distributions of population Q on the test form Y and the same anchor V. Assume that $f_P(x, v)$ and $g_Q(y, v)$ follow bivariate normal distributions $N((\mu_x, \mu_{v,P}), (\sigma_x^2, \sigma_{v,P}^2, \rho_P))$ and $N((\mu_y, \mu_{v,Q}), (\sigma_y^2, \sigma_{v,Q}^2, \rho_Q))$, respectively. For simplicity, let the variances be 1 and $\rho_P = \rho_Q = \rho > 0$ in the following discussion. Then the conditional density of X, given V, is a normal distribution with conditional mean and conditional variance (Kendall & Stuart, 1977, p. 411), as follows:

$$\mu(X|V = v) = \mu_{v1} + \rho\sigma_{v1}(v - \mu_x)/\sigma_x$$

$$\sigma^2(X|V = v) = \sigma_{v1}^2(1 - \rho^2).$$

The conditional density of Y, given V, is also a normal distribution, with

$$\mu(Y|V = v) = \mu_{v2} + \rho\sigma_{v2}(v - \mu_y)/\sigma_y$$

$$\sigma^2(Y|V = v) = \sigma_{v2}^2(1 - \rho^2).$$

Then $f_P(x|v)$ and $g_Q(y|v)$ are the same if and only if $\mu_{v1} - \mu_{v2} = \rho(\mu_x - \mu_y)$. In other words, when the two groups are different, that is, when $\mu_{v1} \neq \mu_{v2}$, the observed conditional distributions can be the same if the difference in the means of the observed form scores multiplied by the correlation between the anchor and the form scores (note $\rho_P = \rho_Q$) equals the difference in the observed anchor score means.

3 Comparison of Distributions

To use the arguments in section 2 and determine whether the FE equipercentile equating method is appropriate for use in equating practice, we need to test whether the item difficulty distributions of form X and form Y are the same, and whether the observed conditional distributions $f_P(x|v)$ and $g_Q(y|v)$ are the same.

3.1 Comparison of Two Test Form Distributions

When the two groups are equivalent, testing form equivalence can be demonstrated by testing the equivalence of the two score distributions on form X and form Y, as discussed by Hanson (1996). Alternatively, Dorans and Lawrence (1990) used the standard error of linear equating for equivalent groups to determine if the two nearly identical test forms are the same. However, when the group abilities are very different, the preceding approaches are not applicable. Instead, reliable item statistics for items on the test forms are required to compare the forms. We assume that the item difficulty statistics, such as item difficulty parameters from an item response theory calibration or P-plus or Delta from conventional analysis (see Swineford, 1980), are available and reliable.

Many statistical methods can be used to test the homogeneity of two discrete distributions in a contingency table (see Agresti, 2002). One of them is the simple goodness-of-fit test for homogeneity, which may help to identify whether two test forms are the same from the perspective of test statistical specifications. However, one issue is that usually, at each item difficulty level, there are only a few items. If the sample size compared to the number of categories is small, the accuracy of this test is compromised.

Another option is to carry out the Kolmogorov–Smirnov (KS) two-sample test (see Conover, 1971) or other related tests using the item difficulty statistics directly, instead of contingency tables. The KS test is one of the most useful and general nonparametric methods for comparing two samples as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions $F_n(x)$ and $F'_n(x)$ of the two samples, where n is the number of items in a test form in our setting. The KS statistic is

defined as

$$D_n = \sup_x |F_n(x) - F'_n(x)|,$$

which measures the supreme difference between the two empirical cumulative distributions. When the sample size n increases, D_n converges to the Kolmogorov distribution (see Conover, 1971, p. 309). Larger values of the KS statistic indicate larger differences between the item difficulty distributions of the two test forms. Because the KS statistic is based on asymptotic results, it is not wise to use it when the number of items on the test is small. Besides statistical tests, graphic comparisons of the empirical item difficulty distributions are also recommended.

3.2 Comparison of Two Conditional Distributions

Because a complete examination of conditional distributions is too tedious when the score range is large, we focus on comparison of summary statistics: the conditional mean, the conditional standard deviation, and the conditional skewness. In addition, we modify those statistics suggested by Holland and Thayer (2000, section 4) for our purposes. Let Z_1, Z_2, and Z_3 be the discrepancies of conditional means, conditional variances, and conditional skewnesses of the two conditional distributions for a given anchor score, respectively, defined as

$$\begin{aligned} Z_1 &= \frac{M_{1X} - M_{1Y}}{\sqrt{(M_{2X} + M_{2Y})/N}}, \\ Z_2 &= \frac{\log M_{2X} - \log M_{2Y}}{\sqrt{\left(\frac{M_{4X}}{M_{2X}^2} + \frac{M_{4Y}}{M_{2Y}^2} - 2\right)/N}}, \\ Z_3 &= \frac{\frac{M_{3X}}{M_{2X}^{3/2}} - \frac{M_{3Y}}{M_{2Y}^{3/2}}}{\sqrt{3/N}}, \end{aligned} \tag{2}$$

where M_{iX} and M_{iY} are the ith conditional central sample moments for $i = 1, 2, 3, 4$; N denotes the number of form scores for the given anchor score. When the two sets of conditional distributions are the same, that is, when $f_P(x|v)$ and $g_Q(y|v)$ are the same, Z_1 and Z_2 follow an approximate standard normal distribution. As mentioned by Holland and Thayer (2000), log variances are used because they often exhibit more approximate

normality than variances themselves. Therefore we will expect Z_1 and Z_2 to fall in the $[-2, 2]$ confidence band with a 95% confidence level across the anchor score range. The denominator of Z_3 is the correct asymptotic variance value for data from the normal distribution, and Z_3 gives a rough index of the difference between the two conditional distributions. Derivation of (2) is detailed in the appendix.

4 An Application

This section illustrates how the proposed method can be applied in a real test equating situation.

We consider equating for a large-scale admissions test. A new test form X is equated back to four old forms, $Y1$, $Y2$, $Y3$, and $Y4$, through the NEAT design. The standardized mean difference is used with a t test to see whether the mean difference on anchor scores between the old and new groups is statistically significant. In this admission test, we assume that the samples in equating are representative. The equating sample size is around 5,000. The number of items on each test form is around 70. We noticed, from Table 1, that the new group is very close in ability to the old group, who took $Y4$ based on the standardized mean difference, so this link is not considered until the end of this section because we primarily want to look at the situations in which the abilities are different. We focus on FE equating with relatively large group differences, that is, equating X to $Y1$, $Y2$, and $Y3$, respectively.

To evaluate test form difficulties, we use the equated deltas (Swineford, 1980) for the test items. Delta measures the item difficulty and is placed on the same scale for different administrations via delta equating. In Table 2, we summarize the means and standard deviations of item difficulty or equated deltas for the four forms. Comparison of the

Table 1
Ability Comparison of New and Old Groups

On anchor	X to $Y1$	X to $Y2$	X to $Y3$	X to $Y4$
Std. mean differences	-0.21	0.15	0.13	-0.04
Ratio of variances	1.07	0.98	1.03	1.02

Table 2
Summary of New and Old Forms

Form difficulty	X	$Y1$	$Y2$	$Y3$	$Y4$
M	11.4	11.4	11.5	11.3	11.4
SD	2.3	2.4	2.4	2.4	2.4
KS statistic		.0896	.0746	.1045	.1194
KS p value		.951	.992	.858	.726

Note. KS = Kolmogorov–Smirnov.

summary statistics indicates that forms $Y1$, $Y2$, and $Y3$ are all close to X. The KS test is given in the last two rows of Table 2, where the KS statistic is the calculated test statistic and the KS p value is the corresponding p value. Smaller KS values indicate that test forms are closer in difficulty. We observe, as expected from the KS test, that the old and new forms are comparable in difficulty. This is because all the items are pretested and the forms are well constructed for this admission test. Among the three old forms, $Y2$ is the closest one in difficulty to the new form X. Figure 1 shows the empirical distributions of item difficulties on the new form and the old forms. The solid line is the cumulative distribution of the new form's difficulty, and the dashed line is that of the old forms' difficulty. When the solid line is lower than the dashed line at a certain delta value, it indicates that up to that delta value, the new form has fewer easier items. In Figure 1, we observe that the $Y2$ and $Y1$ lines closely intertwine with the X line, and the differences are relatively small, which implies that $Y2$ and $Y1$ are close to X in form difficulty; $Y3$ is easier overall than X in form difficulty, except for a few hard items at the top; and overall, $Y4$ is also close to X in difficulty but has a relatively large difference around the delta value of 13. Hence, from the preceding form difficulty analysis, $Y2$ is the closest to X in form difficulty, and $Y1$ is the second closest. X and $Y3$ are not that close in form difficulty. One would expect that X-to-$Y2$ equating would be close to the identity function, according to Theorem 2, regardless of which equating methods are used.

To see whether FE equating can be used for the equating, we further analyze the conditional score distributions. The conditional means, conditional standard deviations, and conditional skewnesses, given anchor scores, are plotted in Figures 2–4, respectively. As can be seen, among $Y1$, $Y2$, and $Y3$, the conditional mean of $Y1$ is more discrepant

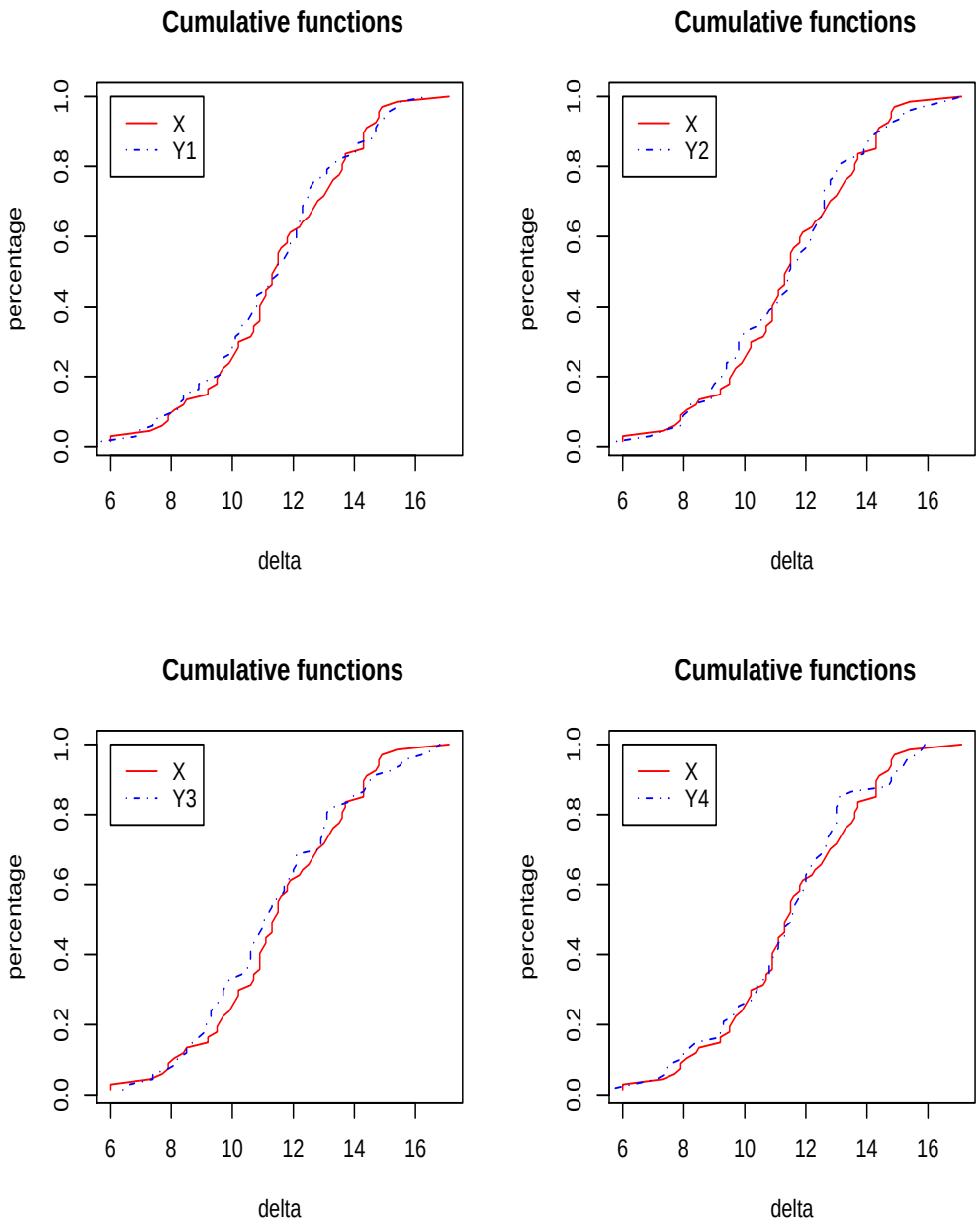

Figure 1. **Cumulative distributions of form difficulty.**

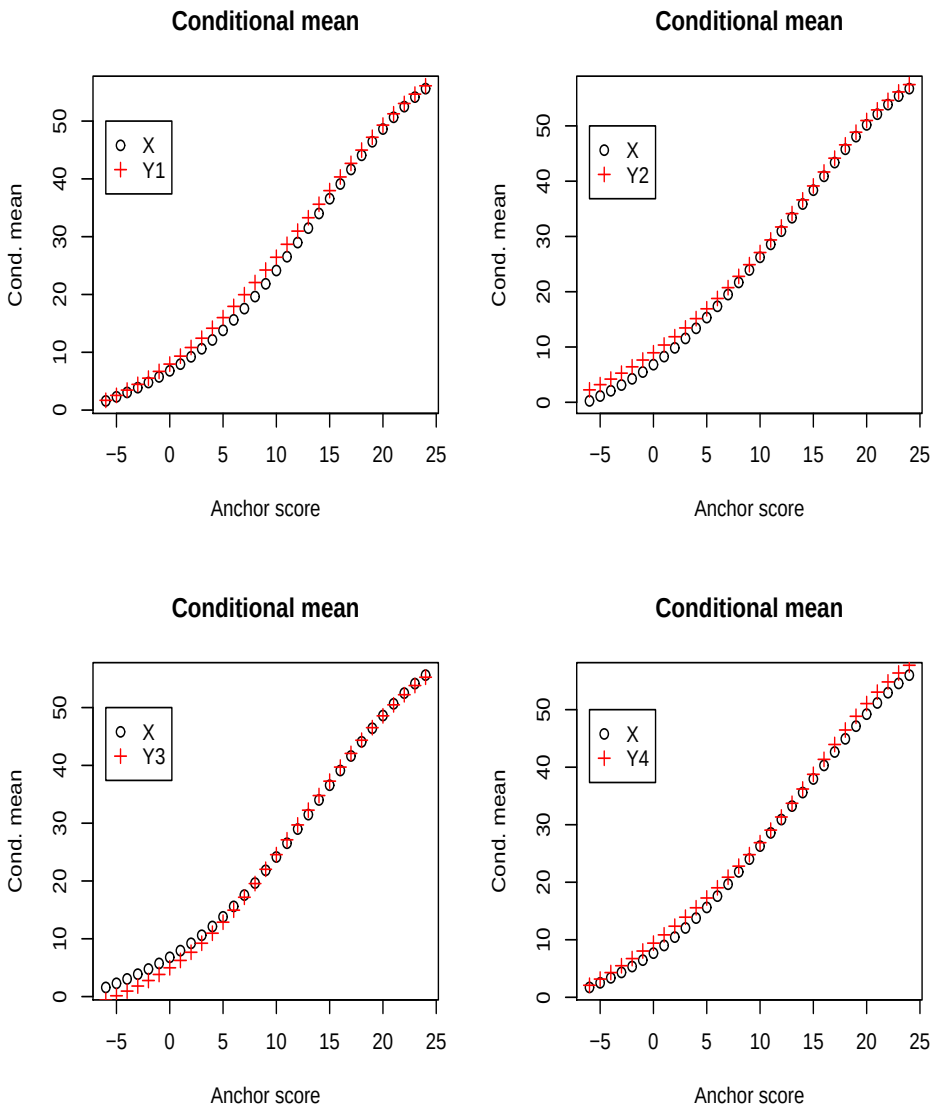

Figure 2. **Comparison of conditional means.**

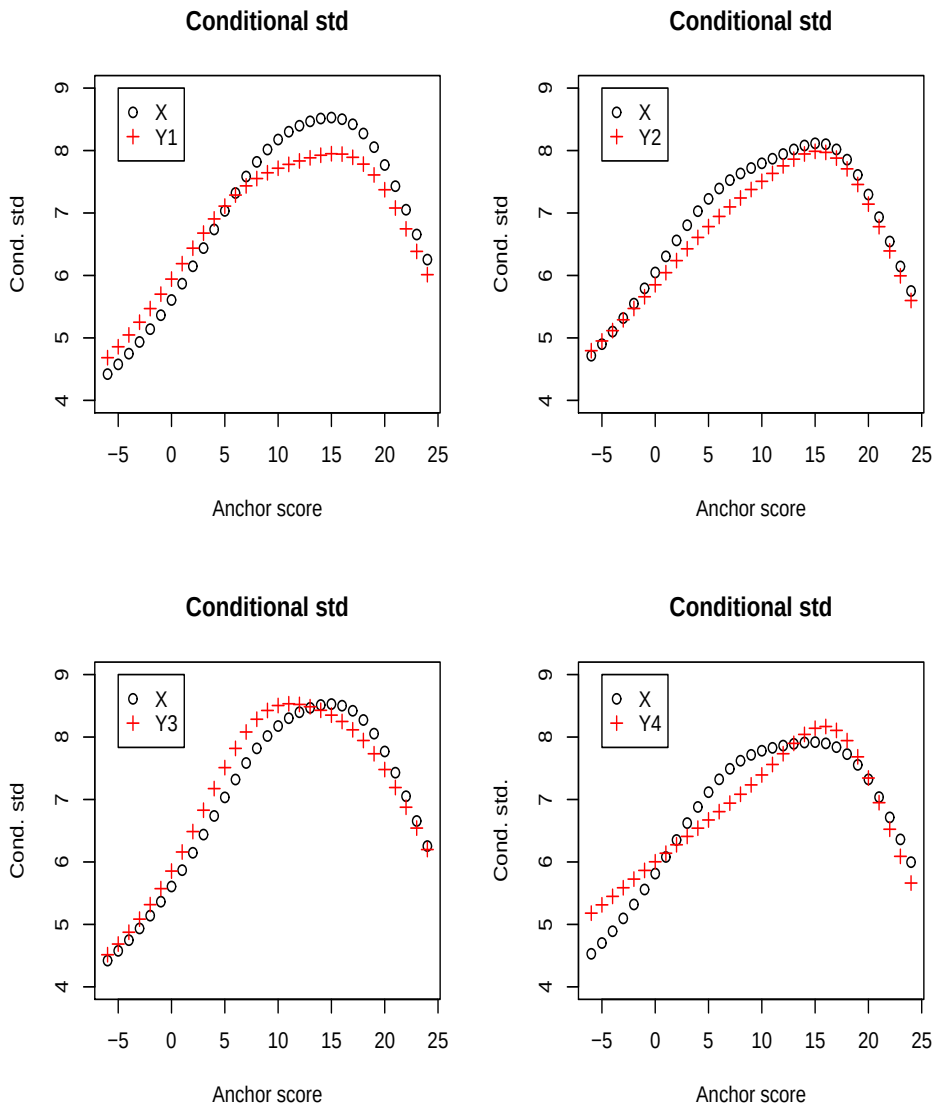

Figure 3. **Comparison of conditional standard deviations.**

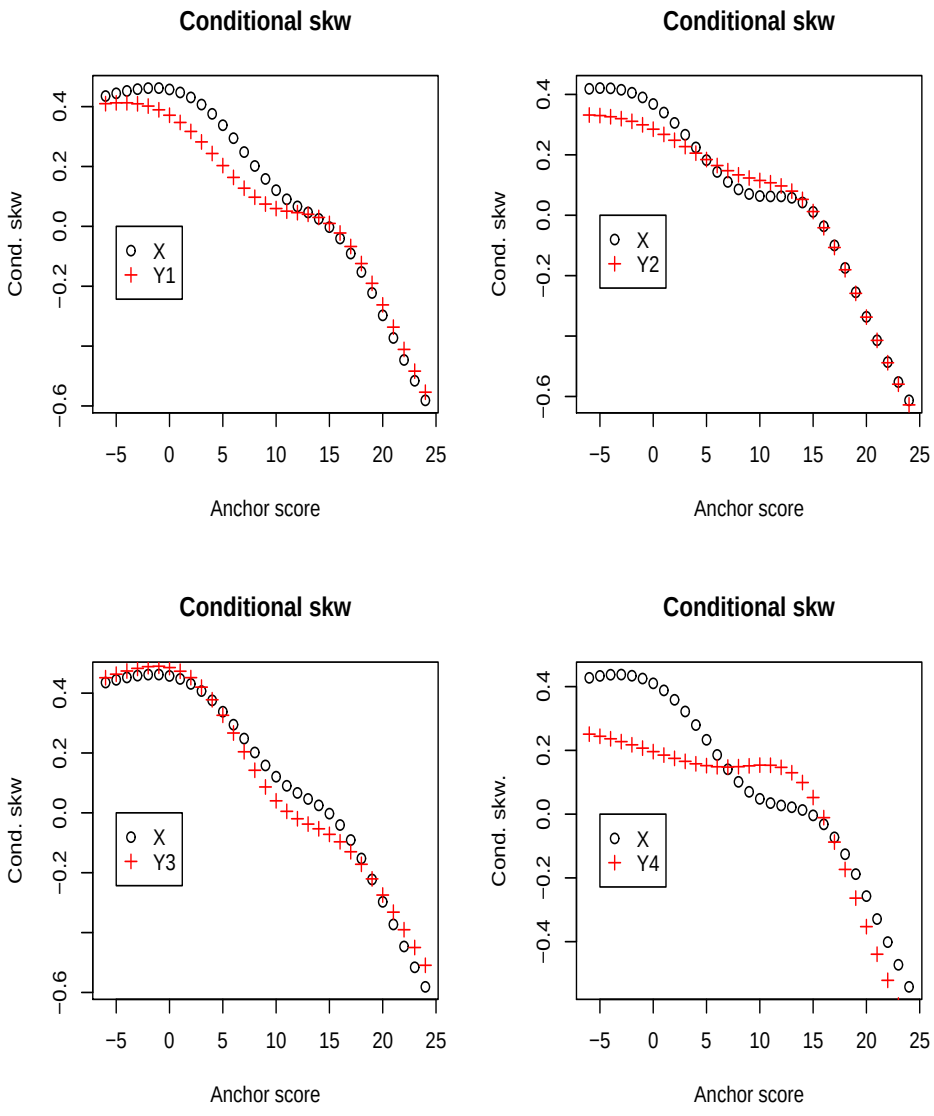

Figure 4. **Comparison of conditional skewness.**

from the new form, especially in the middle range of anchor scores. $Y2$ and $Y3$ have some differences from the new form at negative anchor scores, but the number of examinees having negative anchor scores is few. When looking at conditional standard deviations and skewnesses, $Y2$ and $Y3$ are also relatively close to the new form X. In addition, we calculate the statistics introduced in (2). Figures 5–7 are the realizations of Z_1, Z_2, and Z_3 for X-to-$Y4$ (circles), X-to-$Y1$ (triangles), X-to-$Y2$ (plusses), and X-to-$Y3$ (crosses) equating.

From Figure 5, Z_1 for X-to-$Y2$ and Z_1 for X-to-$Y3$ are close to zero above anchor score 5; Z_1 for X-to-$Y1$ is close to zero below anchor score 0 and above 15, and it is near -2 in the rest of the range. From Figures 6 and 7, Z_2 and Z_3 for X-to-$Y2$ are closer to the zero line than are the other two (X-to-$Y1$ and X-to-$Y3$), and all lines are within the $[-2, 2]$ band.

As mentioned before, because of the well-constructed and thus similar test forms, it is hard to rule out FE equating in all four links, and the equating functions should not differ too much from the identity function. However, given that the forms and conditional distributions are so close for X-to-$Y2$, and the assumptions of FE equipercentile equating

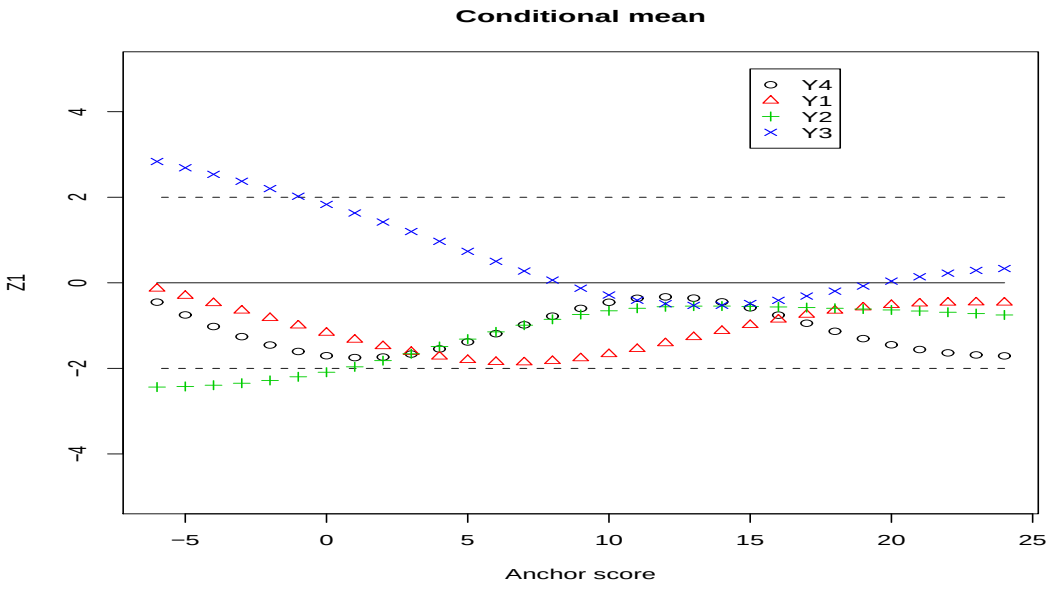

Figure 5. **Discrepancy of conditional means.**

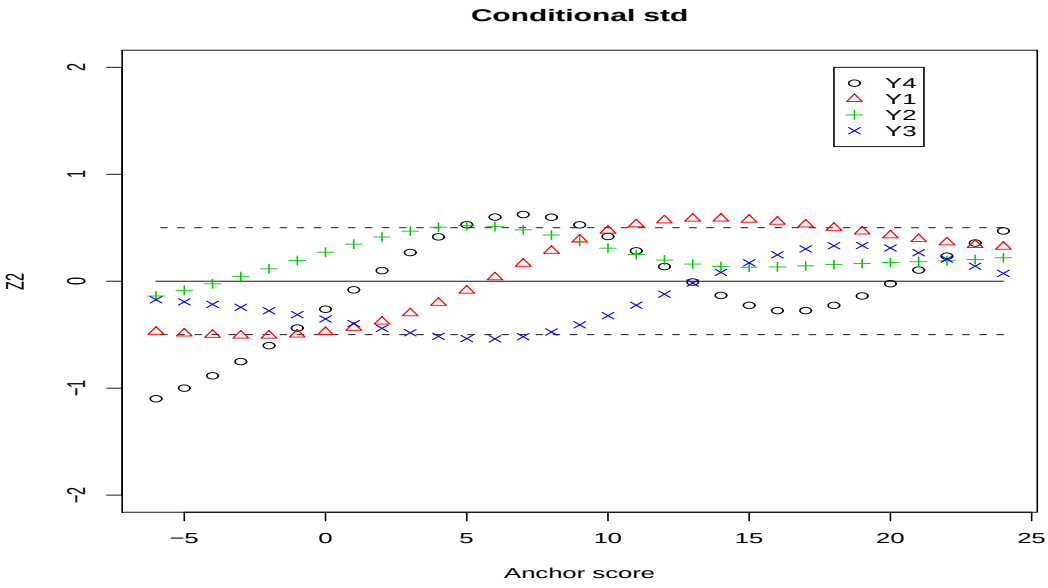

Figure 6. **Discrepancy of conditional standard deviations.**

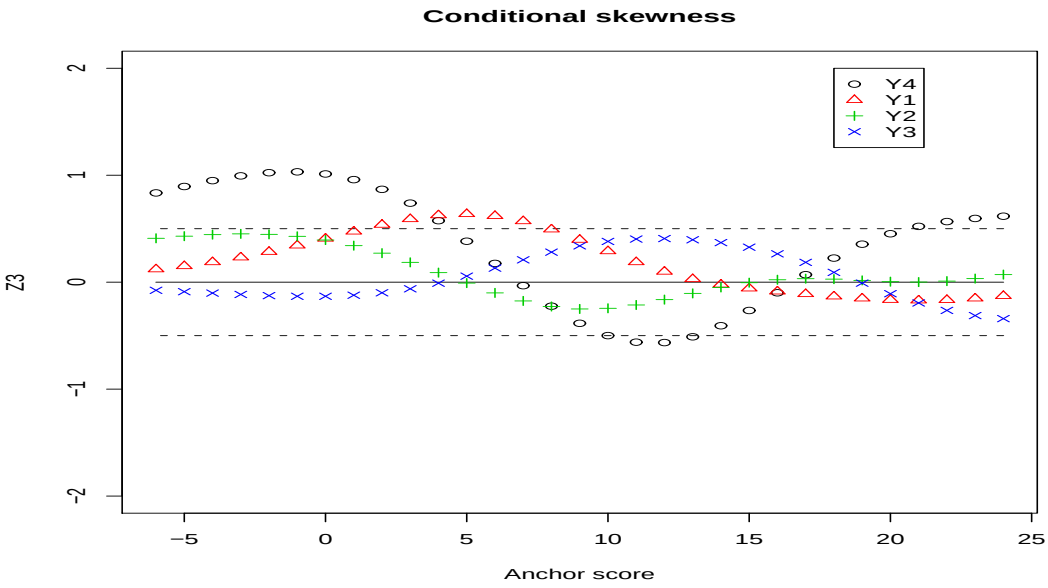

Figure 7. **Discrepancy of conditional skewness.**

are most likely to be true, FE equipercentile equating should be considered when equating X to $Y2$ under the NEAT design. In fact, the raw-to-raw conversion line from FE equating is the closest to the identity line in absolute value when compared to the other three methods (see Figure 8, in which the plotted lines show differences between the identity functions and the raw-to-raw conversions from four equating methods: Tucker, Levine, CE, and FE; Kolen & Brennan, 2004). For the actual equating that was done, FE equipercentile equating was not considered because of the convention that FE equipercentile equating "should be conducted only when the two populations are reasonably similar to one another" (Kolen & Brennan, 2004, p. 139). The operational or actual conversion for this link was chosen to be the Levine method based on the fact that the ability difference is large and a linear relationship between scores on the old and new forms seems to be supported by the data.

Forms X and $Y3$ are not as close in form difficulty as are forms X and $Y2$, as seen in Figure 1 and Table 2, but the observed conditional distributions are close, as seen in Figures 2–7. For this link, FE equating may not be appropriate. For the actual equating,

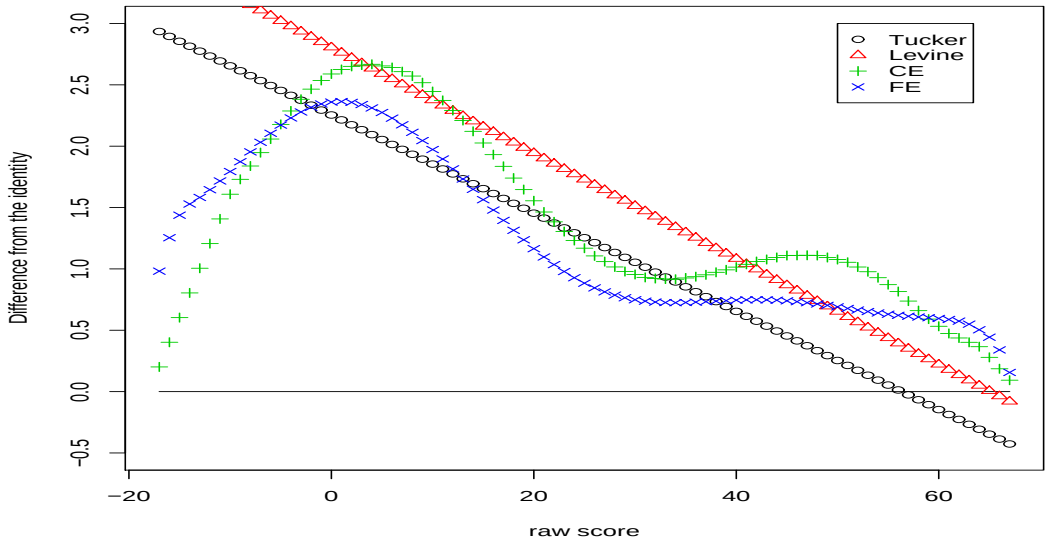

Figure 8. **Difference of conversions for X-to-$Y2$ equating.**

FE was not considered for the same reasons mentioned previously.

The similarities in form difficulties and the observed conditional distributions between X and $Y1$ are less well defined when compared to the preceding two links. It is not clear whether FE equating is appropriate for equating X to $Y1$. For the actual equating, FE was not considered.

Now we consider X-to-$Y4$ equating. As noticed before, the new and old groups have similar abilities, and therefore the FE equating assumptions are likely to be true whether or not forms and conditional distributions are close. For the actual equating, FE was considered. Table 3 summarizes the statistical test results and suggestions as to whether the FE method could be used in this application.

5 Conclusion

When two populations P and Q are of the same ability, the FE equating assumptions are true. Theoretically, equating under the EG design should be carried out to adjust for the form difference. However, when the two groups have similar ability, FE assumptions are likely to be true. Then FE equipercentile equating is used in practice, instead of EG equating, in the hope that FE equipercentile equating can adjust for systematic error, which is introduced if the procedure used in equating cannot adequately deal with group or form differences. On the other hand, when the two test forms are the same, no equating is necessary; that is, the equating function is the identity function. In the case in

Table 3
Summary of Forms and Conditional Distributions

	X to $Y1$	X to $Y2$	X to $Y3$	X to $Y4$
Are the abilities the same or different?	Different	Different	Different	Same
Are the forms equivalent or different?	Not clear	Equivalent	Somewhat different	
Are the conditional distributions the same or different?	Not clear	Same	Same	
Is frequency estimation equating recommended?	Not clear	Yes	No	Yes

which differences in the two form score distributions are indistinguishable from differences expected from sampling error, previous studies suggest that the difference in an observed score equating function from an identity function would also be due to sampling error. However, systematic error is not considered in these studies, and the approach used is not applicable when two populations have different abilities. So we have to rely on reliable item difficulty statistics. As discussed in this report, when the two forms are similar, that is, the two forms have the same item difficulty distributions, and the observed conditional distributions are the same, the FE equipercentile equating assumptions are likely to be true. Then FE equipercentile equating can be used in practice, instead of no equating, in the hope that FE equipercentile equating can adjust for systematic error.

This study is aimed at providing guidelines for appropriately using the FE equipercentile equating method when the group difference in ability is large. As mentioned earlier, when the two forms are not significantly different statistically, and when the observed conditional distributions $f_P(x|v)$ and $g_Q(y|v)$ are not significantly different as well, the assumptions of the FE equating method are likely to be true. Therefore, in this situation, the equating function derived from the FE method should not be excluded from consideration. When the two forms are not significantly different statistically but $f_P(x|v)$ and $g_Q(y|v)$ are different, or when the two forms are statistically different but the observed conditional distributions are statistically the same, then the assumptions of the FE equating method are likely to be incorrect. In these two cases, the FE method is not recommended for equating.

Judgment of whether the group difference in ability is sufficiently large is determined by the individual testing program. One helpful statistical tool is to use a t test on the anchor score means to see if there is a statistically significant difference between the old and new group abilities. The standardized mean difference of the old and new groups used in this report is an exemplary index. Also, in comparisons of test form item difficulty distributions, it is crucial to use reliable, or at least comparable, item difficulty estimates. Statistical tests are important tools, but different criteria for confirming homogeneity of distributions are recommended and are program-specific.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: John Wiley.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

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: Academic Press.

Conover, W. (1971). *Practical nonparametric statistics*. New York: John Wiley.

Dorans, N., & Lawrence, I. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education*, *3*, 245–254.

Hanson, B. (1996). Testing for differences in test score distributions using loglinear models. *Applied Measurement in Education*, *9*, 305–321.

Harris, D., & Crouse, J. (1993). A study of criteria used in equating. *Applied Measurement in Education*, *6*, 195–240.

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

Holland, P., von Davier, A., Sinharay, S., & Han, N. (2008). An approach to evaluating the missing data assumptions of the chain and post-stratification equating methods for the NEAT design. *Journal of Educational Measurement*, *45*, 17–43.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed.). New York: Macmillan.

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

Swineford, F. (1980). *Item analysis at ETS* (ETS Statistical Rep. No. SR-80-05). Princeton, NJ: ETS.

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

Note

[1] Defined as $(\text{mean}(P) - \text{mean}(Q))/\text{Std}(P + Q)$ on the anchor.

Appendix

The following discusses the derivation of (2). Let n be the number of examinees taken from a population. The sample variance M_2 is then given by $M_2 = \sum_i (X_i - \bar{X})^2/n$. The expected value of M_2 for a sample of size n is then given by

$$\mathrm{E}(M_2) = \frac{n-1}{n}\mu_2.$$

Similarly, the expected variance of the sample variance is given by

$$\mathrm{Var(M_2)} = \frac{\mathrm{(n-1)^2}}{\mathrm{n^3}}\mu_4 - \frac{\mathrm{(n-1)(n-3)}}{\mathrm{n^3}}\mu_2^2$$

(Kendall & Stuart, 1977, p. 260). By the Delta method (see, e.g., Kendall & Stuart, 1977, chap. 10), for large n,

$$\mathrm{Var(\log M_2)} \sim \{\frac{\mu_4}{\mu_2^2} - 1\}/\mathrm{n}.$$

The sample skewness is given by $\gamma = M_3/M_2^{3/2}$. For a normal distribution population with a sample size of n, the variance is given by $\mathrm{Var}(\gamma) \approx 6/\mathrm{n}$ (Kendall & Stuart, 1977, p. 316).