ACT Research Report Series 94-4

A Comparison of Presmoothing and Postsmoothing Methods in Equipercentile Equating

Bradley A. Hanson

Lingjia Zeng

Dean Colton

ACT

November 1994

A Comparison of Presmoothing and Postsmoothing Methods in Equipercentile Equating

Bradley A. Hanson
Lingjia Zeng
Dean Colton

American College Testing

Abstract

This paper compares various methods of smoothed equipercentile equating and linear equating in the random groups equating design. Three presmoothing methods (based on the beta binomial model, four-parameter beta binomial model and a log-linear model) are compared to postsmoothing using cubic splines, linear equating and unsmoothed equipercentile equating. Performance of these methods is evaluated by specifying several pairs of population distributions and estimating equating error by Monte Carlo methods. The results indicate that both presmoothing and postsmoothing methods can improve estimation of the equipercentile equating function and that presmoothing and postsmoothing methods provide comparable levels of performance in terms of equating error.

A Comparison of Presmoothing and Postsmoothing Methods in Equipercentile Equating

Two types of smoothing methods have been suggested for improving equipercentile equating results in the random groups equating design. Presmoothing methods involve smoothing the score distributions for the two test forms to be equated before equipercentile equating is performed. Postsmoothing methods involve smoothing the equipercentile equating function produced from the unsmoothed score distributions. Both presmoothing and postsmoothing methods have been found to reduce equating error as compared with unsmoothed equipercentile equating. Kolen (1984) concluded that a postsmoothing method based on cubic splines is preferred to unsmoothed equipercentile equating. Fairbank (1987) studied seven presmoothing methods and seven postsmoothing methods and concluded that of the methods he investigated a presmoothing method based on the negative hypergeometric distribution and a postsmoothing method based on cubic splines were the preferred presmoothing and postsmoothing methods.

Cope & Kolen (1990) and Hanson (1990) investigated various methods of smoothing distributions of test scores (these methods could be used for presmoothing in equipercentile equating). Cope & Kolen (1990) found that smoothing based on the four-parameter beta binomial model (Lord, 1965) provided more accurate results than smoothing based on the beta binomial (negative hypergeometric) model. Hanson (1990) found the four-parameter beta binomial model and a log-linear model (Holland & Thayer, 1987) provided the most accurate results of the methods examined. The results presented in the papers of Cope & Kolen (1990) and Hanson (1990) suggest that the four-parameter beta binomial model and the log-linear model, when used in presmoothing of score distributions, may produce less equating error than presmoothing based on the beta binomial model which produced the most accurate results in Fairbank (1987).

This paper investigates the relative performance of unsmoothed equipercentile equating, several presmoothing methods (using the two and four parameter beta binomial models, and a log-linear model), a postsmoothing method based on cubic splines, and linear equating (Angoff, 1971). The relative performance of these methods is investigated using several example data sets.

Equipercentile Equating

The focus of this paper is on the random groups equating design. In the random groups equating design the new and old forms are each administered to a random sample from a common population. Let the random variables X and Y represent the test scores on the new and old forms

of the test, respectively, for a random examinee from the population of interest. The test score X is to be equated to the test score Y.

The equipercentile equating function is determined by the cumulative distribution functions of X and Y. If the random variables X and Y were continuous, then the equipercentile equating function would be given by $F_Y^{-1}[F_X(x)]$ where $F_Y(y) \equiv \Pr(Y < y)$ and $F_X(x) \equiv \Pr(X < x)$. Because X and Y are discrete random variables the equipercentile equating function is not defined. To define an equipercentile equating function based on X and Y the common practice is to use the equipercentile equating function based on continuous approximations of X and Y. The most widely used continuous approximation is based on a uniform kernel being applied to X and Y to produce approximating continuous distributions (Holland & Thayer, 1989). The uniform kernel spreads the density at each score point uniformly in a unit interval one-half point above and below the score point. This results in a continuous distribution on the interval $(-1/2,\ K + 1/2)$, where K is the number of items on the test. Based on the continuous distribution given by the uniform kernel the equipercentile equivalent of raw score i on the new form is given by:

$$\frac{p^*(i) - \Pr\big(Y < u^*(i)\big)}{\Pr\big(Y = u^*(i)\big)} + u^*(i) - .5\,, \tag{1}$$

where

$$p^*(i) = \Pr(X < i) + .5\ \Pr(X = i)\,,$$

and $u^*(i)$ is the smallest integer such that $p^*(i) < \Pr\big(Y \leq u^*(i)\big)$.

For presmoothing methods of equipercentile equating the distributions of X and Y are smoothed before the equipercentile function given in Equation 1 is used. Presmoothing methods of equipercentile equating are based on the premise that smoothing the distributions of X and Y has the potential to improve estimation of these distributions and, presumably, the equipercentile equating functions based on these distributions. Three general methods of smoothing the observed raw score distributions are used in this paper: a log-linear model, the beta binomial model, and the four-parameter beta binomial model.

For postsmoothing methods a smoothing procedure is applied to the equipercentile equating function produced with Equation 1 using the observed (unsmoothed) raw score distributions. The postsmoothing method used in this paper uses cubic splines to smooth the equipercentile equating function (Kolen, 1984).

Log-Linear Model Smoothing

Rosenbaum & Thayer (1987) suggested using log-linear models to estimate the bivariate distributions needed for equipercentile equating in the common item equating design. Similar log-linear models can also be used to smooth the univarate distributions of X and Y in the random groups equating design. These log-linear models are discussed by Holland & Thayer (1987) and Haberman (1974). For the distribution of X (the same model would be used for the distribution of Y) the model used in this paper can be written as

$$\log\left[N \Pr(X = i)\right] = \beta_0 + \sum_{k=1}^{m} \beta_k i^k , \tag{2}$$

where N is the sample size, and $m \leq K$. Estimates of the raw score probabilities based on the maximum likelihood estimates of the parameters of the model given in Equation 2 have the property that the first m moments of the fitted distribution are identical to the first m moments calculated from the observed frequencies. For example, if $m = 4$ then the mean, variance, skewness and kurtosis of the fitted distribution will equal the mean, variance, skewness and kurtosis computed from the observed frequencies.

In this paper maximum likelihood is used to estimate the parameters in Equation 2. The procedure used is that given by Haberman (1974), which is also discussed in Holland & Thayer (1987).

Beta Binomial and Four-parameter Beta Binomial Model Smoothing

The beta binomial and four-parameter beta binomial models are strong true score models described in Lord & Novick (1968) and Lord (1965). Under these strong true score models the probability that a raw score random variable Z (which may be the raw score on either the new or old form) equals i ($i = 0, \ldots, K$, where there are K items on the test), is given by:

$$\Pr(Z = i) = \int_l^u \Pr(Z = i \mid \tau)\, g(\tau)\, d\tau, \tag{3}$$

where τ is the proportion correct true score. The conditional error distribution $[\Pr(Z = i \mid \tau)]$ is assumed to be binomial (with parameters K and τ). For the beta binomial model the true score density $g(\tau)$ is assumed to be a beta distribution. For the four-parameter beta binomial model the true score density is assumed to belong to the four-parameter beta class of densities. The four-parameter beta distribution is a generalization of the beta distribution that in addition to the two

shape parameters (α and β) has parameters for the lower (l) and upper (u) limits of the distribution ($l \geq 0$ and $u \leq 1$).

In this paper estimates of parameters in the beta binomial and four-parameter beta binomial models are obtained using the method of moments. For the beta binomial model the observed score mean and variance are used to calculate estimates of the two parameters of the true score distribution (Lord & Novick, 1968, page 517). For the four-parameter beta binomial model the first four moments of the observed score distribution are used to calculate estimates of the four parameters of the true score distribution as described by Hanson (1991). After estimates of the parameters of the true score distribution are obtained the observed score distribution is calculated using the procedure described by Hanson (1991).

Postsmoothing Using Cubic Splines

A cubic spline method for smoothing an equipercentile equating function has been described by Kolen (1984) and Kolen & Jarjoura (1987). A smoothing cubic spline function is fit to the equipercentile equating function (computed from the observed raw score distributions) relating scores on the new form to scores on the old form. The degree of smoothing is controlled by a smoothing parameter s. Selecting $s = 0$ leaves the equipercentile equating function computed from the observed distributions unchanged. Selecting s to be very large (e.g., $s = 999$) results in a linear function.

A second smoothing cubic spline function is fit to the equipercentile equating function relating scores on the old form to scores on the new form (using the same smoothing parameter as that used for first smoothing cubic spline). The smoothed equipercentile equating function is taken as the average of the smoothing cubic spline relating scores on the new form to scores on the old form and the inverse of the smoothing cubic spline relating scores on the old form to scores on the new form. The smoothing cubic spline is not computed for some scores at the extremes (both low and high scores) — equivalents for these scores are found by linear interpolation.

Method

To investigate the performance of smoothed equipercentile equating methods, population distributions of test scores for new and old forms are specified and equating error is estimated by Monte Carlo methods. The population distributions are defined using data from test administrations. Results will be given for five pairs of population distributions.

The first pair of population distributions are defined using data from two 30-item subsets from a professional licensure exam. The 30-item subsets were used as a set of common items for equating to two different links (this exam is equated using a common item equating design). The two sets of common items were chosen to result in scores that are approximately parallel to one another. The two sets of common items were in separately timed test sections. One of these two subsets of items is designated as the new form and the other subset of items is designated as the old form for the purposes of this study. Table 1 gives the sample statistics for each form for a sample of 39,765 examinees (each examinee took both sets of items). The distributions of scores for the new and old forms are presented in Figure 1. For these two score distributions the population distributions are taken to be the observed distributions as presented in Figure 1 mixed with a uniform distribution. The purpose of this minimal smoothing of the distributions (via the mixing with a uniform distribution) was to smooth large fluctuations in the equipercentile equating function (which is to be considered the population equating function) for low scores. If p_i is the probability of raw score i based on the observed data then the smoothed probability (p_i^*) to be used as the population distribution is given by $p_i^* = .999\, p_i + \frac{.001}{31}$.

The second pair of population distributions is defined using data from two forms of a 20-item Reading test which is the basis of one of the reading subscores reported on the ACT Assessment. The two forms contain the same items but the items are in a different order in each form. One of these forms is designated as the new form and the other form is designated as the old form for the purposes of this study. Table 1 gives the sample statistics for each form for randomly equivalent samples of 82,073 examinees for the new form and 83,709 examinees for the old form. The distributions of scores for the new and old forms are presented in Figure 2. For these two score distributions the population distributions are taken to be the observed distributions as presented in Figure 2.

The third and fourth pairs of population distributions are based on samples used in equating forms of the ACT Assessment. The third and fourth pairs of population distributions are defined with data used in equating two forms of the ACT English (75 items) and Science Reasoning (40 items) tests, respectively (the two forms for English and the two forms for Science Reasoning are different forms). For both pairs of forms one of the forms is designated as the new form and the other form is designated as the old form for the purposes of this study. Table 1 gives sample statistics for the two forms for randomly equivalent samples of 3,158 (new form) and 3,293 (old form) for

the English test and randomly equivalent samples of 2,831 (new form) and 2,898 (old form) for the Science Reasoning test (the samples for the English test and Science Reasoning test are from a different test dates).

Because of the rough shape of the equipercentile equating function based on the observed distributions (see the top panels of Figures 8 and 9), the pairs of population distributions based on the English and Science Reasoning test data will be defined with model based fitted distributions. For the English test the log-linear model given in Equation 2 with $m = 9$ is used as the fitted distribution for the new form and the four-parameter beta binomial model is used as the fitted distribution for the old form. For the Science Reasoning test the four-parameter beta binomial model is used as the fitted distribution for both the new and old forms. Figures 3 and 4 present the observed and fitted distributions for the English and Science Reasoning tests, respectively.

The likelihood ratio chi-squared goodness of fit statistics for the new and old forms of the English test are 52.96 (with 65 degrees of freedom) and 90.85 (with 70 degrees of freedom), respectively. The likelihood ratio chi-squared goodness of fit statistics for the new and old forms of the Science Reasoning test are 32.96 (with 35 degrees of freedom) and 39.91 (with 35 degrees of freedom), respectively. Only the chi-squared statistic for the old form of the English test is large when compared with a chi-squared distribution with the appropriate degrees of freedom. Based on the relatively large sample size used in computing the goodness of fit statistic and the visual display of the fitted distribution given in Figure 3 it is concluded that the fitted distribution for the old form of the English test is, for practical purposes, adequate.

The fifth pair of population distributions are based on samples used for equating two forms of the PLAN Math test (40 items). One of the forms is designated as the new form and the other form is designated as the old form for the purposes of this study. Table 1 gives sample statistics for the two forms for randomly equivalent samples of 2,190 for the new form and 2,176 for the old form. Again, because of the relatively small sample sizes, the population distributions based on these two sample distributions will be defined with model based fitted distributions. For the new form the four-parameter beta binomial model is used and for the old form the log-linear model given in Equation 2 with $m = 6$ is used. The likelihood ratio chi-squared goodness of fit statistics for the new and old forms are 41.39 (with 35 degrees of freedom) and 43.2 (with 33 degrees of freedom), respectively. Figure 5 presents the observed and fitted distributions for the new and old forms. The fitted distributions appear to provide a reasonable fit to the data.

For each of the pairs of population test score distributions, 500 samples for each of five sample sizes (100, 250, 500, 1000 and 3000) are drawn. For each of the 12,500 pairs of sample distributions (5 pairs of population distributions by 5 sample sizes by 500 samples) ten estimated equating functions are computed.

One of the estimated equating functions computed is the equipercentile equating function based on the observed data (the unsmoothed equipercentile equating function). Applying Equation 1 to compute the equipercentile equating function can be problematic when there are zero frequencies in one or both of the raw score distributions. Consequently, before computing the equipercentile equating function using Equation 1, the observed distributions are mixed with a uniform distribution to eliminate score combinations with zero probability. If p_i is the probability of raw score i based on the observed data then the modified probability (p_i^*) for that raw score is given by $p_i^* = .999999\, p_i + .000001\, t^{-1}$, where t is the number of score categories (for the number correct test score this is the number of items plus one).

Five equipercentile equating functions based on presmoothing methods are computed. Three of the estimated equating functions are based on log-linear model smoothing (using three alternate models). The models are distinguished by the highest degree polynomial used in the model (m of Equation 2). The three models to be used will be those corresponding to $m = 3$, $m = 4$ and $m = 6$. The value $m = 3$ was chosen as the minimum value to be used since it has been our experience in practice that $m = 3$ is the smallest value that provides an adequate fit for test score distributions. The value $m = 6$ was chosen as the maximum value to be used since it has been our experience that in most cases a model with $m \leq 6$ provides an adequate fit to score distributions. The two other estimated equating functions using presmoothing methods are based on the beta binomial model and the four-parameter beta binomial model.

Three equipercentile equating functions are based on postsmoothing methods. These three methods correspond to smoothing parameters $s = .50$, $s = .25$ and $s = .10$.

The remaining equating method is linear equating. For linear equating the raw score equivalent on the new form of raw score i on the old form is given by the linear function

$$\frac{\sigma_Y}{\sigma_X} i + \mu_Y - \frac{\sigma_Y}{\sigma_X} \mu_X \,, \tag{5}$$

where μ_X and μ_Y are the means and σ_X and σ_Y are the standard deviations of X and Y. An estimate of the linear equating function is obtained by substituting sample moments for population moments given in Equation 5.

Figures 6 through 10 display the equating functions for the 10 equating methods to be studied for the Licensure test, ACT Reading subscore, ACT English test, ACT Science Reasoning test and PLAN Math test, respectively. The equating functions are computed based on the observed score distributions. In Figures 6, 8, 9 and 10 the population equating functions based on the fitted distributions described previously are also presented (in Figure 9 the population equating function is identical to the equating function given by four-parameter beta binomial model).

Criteria

If $\tilde{e}(i)$ is the estimated old form raw score equivalent to new form raw score i given by a particular equating method and $e(i)$ is the raw score equivalent given by the population equating function, then the mean squared error for the equating method at raw score i is given by

$$E\left[\tilde{e}(i) - e(i)\right]^2, \tag{6}$$

where E stands for expected value (the expected value is over the pair of random variables used to determine $\tilde{e}(i)$). The mean squared error can be written as

$$E\left[\tilde{e}(i) - \mu_{\tilde{e}(i)}\right]^2 + \left[e(i) - \mu_{\tilde{e}(i)}\right]^2, \tag{7}$$

where

$$\mu_{\tilde{e}(i)} \equiv E\left[\tilde{e}(i)\right].$$

The first term in Equation 7 is the variance of $\tilde{e}(i)$ and the second term is the squared bias of $\tilde{e}(i)$.

The average mean squared error for an equating method is given by

$$\sum_{i=0}^{K} E\left[\tilde{e}(i) - e(i)\right]^2 \Pr(X = i). \tag{8}$$

This can be written as the sum of the average variance and average squared bias

$$\sum_{i=0}^{K} E\left[\tilde{e}(i) - \mu_{\tilde{e}(i)}\right]^2 \Pr(X = i) + \sum_{i=0}^{K} \left[e(i) - \mu_{\tilde{e}(i)}\right]^2 \Pr(X = i). \tag{9}$$

For each of the ten equating methods for a particular sample size and pair of population distributions the mean squared error at raw score i is estimated using the 500 pairs of sample distributions as

$$\frac{1}{500} \sum_{s=1}^{500} \left[\tilde{e}_s(i) - e(i)\right]^2, \tag{10}$$

where $\tilde{e}_s(i)$ is the old form raw score equivalent of new form raw score i for sample s. The variance and squared bias of $\tilde{e}(i)$ are estimated in a similar manner. Estimates of the average mean squared error are obtained by substituting the estimates of the mean squared error for each raw score into Equation 8. An estimate of the standard error of these estimates of the average mean squared error is obtained from the usual estimate of the standard error of a mean (the standard deviation divided by $\sqrt{500}$). Estimates of the average variance and average squared bias are obtained analogously using Equation 9 with an estimate of $\mu_{\tilde{e}(i)}$ given by

$$\frac{1}{500}\sum_{s=1}^{500}\tilde{e}_s(i)\,.$$

Results

Estimates of average squared bias and variance (Equation 9), average mean squared error (Equation 8), and the standard error of the average mean squared error for each of the ten equating methods and five sample sizes for population distributions based on the Licensure test, ACT Reading subscore, ACT English test, ACT Science Reasoning test and PLAN Mathematics test are given in Tables 2 through 6, respectively. The average mean squared error, if no equating were performed (i.e., using an identity function as the equating function) are 0.18, 0.20, 57.64, 0.51 and 23.40 for the population equating functions corresponding to Tables 2 through 6, respectively. If a value of average mean squared error for a particular equating method is larger than the corresponding value of average mean squared error under no equating then not equating would be preferable to equating using the equating method in question. In Table 2 none of the equating methods have average mean squared error less than 0.18 for sample sizes of 100 and 250, so that for those sample sizes not equating would be preferable to equating using any of the ten equating methods. In Tables 3 and 5 no equating method has lower average mean squared error than not equating for the sample size of 100. In Tables 4 and 6 equating is preferable to not equating for all sample sizes.

A summary of the results presented in Tables 2 through 6 is presented in Table 7. Table 7 gives, for each pair of population distributions and sample size, the equating method with the lowest average mean squared error from Tables 2 through 6 and all methods with average mean squared errors within two standard errors of this minimum average mean squared error (using the standard error corresponding to the method with the minimum average mean squared error). The two standard error rule is arbitrary and there are some cases, especially for the larger sample sizes,

in which methods not listed in Table 7 we would judge to provide results comparable in practical terms to the methods listed.

The results in Tables 2 through 6 indicate that for all population distributions and sample sizes at least one of the methods of smoothed equipercentile equating produced less equating error than unsmoothed equipercentile equating. Each of the methods of smoothed equipercentile equating and linear equating had an average mean squared error that was close to the minimum average mean squared error in at least one case. In other words, each of the methods of smoothed equipercentile equating and linear equating worked well for at least one sample size by population distribution combination.

Average variance tends to dominate average squared bias for lower sample sizes. Consequently, for lower sample sizes the methods that tend to have lower average variance (at the expense of higher average squared bias) tend to perform best. At higher sample sizes variance decreases and bias becomes more of a factor in the average mean squared error. Consequently, at higher sample sizes the bias of a method is an important factor in the performance of the method. Because the bias of a method depends on the population equating function, different methods perform best with different population equating functions for larger sample sizes.

Figures 11 through 15 present estimates of the mean squared error by score point (Equation 6) for eight of the equating methods for sample sizes of 250 and 1000 for the population distributions based on the Licensure test, ACT Reading subscore, ACT English test, ACT Science Reasoning test and PLAN Mathematics test, respectively. Mean squared errors are not presented in Figures 11 through 15 for the presmoothing method based on the Log-Linear model of degree 4 and the postsmoothing method using a smoothing parameter of .25 due to the similarity of the curves for these methods to the curves that are presented in the figures. The range of raw scores given in Figures 11 through 15 exclude scores at the bottom of the scale with low probabilities of occurrence in the population taking the new form. The general level of the mean squared error curves reflects the average mean squared errors given in Tables 2 through 6.

Discussion

The results provide evidence that both presmoothing and postsmoothing methods can improve estimation of the equipercentile equating function in the random groups design. An improvement in the estimation of the equipercentile equating function resulted from smoothing for all population distributions and sample sizes considered.

The results indicated that presmoothing and postsmoothing methods can produce results comparable with one another. The results do not support a conclusion that either presmoothing or postsmoothing methods should be preferred in all cases. The only cases in which there was not a postsmoothing method that performed as about as well as the best performing presmoothing method, or vice versa, was for sample sizes of 100 where, in a couple of cases, the beta binomial presmoothing method performed significantly better than any of the three postsmoothing methods. It is possible that if postsmoothing methods with a larger smoothing parameter had been included in this study at least one of these methods would have performed as well as the beta binomial method in these cases.

Presmoothing based on the beta binomial model (along with linear equating in some cases) resulted in the smallest equating error when the sample sizes were small. In these cases the bias introduced by the beta binomial and linear methods was not large relative to the variance of the other smoothed equipercentile methods, although there were cases in which linear equating resulted in large equating error for even small sample sizes when the bias of linear equating was large (e.g., the ACT English test and PLAN Mathematics test). With large sample sizes the bias of the beta binomial method was typically large compared to the variance of the other smoothed equipercentile methods resulting in larger average mean squared error for the beta binomial method relative to other smoothed equipercentile methods in this situation. This result suggests other presmoothing methods will probably result in smaller equating error than using the beta binomial model for large sample sizes.

For the method based on the log-linear model, adding parameters to the model will generally result in lower bias but greater variance of the resulting estimated equating functions. For smaller sample sizes a model with fewer parameters can result in lower average mean squared error than a model with more parameters if the bias introduced by the simpler model is small compared to the variance of the more complex model. Effects analogous to adding parameters to the log-linear model are achieved by decreasing the smoothing parameter in postsmoothing.

It is likely that an important factor in the performance of the presmoothing methods in practical situations is the appropriateness of the models used for smoothing the score distributions. In using any of the presmoothing methods, the fit of the model to the raw score distributions should be evaluated. A necessary condition for using a particular presmoothing method in practice would be that the model fit the data well. Assessment of model fit may involve formal tests of model fit

(chi-squared goodness of fit statistics) and informal analyses of model fit such as residual analyses and various graphical displays.

The results of this study and practical experience with data from several testing programs indicate that presmoothing based on the four-parameter beta binomial model and the log-linear model will usually provide an adequate fit to observed distributions of test scores if the sample sizes are large enough (around 1000 or more). The log-linear model has an advantage over the four-parameter beta binomial model in that it can potentially fit a wider class of distributions than the four-parameter beta binomial model. The cost involved in this greater flexibility is that a model selection process must be used to choose a particular log-linear model to use. In this paper three fixed log-linear models were used. In applied settings the user would likely evaluate several log-linear models and pick the simplest model that fit the data adequately. Haberman (1974) discusses model selection for models such as those given in Equation 3. Agresti (1990, Chapter 7) discusses some general methods for selecting log-linear models.

The process of selecting a log-linear model, or selecting a smoothing parameter in cubic spline postsmoothing, could introduce errors that were not present in the results reported in this paper. For example, Hanson (1990) compared smoothing of univariate test score distributions based on the four-parameter beta binomial model and the log-linear model given in Equation 2. In Hanson (1990) a model selection process was used for each sample to select a log-linear model to use. It was found that the four parameter beta binomial model provided more accurate results than the log-linear model for all sample sizes less than 5000. It is likely the log-linear model would have performed better in Hanson (1990) if a procedure like the one followed in this paper had been used with a fixed model being used for all samples. Conversely, the accuracy of the log-linear model in this paper may have been less if some model selection procedure had been used for each sample to select a model. Similarly, the accuracy of postsmoothing when a smoothing parameter is chosen for each sample may be less than the procedures studied in this paper in which the same smoothing parameter was used for each sample.

The assumption made for the beta binomial and four-parameter beta binomial models that the conditional error distribution is binomial may not in many cases be very realistic. It has been our experience that using the more general conditional error distribution given by Lord's two-parameter approximation to a compound binomial distributions (Lord, 1965) does not seem to improve the fit of the model using a four-parameter beta true score distribution in the examples we have looked

at. Consequently, for the purpose of smoothing a univariate raw score distribution the assumption of a binomial error distribution seems adequate. If an estimate of the true score distribution is needed (other than for simply computing the estimated observed score distribution) a more realistic error distribution than the binomial distribution should probably be used in most cases (e.g., Lord's two-parameter approximation to a compound binomial distribution).

C language source code for functions that compute all the equating function estimates discussed in this paper is available from the first author.

References

Agresti, A. (1990). *Categorical data analysis.* New York: John Wiley and Sons.

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.

Cope, R. T., & Kolen, M. J. (1990). *A study of methods for estimating distributions of test scores.* American College Testing Research Report 90-5. Iowa City, IA: American College Testing.

Fairbank, B. A. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. *Applied Psychological Measurement, 11*, 245-262.

Hanson, B. A. (1990). *An investigation of methods for improving estimation of test score distributions.* American College Testing Research Report 90-4. Iowa City, IA: American College Testing.

Hanson, B. A. (1991). *Method of moments estimates of the four-parameter beta compound binomial model and the calculation of classification consistency indices.* American College Testing Research Report 91-5. Iowa City, IA: American College Testing.

Haberman, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics. 30*, 589-600.

Holland, P. W. & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions.* Educational Testing Service Research Report 87-31. Princeton, NJ: Educational Testing Service.

Holland, P. W. & Thayer, D. T. (1989). *The kernel method of equating score distributions.* Educational Testing Service Research Report 89-7. Princeton, NJ: Educational Testing Service.

Kolen, M. J. (1984). Effectiveness of analytical smoothing in equipercentile equating. *Journal of Educational Statistics. 9*, 25-44.

Kolen, M. J. & Jarjoura, D. (1987). Analytic smoothing for equipercentile equating under the common item nonequivalent populations design. *Psychometika. 52*, 43-59.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30*, 239-270.

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

Table 1. Descriptive Statistics for Observed Data.

Licensure Test (30 items)

	Mean	s.d.	Skewness	Kurtosis	Sample Size
New Form	18.88	3.68	-0.13	2.79	38,765
Old Form	19.16	3.43	-0.31	3.05	38,765

ACT Reading (20 items)

	Mean	s.d.	Skewness	Kurtosis	Sample Size
New Form	12.30	3.76	-0.21	3.40	82,073
Old Form	12.69	3.58	-0.29	2.54	83,709

ACT English (75 items)

	Mean	s.d.	Skewness	Kurtosis	Sample Size
New Form	52.50	12.10	-0.37	2.55	3,158
Old Form	45.07	12.99	-0.02	2.25	3,293

ACT Science Reasoning (40 items)

	Mean	s.d.	Skewness	Kurtosis	Sample Size
New Form	21.63	5.70	0.06	2.63	2,831
Old Form	22.17	5.38	0.23	2.55	2,898

P-ACT+ Mathematics (40 items)

	Mean	s.d.	Skewness	Kurtosis	Sample Size
New Form	19.54	7.72	0.28	2.28	2,190
Old Form	14.93	6.79	0.81	3.42	2,176

Table 2. Average Squared Bias, Variance and Mean Squared Error for the Licensure Test.

	Sample Size = 100				Sample Size = 250			
	Squared Bias	Var.	MSE	s.e.(MSE)	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.006	0.590	0.597	0.023	0.002	0.258	0.260	0.009
Linear	0.028	0.373	0.401	0.020	0.027	0.155	0.181	0.008
Beta Binomial	0.027	0.369	0.395	0.019	0.026	0.154	0.180	0.007
4-para. Beta Binomial	0.004	0.466	0.470	0.021	0.002	0.198	0.200	0.008
Log-Linear 3	0.004	0.444	0.447	0.022	0.001	0.185	0.187	0.008
Log-Linear 4	0.006	0.534	0.540	0.025	0.001	0.216	0.217	0.008
Log-Linear 6	0.009	0.614	0.623	0.028	0.003	0.249	0.252	0.010
Post Smooth 0.50	0.026	0.469	0.496	0.026	0.010	0.196	0.206	0.009
Post Smooth 0.25	0.017	0.477	0.494	0.025	0.006	0.200	0.207	0.009
Post Smooth 0.10	0.013	0.507	0.520	0.024	0.005	0.214	0.220	0.009

	Sample Size = 500				Sample Size = 1000			
	Squared Bias	Var.	MSE	s.e.(MSE)	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.003	0.145	0.148	0.005	0.002	0.067	0.068	0.002
Linear	0.027	0.081	0.109	0.004	0.027	0.034	0.061	0.002
Beta Binomial	0.027	0.082	0.109	0.004	0.027	0.034	0.061	0.002
4-para. Beta Binomial	0.002	0.108	0.111	0.005	0.002	0.046	0.048	0.002
Log-Linear 3	0.002	0.100	0.102	0.004	0.001	0.042	0.043	0.002
Log-Linear 4	0.002	0.122	0.125	0.005	0.001	0.053	0.054	0.002
Log-Linear 6	0.004	0.138	0.142	0.005	0.002	0.061	0.063	0.002
Post Smooth 0.50	0.011	0.109	0.120	0.005	0.009	0.042	0.051	0.002
Post Smooth 0.25	0.007	0.111	0.118	0.005	0.005	0.044	0.048	0.002
Post Smooth 0.10	0.006	0.118	0.123	0.005	0.004	0.048	0.051	0.002

	Sample Size = 3000			
	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.001	0.024	0.025	0.001
Linear	0.027	0.012	0.038	0.001
Beta Binomial	0.026	0.011	0.038	0.001
4-para. Beta Binomial	0.001	0.015	0.016	0.001
Log-Linear 3	0.001	0.014	0.015	0.001
Log-Linear 4	0.001	0.019	0.020	0.001
Log-Linear 6	0.001	0.021	0.022	0.001
Post Smooth 0.50	0.007	0.014	0.021	0.001
Post Smooth 0.25	0.004	0.015	0.019	0.001
Post Smooth 0.10	0.003	0.016	0.019	0.001

Table 3. Average Squared Bias, Variance and Mean Squared Error for the ACT Reading Subscore.

	Sample Size = 100				Sample Size = 250			
	Squared Bias	Var.	MSE	s.e.(MSE)	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.004	0.506	0.511	0.020	0.001	0.225	0.226	0.008
Linear	0.010	0.336	0.345	0.020	0.007	0.146	0.153	0.008
Beta Binomial	0.004	0.340	0.344	0.020	0.002	0.149	0.152	0.008
4-para. Beta Binomial	0.004	0.418	0.423	0.020	0.001	0.182	0.183	0.008
Log-Linear 3	0.003	0.392	0.396	0.020	0.001	0.171	0.172	0.008
Log-Linear 4	0.004	0.432	0.437	0.021	0.001	0.189	0.190	0.008
Log-Linear 6	0.005	0.477	0.482	0.021	0.002	0.207	0.208	0.008
Post Smooth 0.50	0.013	0.326	0.339	0.021	0.005	0.153	0.158	0.008
Post Smooth 0.25	0.008	0.360	0.368	0.020	0.002	0.165	0.167	0.008
Post Smooth 0.10	0.006	0.414	0.420	0.020	0.002	0.184	0.186	0.008

	Sample Size = 500				Sample Size = 1000			
	Squared Bias	Var.	MSE	s.e.(MSE)	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.001	0.109	0.110	0.004	0.000	0.058	0.058	0.002
Linear	0.007	0.070	0.077	0.004	0.007	0.038	0.045	0.002
Beta Binomial	0.003	0.071	0.074	0.004	0.002	0.037	0.040	0.002
4-para. Beta Binomial	0.001	0.087	0.088	0.004	0.000	0.047	0.047	0.002
Log-Linear 3	0.001	0.081	0.082	0.004	0.001	0.043	0.044	0.002
Log-Linear 4	0.001	0.090	0.091	0.004	0.000	0.048	0.049	0.002
Log-Linear 6	0.001	0.100	0.101	0.004	0.000	0.052	0.053	0.002
Post Smooth 0.50	0.005	0.071	0.076	0.004	0.004	0.037	0.041	0.002
Post Smooth 0.25	0.002	0.078	0.080	0.004	0.002	0.041	0.043	0.002
Post Smooth 0.10	0.001	0.088	0.089	0.004	0.001	0.047	0.047	0.002

	Sample Size = 3000			
	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.000	0.018	0.019	0.001
Linear	0.007	0.012	0.019	0.001
Beta Binomial	0.002	0.011	0.014	0.001
4-para. Beta Binomial	0.000	0.015	0.015	0.001
Log-Linear 3	0.001	0.013	0.014	0.001
Log-Linear 4	0.000	0.015	0.015	0.001
Log-Linear 6	0.000	0.016	0.017	0.001
Post Smooth 0.50	0.003	0.012	0.015	0.001
Post Smooth 0.25	0.001	0.013	0.015	0.001
Post Smooth 0.10	0.001	0.015	0.015	0.001

Table 4. Average Squared Bias, Variance and Mean Squared Error for the ACT English Test.

Sample Size = 100

	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.015	6.583	6.598	0.223
Linear	1.853	4.295	6.148	0.221
Beta Binomial	0.217	4.258	4.475	0.212
4-para. Beta Binomial	0.081	5.198	5.280	0.220
Log-Linear 3	0.268	4.936	5.204	0.222
Log-Linear 4	0.192	5.469	5.661	0.224
Log-Linear 6	0.086	6.007	6.094	0.226
Post Smooth 0.50	0.023	5.144	5.167	0.215
Post Smooth 0.25	0.018	5.555	5.573	0.219
Post Smooth 0.10	0.018	5.967	5.985	0.221

Sample Size = 250

	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.010	2.822	2.832	0.088
Linear	1.852	1.795	3.647	0.082
Beta Binomial	0.238	1.818	2.056	0.084
4-para. Beta Binomial	0.073	2.167	2.240	0.088
Log-Linear 3	0.203	2.093	2.296	0.087
Log-Linear 4	0.178	2.291	2.469	0.088
Log-Linear 6	0.047	2.505	2.552	0.089
Post Smooth 0.50	0.044	2.149	2.193	0.085
Post Smooth 0.25	0.023	2.321	2.343	0.086
Post Smooth 0.10	0.017	2.531	2.548	0.088

Sample Size = 500

	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.004	1.493	1.497	0.057
Linear	1.843	0.955	2.798	0.053
Beta Binomial	0.223	0.969	1.192	0.055
4-para. Beta Binomial	0.067	1.150	1.216	0.057
Log-Linear 3	0.178	1.110	1.288	0.057
Log-Linear 4	0.170	1.215	1.386	0.057
Log-Linear 6	0.029	1.302	1.331	0.057
Post Smooth 0.50	0.039	1.128	1.167	0.057
Post Smooth 0.25	0.016	1.211	1.227	0.056
Post Smooth 0.10	0.009	1.326	1.335	0.057

Sample Size = 1000

	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.002	0.750	0.752	0.025
Linear	1.844	0.486	2.331	0.024
Beta Binomial	0.220	0.489	0.709	0.024
4-para. Beta Binomial	0.065	0.562	0.627	0.024
Log-Linear 3	0.160	0.551	0.711	0.026
Log-Linear 4	0.170	0.602	0.772	0.024
Log-Linear 6	0.025	0.648	0.673	0.024
Post Smooth 0.50	0.034	0.556	0.591	0.024
Post Smooth 0.25	0.016	0.600	0.616	0.024
Post Smooth 0.10	0.007	0.659	0.666	0.024

Sample Size = 3000

	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.002	0.246	0.248	0.009
Linear	1.843	0.157	2.000	0.008
Beta Binomial	0.221	0.159	0.380	0.009
4-para. Beta Binomial	0.060	0.184	0.244	0.009
Log-Linear 3	0.167	0.180	0.347	0.009
Log-Linear 4	0.167	0.195	0.362	0.009
Log-Linear 6	0.021	0.210	0.232	0.009
Post Smooth 0.50	0.027	0.186	0.214	0.009
Post Smooth 0.25	0.014	0.198	0.212	0.009
Post Smooth 0.10	0.007	0.213	0.220	0.009

Table 5. Average Squared Bias, Variance and Mean Squared Error for the ACT Science Reasoning Test.

	Sample Size = 100				Sample Size = 250			
	Squared Bias	Var.	MSE	s.e.(MSE)	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.008	1.610	1.618	0.058	0.004	0.699	0.703	0.024
Linear	0.019	1.017	1.035	0.056	0.016	0.441	0.457	0.023
Beta Binomial	0.023	1.005	1.028	0.055	0.021	0.439	0.461	0.023
4-para. Beta Binomial	0.007	1.272	1.279	0.058	0.003	0.544	0.547	0.024
Log-Linear 3	0.013	1.161	1.174	0.057	0.008	0.504	0.511	0.024
Log-Linear 4	0.011	1.331	1.342	0.058	0.004	0.568	0.572	0.024
Log-Linear 6	0.012	1.511	1.524	0.060	0.004	0.623	0.627	0.024
Post Smooth 0.50	0.023	1.232	1.256	0.059	0.009	0.501	0.510	0.023
Post Smooth 0.25	0.013	1.303	1.316	0.058	0.006	0.557	0.563	0.024
Post Smooth 0.10	0.008	1.412	1.420	0.057	0.004	0.622	0.626	0.025

	Sample Size = 500				Sample Size = 1000			
	Squared Bias	Var.	MSE	s.e.(MSE)	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.001	0.322	0.322	0.010	0.001	0.168	0.168	0.005
Linear	0.014	0.190	0.205	0.009	0.015	0.101	0.115	0.005
Beta Binomial	0.019	0.190	0.209	0.009	0.019	0.100	0.119	0.005
4-para. Beta Binomial	0.001	0.237	0.237	0.010	0.000	0.123	0.123	0.005
Log-Linear 3	0.006	0.220	0.225	0.009	0.006	0.116	0.122	0.005
Log-Linear 4	0.001	0.248	0.249	0.010	0.000	0.129	0.129	0.005
Log-Linear 6	0.001	0.277	0.277	0.010	0.000	0.142	0.143	0.005
Post Smooth 0.50	0.006	0.211	0.216	0.009	0.006	0.106	0.112	0.005
Post Smooth 0.25	0.002	0.241	0.243	0.010	0.002	0.122	0.123	0.005
Post Smooth 0.10	0.001	0.275	0.276	0.010	0.001	0.140	0.141	0.005

	Sample Size = 3000			
	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.000	0.057	0.057	0.002
Linear	0.014	0.034	0.048	0.002
Beta Binomial	0.019	0.034	0.052	0.002
4-para. Beta Binomial	0.000	0.041	0.041	0.002
Log-Linear 3	0.005	0.039	0.044	0.002
Log-Linear 4	0.000	0.043	0.043	0.002
Log-Linear 6	0.000	0.047	0.048	0.002
Post Smooth 0.50	0.006	0.036	0.043	0.002
Post Smooth 0.25	0.002	0.041	0.043	0.002
Post Smooth 0.10	0.000	0.046	0.047	0.002

Table 6. Average Squared Bias, Variance and Mean Squared Error for the PLAN Math Test.

Sample Size = 100

	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.009	2.280	2.289	0.082
Linear	1.099	1.332	2.431	0.067
Beta Binomial	0.350	1.391	1.741	0.072
4-para. Beta Binomial	0.054	1.835	1.889	0.081
Log-Linear 3	0.196	1.762	1.959	0.083
Log-Linear 4	0.069	1.920	1.989	0.082
Log-Linear 6	0.018	2.067	2.085	0.082
Post Smooth 0.50	0.077	1.774	1.851	0.090
Post Smooth 0.25	0.038	1.809	1.846	0.083
Post Smooth 0.10	0.028	1.913	1.942	0.081

Sample Size = 250

	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.006	0.960	0.966	0.032
Linear	1.110	0.534	1.643	0.026
Beta Binomial	0.363	0.561	0.924	0.029
4-para. Beta Binomial	0.054	0.739	0.793	0.031
Log-Linear 3	0.149	0.713	0.862	0.032
Log-Linear 4	0.060	0.773	0.834	0.031
Log-Linear 6	0.009	0.849	0.858	0.031
Post Smooth 0.50	0.056	0.724	0.780	0.033
Post Smooth 0.25	0.021	0.748	0.769	0.032
Post Smooth 0.10	0.012	0.807	0.819	0.031

Sample Size = 500

	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.001	0.520	0.521	0.020
Linear	1.103	0.297	1.400	0.017
Beta Binomial	0.346	0.315	0.661	0.019
4-para. Beta Binomial	0.045	0.403	0.447	0.020
Log-Linear 3	0.190	0.387	0.577	0.021
Log-Linear 4	0.048	0.421	0.469	0.020
Log-Linear 6	0.002	0.457	0.458	0.020
Post Smooth 0.50	0.033	0.408	0.441	0.021
Post Smooth 0.25	0.008	0.419	0.427	0.020
Post Smooth 0.10	0.001	0.446	0.447	0.020

Sample Size = 1000

	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.001	0.228	0.229	0.008
Linear	1.106	0.122	1.228	0.007
Beta Binomial	0.350	0.128	0.478	0.008
4-para. Beta Binomial	0.055	0.169	0.224	0.008
Log-Linear 3	0.179	0.161	0.341	0.010
Log-Linear 4	0.049	0.174	0.223	0.008
Log-Linear 6	0.001	0.194	0.195	0.008
Post Smooth 0.50	0.033	0.171	0.204	0.008
Post Smooth 0.25	0.012	0.176	0.188	0.008
Post Smooth 0.10	0.003	0.189	0.192	0.008

Sample Size = 3000

	Squared Bias	Var.	MSE	s.e.(MSE)
Unsmoothed	0.000	0.083	0.084	0.003
Linear	1.103	0.045	1.148	0.002
Beta Binomial	0.343	0.048	0.391	0.003
4-para. Beta Binomial	0.055	0.063	0.118	0.003
Log-Linear 3	0.180	0.060	0.240	0.005
Log-Linear 4	0.050	0.064	0.114	0.003
Log-Linear 6	0.001	0.072	0.073	0.003
Post Smooth 0.50	0.016	0.064	0.081	0.003
Post Smooth 0.25	0.007	0.066	0.073	0.003
Post Smooth 0.10	0.002	0.070	0.072	0.003

Table 7. Best Performing Equating Methods in Terms of MSE and Equating Methods within 2 Standard Errors of the Best Method.

	Sample Size				
	100	250	500	1000	3000
Licensure	BB,L	BB,L,LL3	LL3,BB,L	LL3	LL3,BB4
ACT Reading	P1,BB,L,P2	BB,L,P1,P2	BB,L,P1,P2,LL3	BB,P1,P2,LL3	BB,LL3,P,LL4,BB4
ACT English	BB	BB,P1	P1,BB,BB4,P2	P1,P2,BB4	P
ACT Science	BB,L	L,BB	L,BB,P1	P1,L,BB,LL3	BB4,P1,P2,LL4,LL3
P-ACT Math	BB,P2,P1	P2,P1,BB4,P3	P2,P1,BB4,P3,LL6	P,LL6	P3,P2,LL6

L = linear
BB = beta binomial
BB4 = four-parameter beta binoimal
LL3 = log-linear 3
LL4 = log-linear 4
LL6 - log-linear 6
P1 = postsmoothing .50
P2 = postsmoothing .25
P3 = postsmoothing .10
P = all three postsmoothing methods

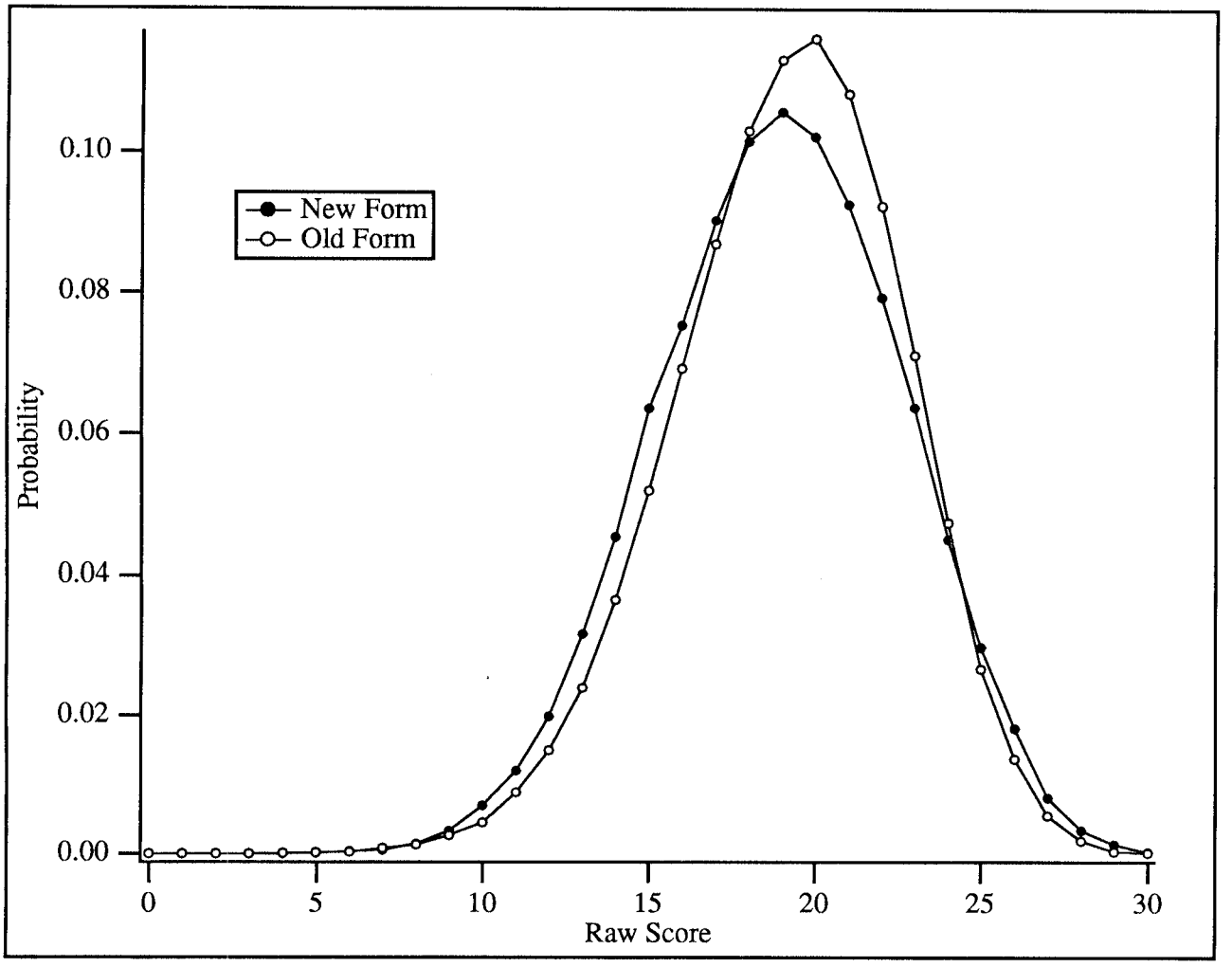

Figure 1. Observed distributions for Licensure test.

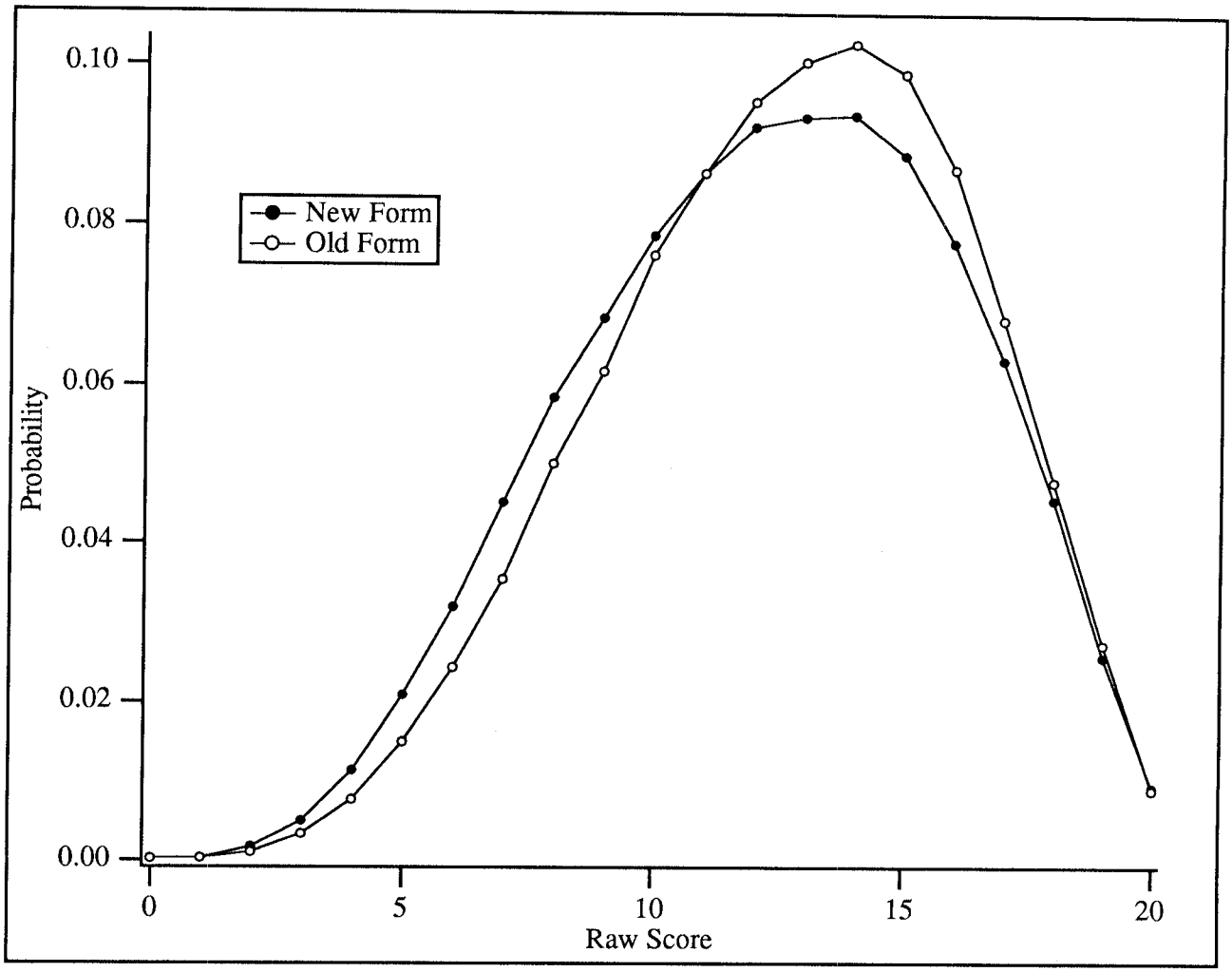

Figure 2. Observed distributions for ACT Reading subscore.

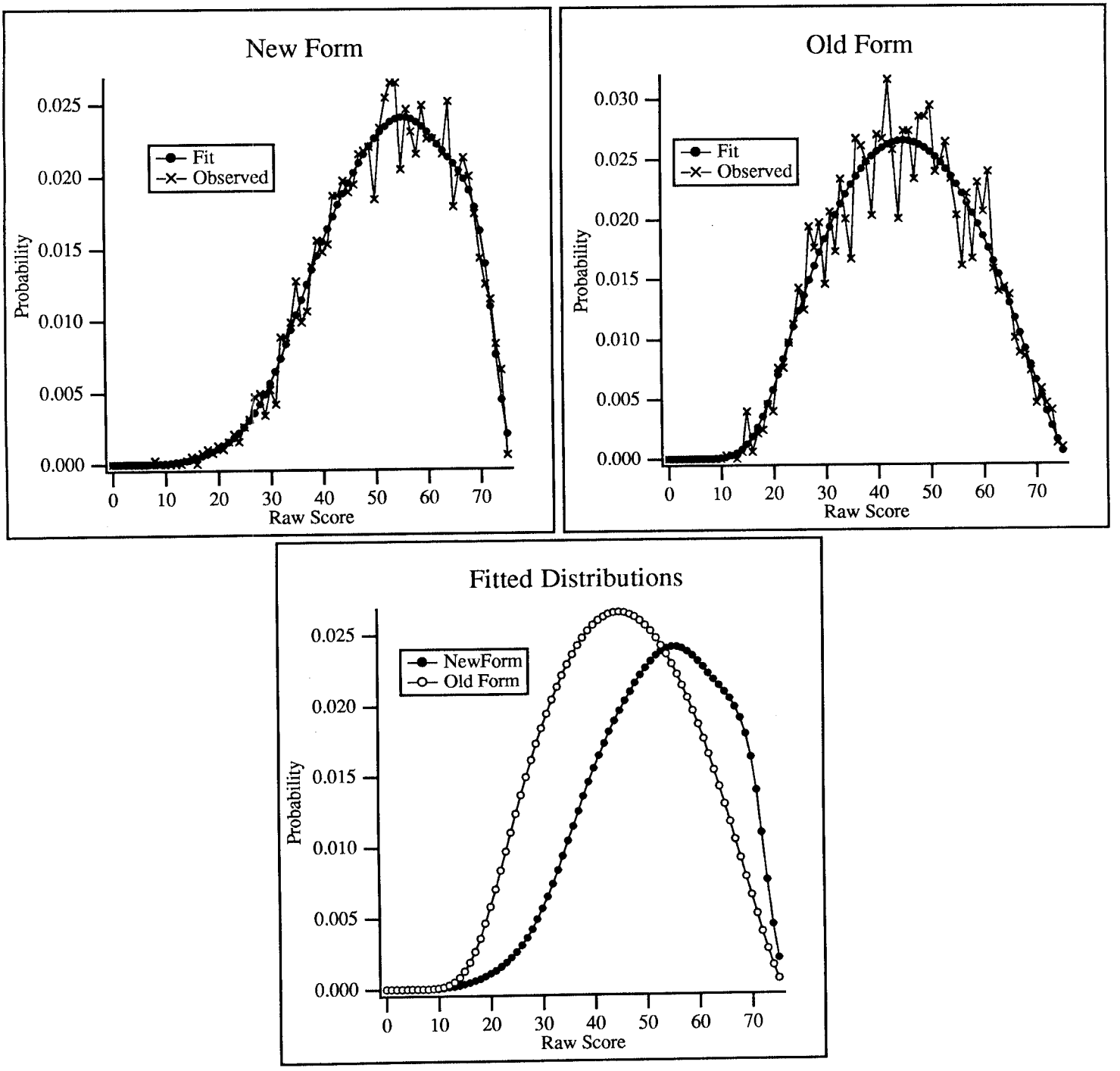

Figure 3. Raw Score Distributions for ACT English test.

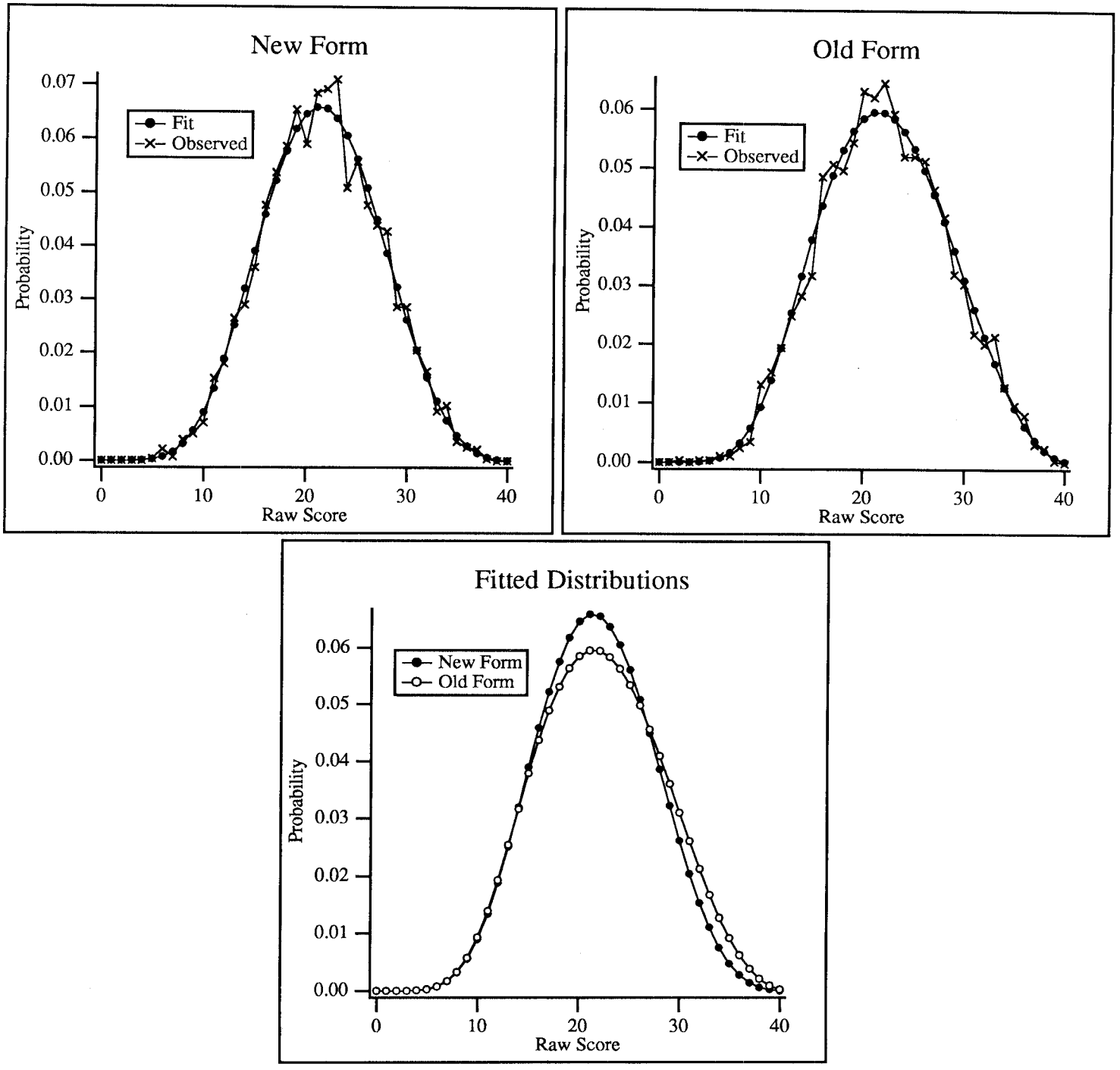

Figure 4. Raw Score Distributions for ACT Science Reasoning test.

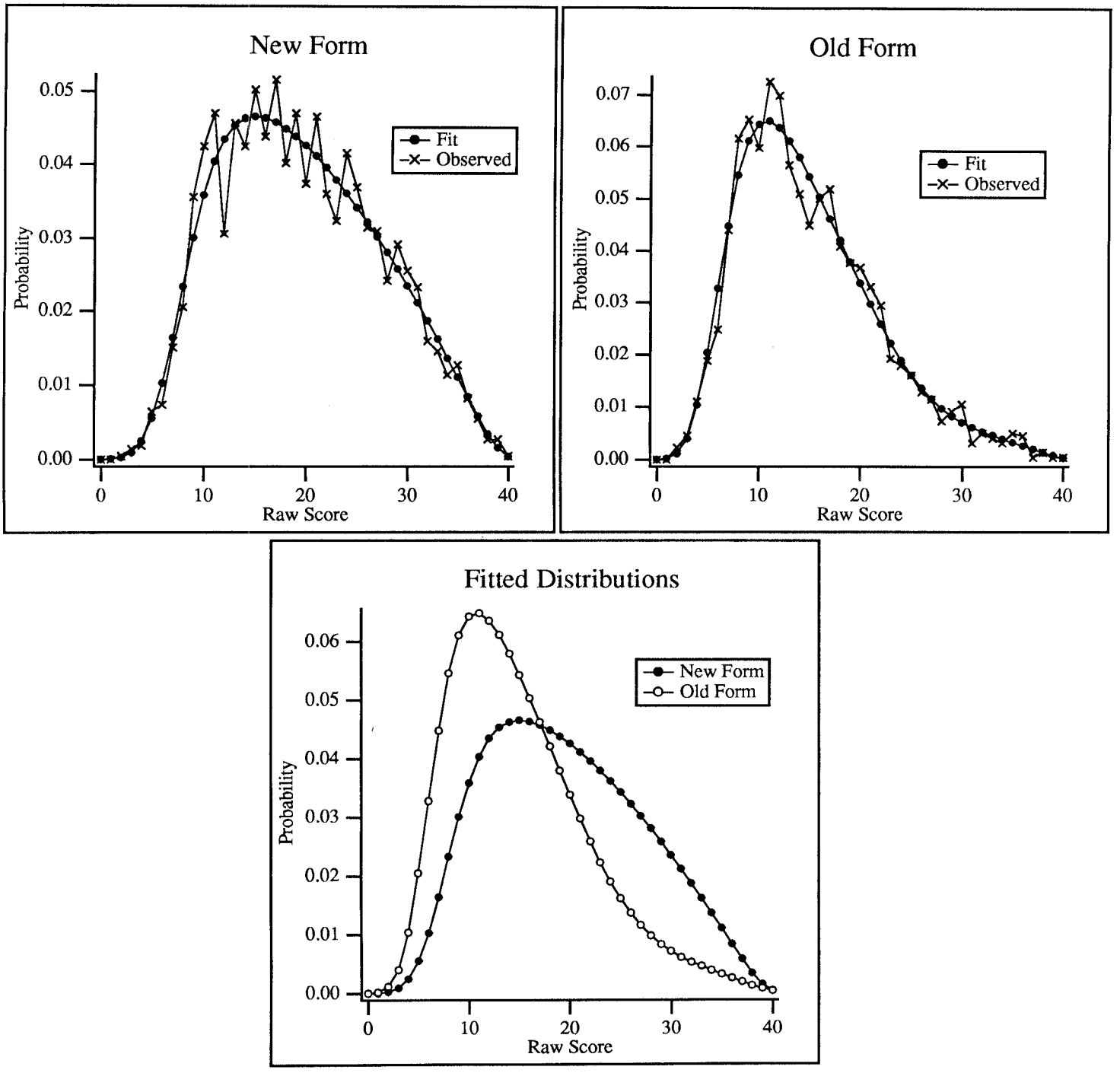

Figure 5. Raw Score Distributions for PLAN Mathematics test.

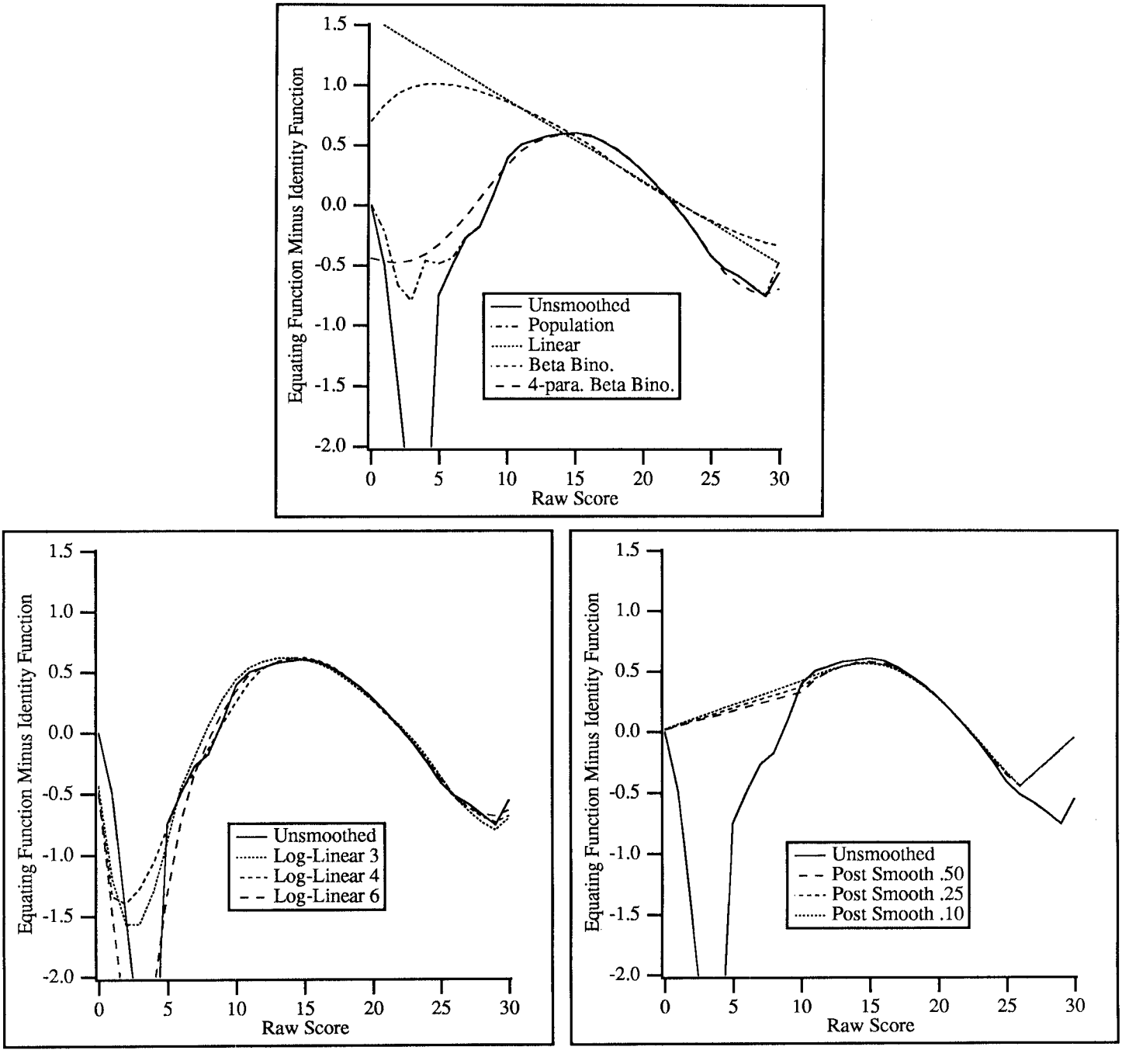

Figure 6. Equating functions for Licensure test using observed data.

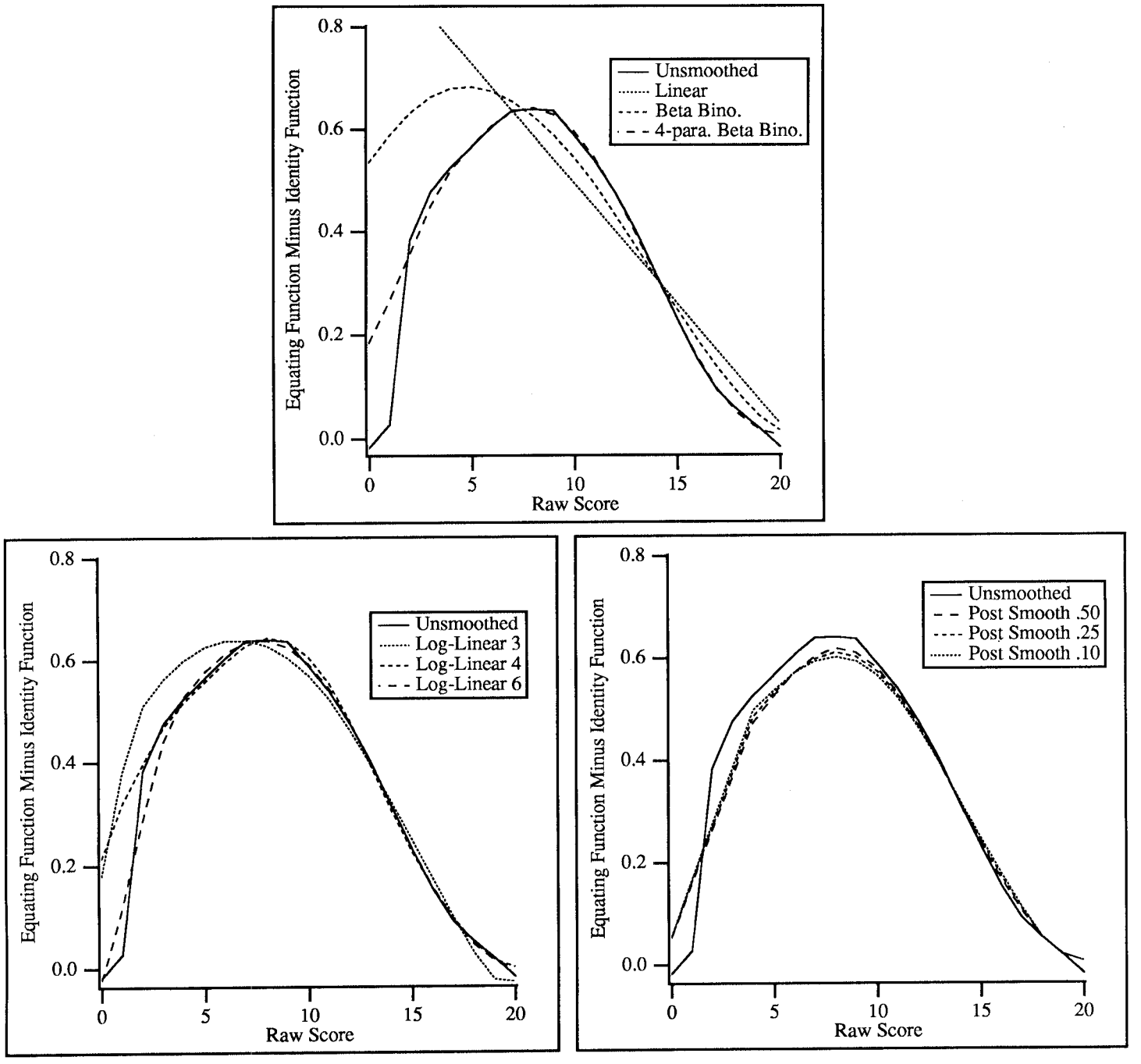

Figure 7. Equating functions for ACT Reading subscore using observed data.

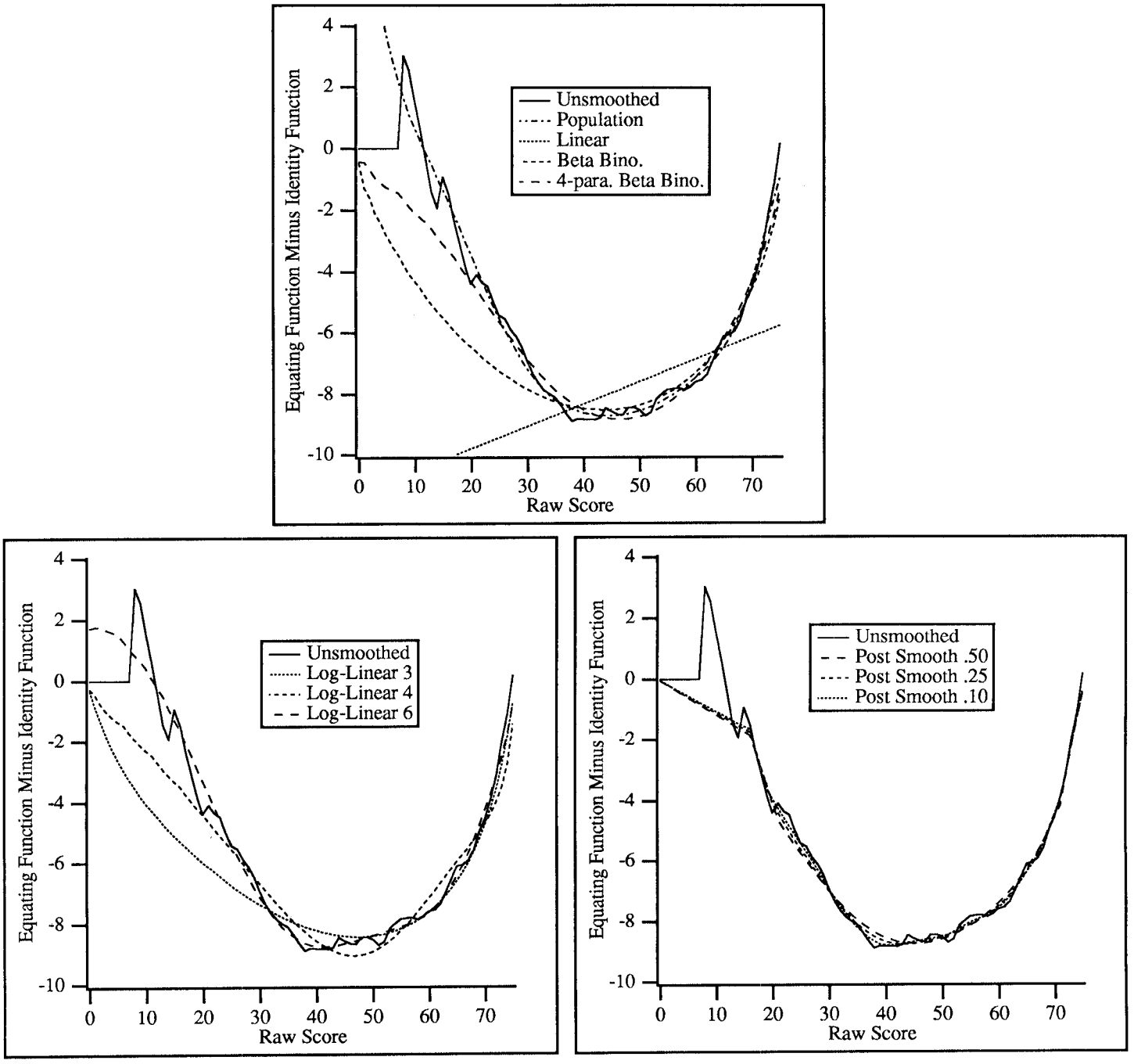

Figure 8. Equating functions for ACT English test using observed data.

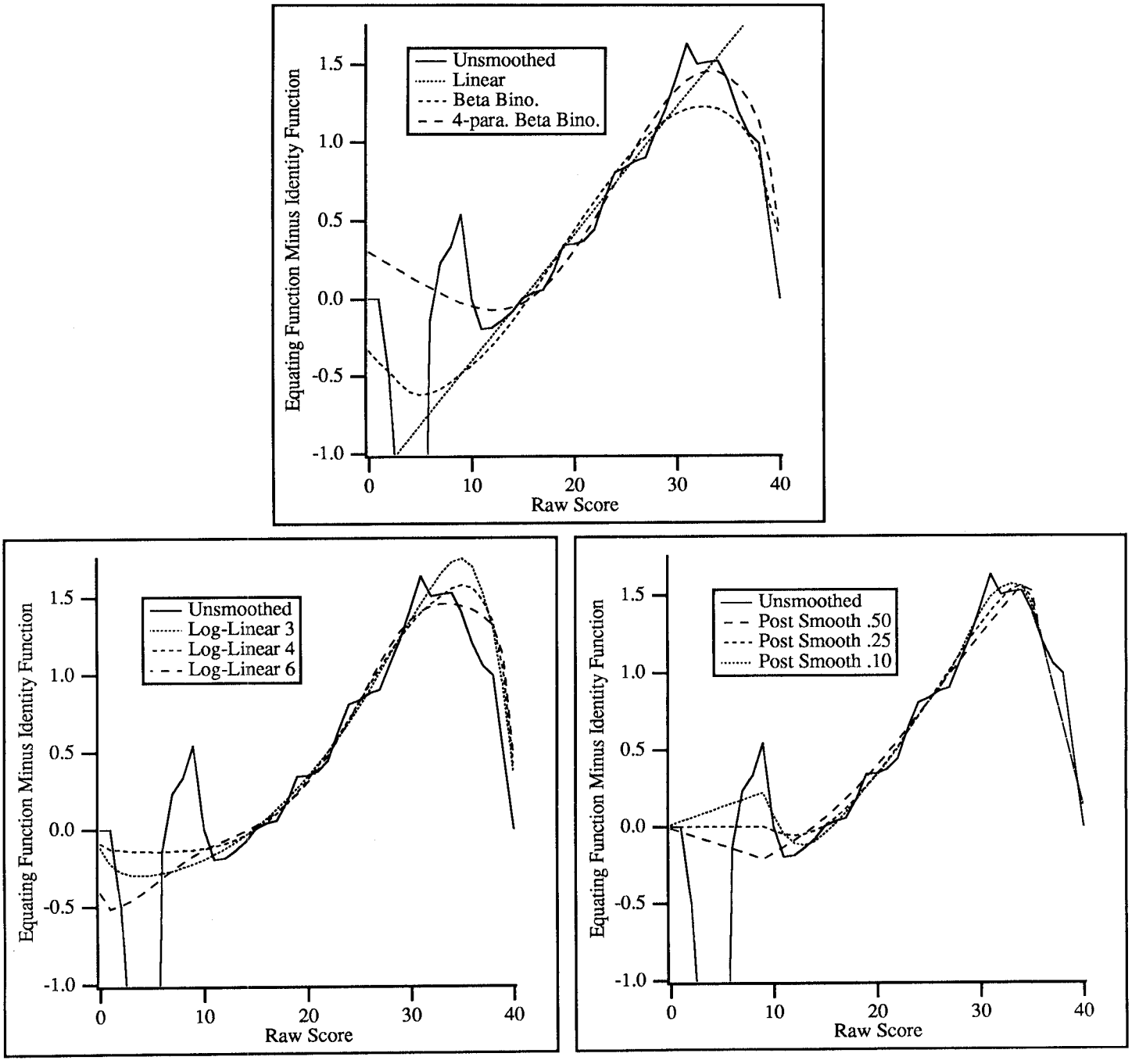

Figure 9. Equating functions for ACT Science Reasoning test using observed data.

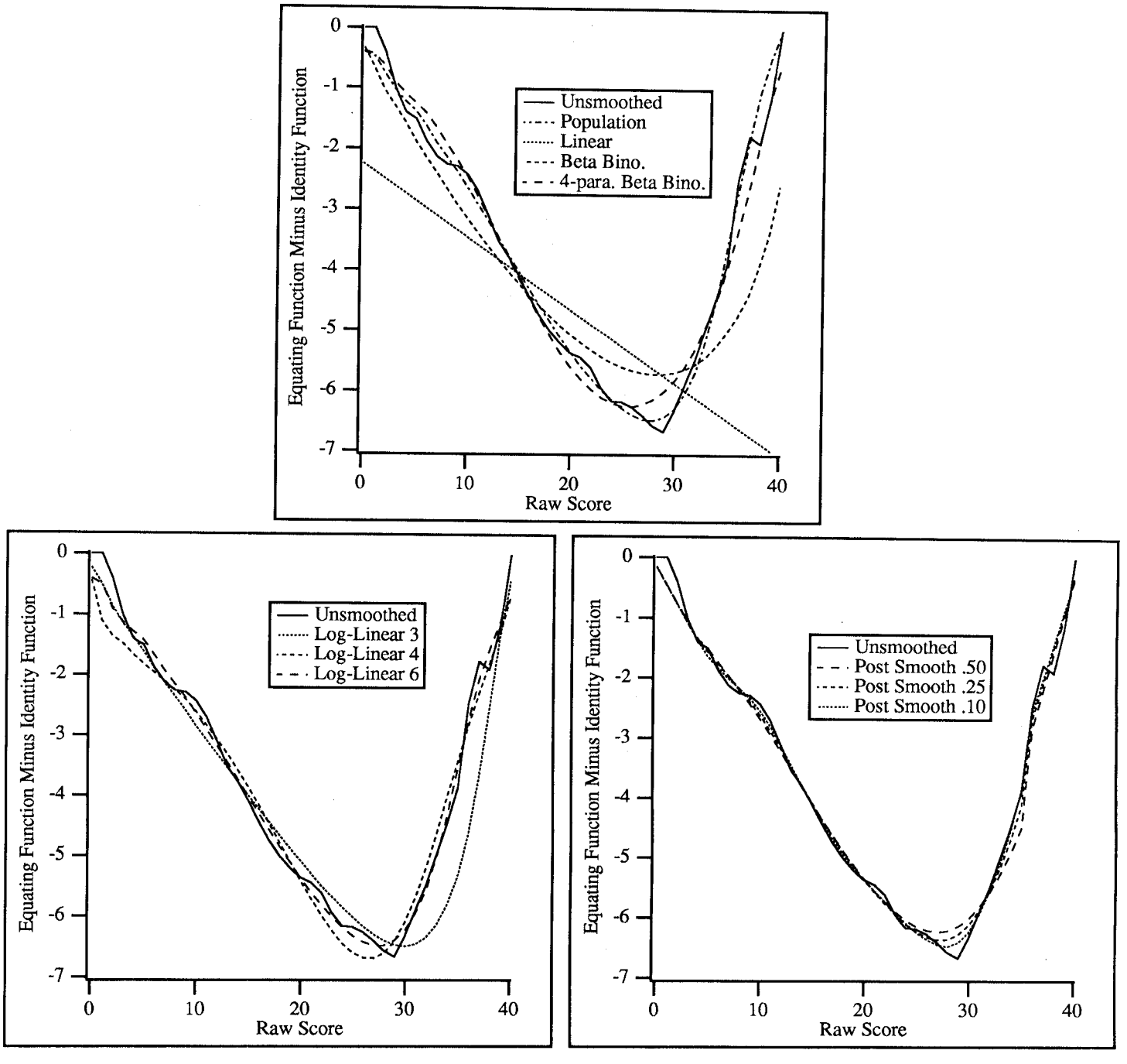

Figure 10. Equating functions for PLAN Mathematics test using observed data.

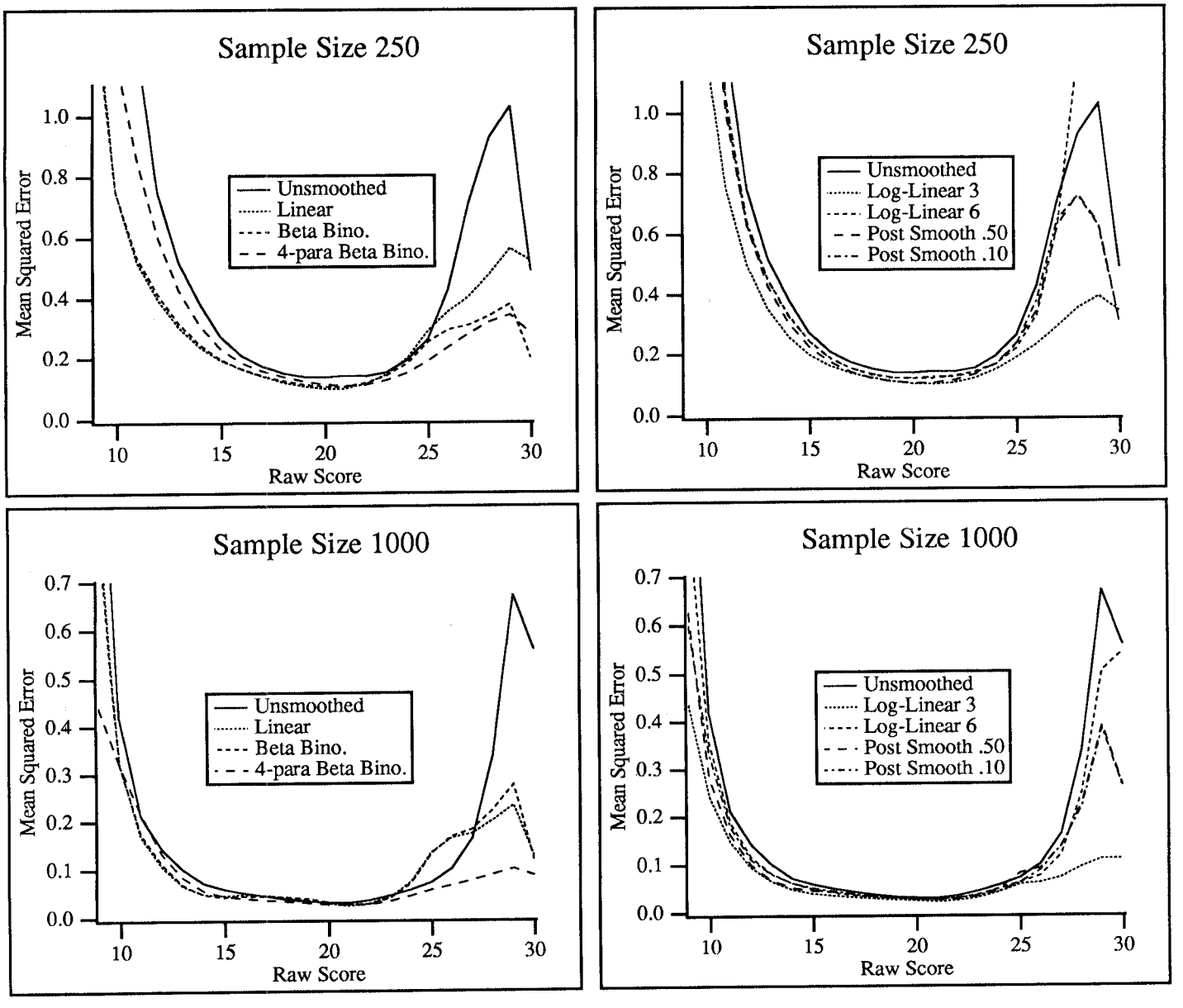

Figure 11. Mean squared error of equating methods for Licensure test.

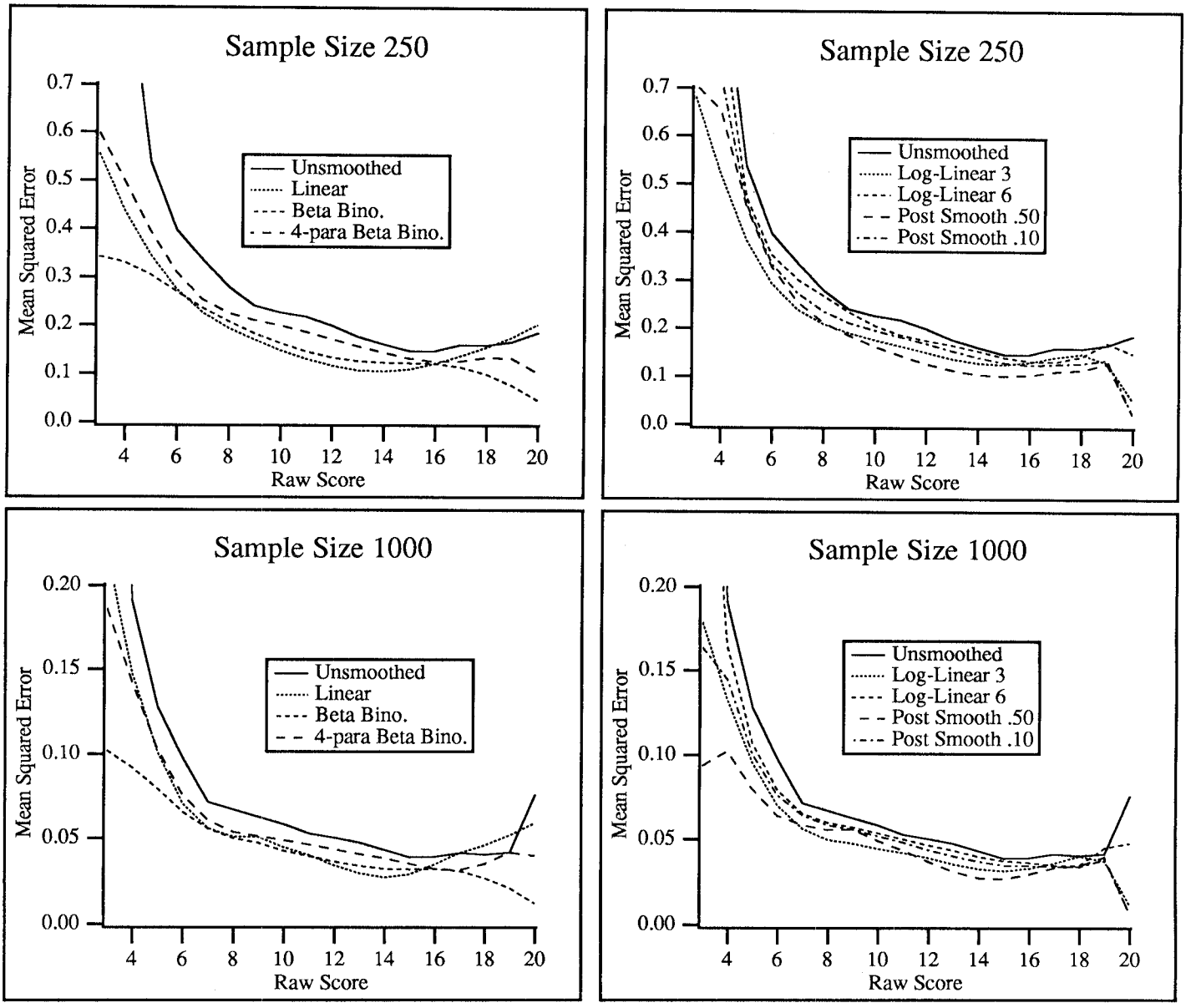

Figure 12. Mean squared error of equating methods for ACT Reading Subscore.

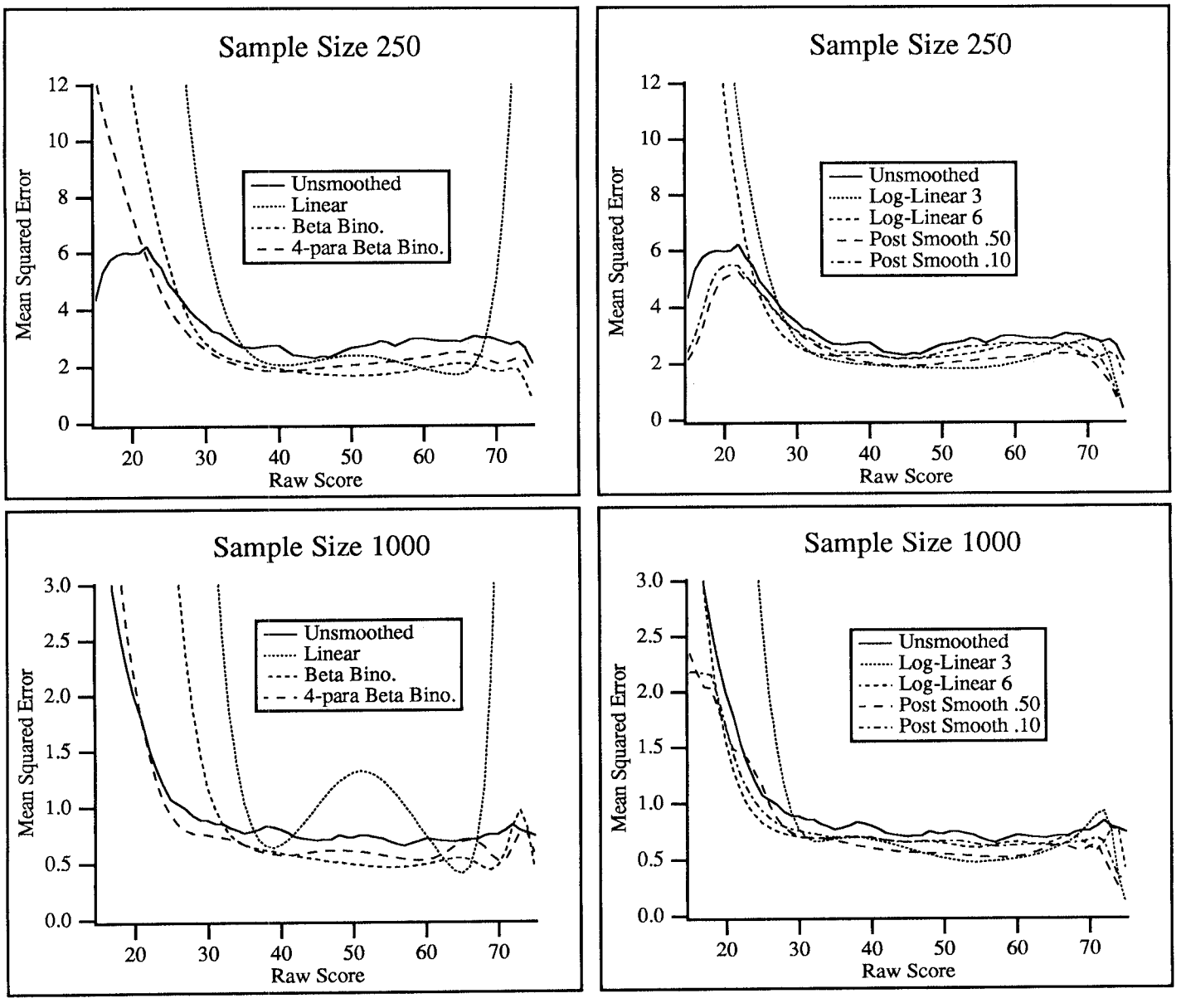

Figure 13. Mean squared error of equating methods for ACT English test.

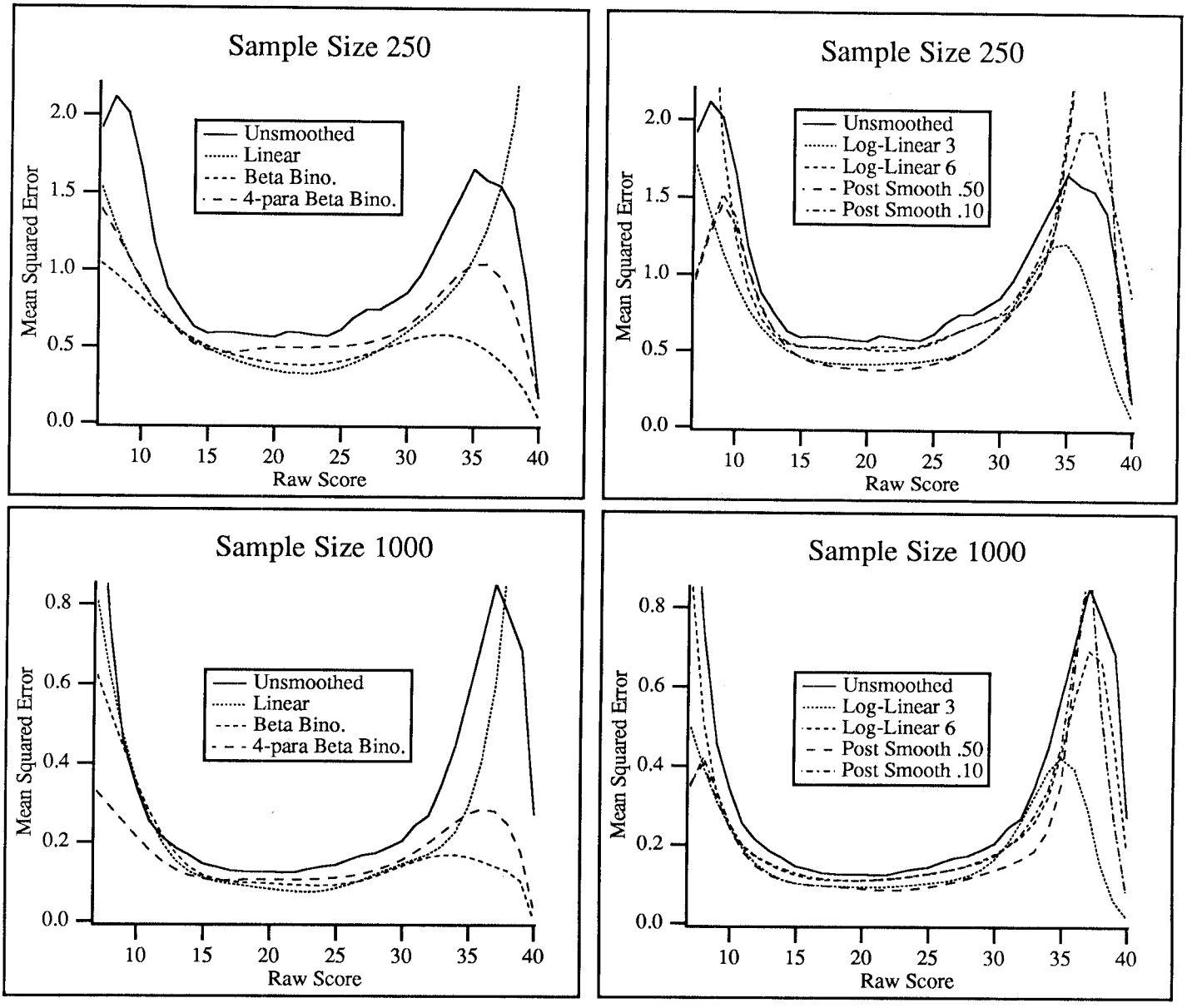

Figure 14. Mean squared error of equating methods for ACT Science Reasoning test.

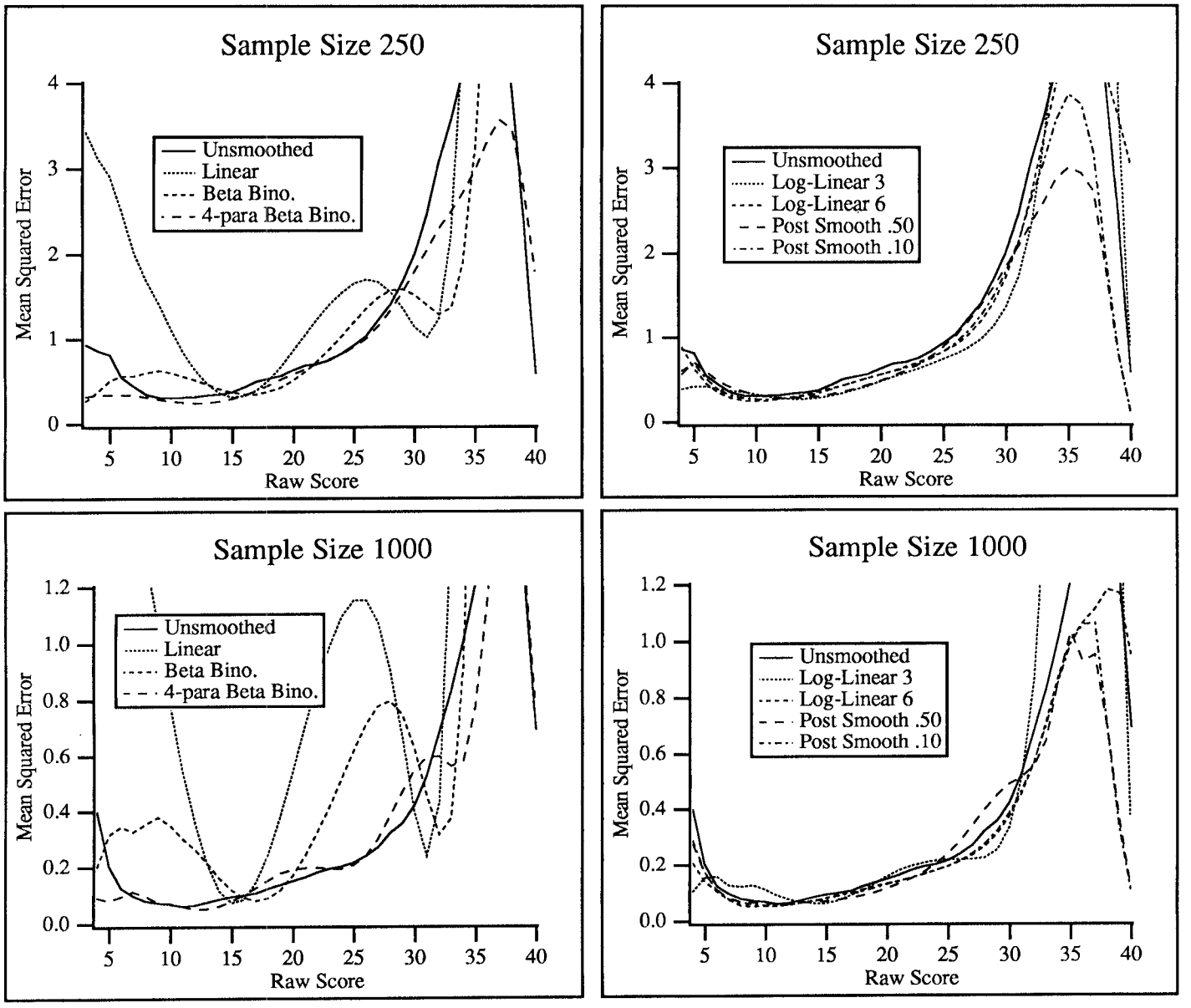

Figure 15. Mean squared error of equating methods for PLAN Mathematics test.