Research Report

Alternative Loglinear Smoothing Models and Their Effect on Equating Function Accuracy

Tim Moses

Paul Holland

December 2009

ETS RR-09-48

Alternative Loglinear Smoothing Models and Their Effect on Equating Function Accuracy

Tim Moses and Paul Holland
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

This simulation study evaluated the potential of alternative loglinear smoothing strategies for improving equipercentile equating function accuracy. These alternative strategies use cues from the sample data to make automatable and efficient improvements to model fit, either through the use of indicator functions for fitting large residuals or by averaging raw and smoothed frequencies. The strategies were studied across equating conditions based on rights-scored and formula-scored test data. Sample sizes were also manipulated. The results showed that the considered strategies produced equating functions with improved on-average accuracy but with added random variability. Of the considered alternative strategies, the frequency averaging strategy produced the most accurate equating functions for most of the evaluations done in the study. The frequency averaging strategy is recommended for circumstances where the desired loglinear model appears to fit the data poorly and where time constraints and/or data conditions make traditional modeling approaches unrealistic.

Key words: Loglinear smoothing, equipercentile equating, residuals

Acknowledgments

The authors thank Dan Eignor, Skip Livingston, and Gautam Puhan for helpful suggestions and Kim Fryer for the editorial work.

Table of Contents

List of Tables

Page

List of Figures

Page

Introduction

Since the introduction of polynomial loglinear models as a smoothing method for discrete test score distributions and equipercentile equating methods (Holland & Thayer, 1987), researchers have wondered how models' parameterizations affect equating function accuracy (Hanson, 1991; Hanson, Zeng, & Colton, 1994; Livingston, 1993; Skaggs, 2004). Research has primarily focused on varying the number of overall features of test score distributions (i.e., moments) that are fit by the loglinear models, showing that the number of moments has a direct relationship with equating variability and an inverse relationship with equating bias. The purpose of this study is to evaluate some alternative loglinear modeling strategies that may improve equipercentile equating accuracy beyond the accuracy improvements offered by the usual focus on moment selection.

Univariate Loglinear Smoothing Models

The loglinear models considered in this study are those used to produce smoothed versions of the frequency distribution for one test, X, with possible scores $x_j = x_1,\ldots,x_J$. The transposed row vector of observed score frequencies, $\mathbf{n} = (n_1,\ldots,n_J)^t$, sums to the total sample size, N. The loglinear model expresses the log of the expected (not observed) score probabilities in terms of a polynomial function of the test scores,

$$\log_e(p_j) = \beta_0 + \sum_{d=1}^{D} \beta_d x_j^d , \qquad (1)$$

where the x_j^d are functions of the possible score values of test X (e.g., $x_j^1, x_j^2,\ldots, x_j^D$), and β_0 is a normalizing constant that forces the sum of the expected probabilities, p_j, to equal 1 and the sum of the smoothed frequencies, m_j, to equal N. The β_d are parameters to be estimated in the model-fitting process. When (1) is fit using maximum likelihood estimation, the value of D determines the number of moments of the observed test score distribution that are preserved in the smoothed distribution.

Test distributions can exhibit complexities that can make models like (1) inadequate even when the models are fit with statistically optimal D values. The distribution of a formula-scored test, where a portion of examinees' incorrect responses is subtracted from their total number of correct responses, typically has abnormally low frequencies occurring at a small number of scores

that are separated by fixed intervals (i.e., teeth).[1] Heterogeneous examinee groups may produce score distributions with lumps, or bimodality, in formula-scored and rights-scored tests (i.e., tests scored as the total number of correct responses). The population models of these data are typically regarded as more complex than models like (1) (Holland & Thayer, 2000; von Davier, Holland, & Thayer, 2004).

Finding population models in samples of complex populations can be difficult and perhaps even unlikely. Research demonstrations suggest that complex population models can and should be found using extensive evaluations of several models' fit to sample data, based on models' residuals, conditional moments, observed versus smoothed plots, and a large variety of overall model fit statistics (e.g., Holland & Thayer, 2000; von Davier et al., 2004). These suggestions can seem fanciful to equating practitioners faced with short score reporting timelines, noisy distributions from small sample sizes, and unanticipated changes to test score distributions and work timelines that arise when items are unexpectedly removed from tests. For example, the equating of formula-scored test data at ETS is done using oversimplified loglinear models like (1) rather than the loglinear models that are most likely to reflect the population. To find plausible population models in search processes that are responsive to the difficulties typically encountered in equating practice, alternative loglinear modeling strategies that use cues from the sample data to make automatable and efficient improvements to model fit may be useful. Four alternative strategies are described and evaluated in this study.

Alternative Loglinear Modeling Strategies

One class of alternative loglinear models considered in this study focuses on directly fitting subsets of a total test score distribution. When a subset (i.e., *S*) of the total score range is identified as having frequencies that do not follow the pattern of most of the score distribution, this subset can be incorporated into a loglinear model with an indicator function,

$$\log_e(p_j) = \beta_0 + \sum_{d=1}^{D} \beta_d x_j^d + \beta_{D+1} I_S(j), \qquad (2)$$

where $I_S(j)$ is an indicator function set equal to 1 when score x_j is in subset *S* and 0 otherwise. The $I_S(j)$ in (2) produces smoothed frequencies that match the first *D* moments of the observed distribution and, for the scores in subset *S*, smoothed frequencies that sum to the total observed

frequencies. The literature's descriptions of indicator functions and loglinear models are limited to anticipated sources of lack of fit, such as distribution structures determined by a test scoring method, by rounding negative scores to zero, or by examinee subgroups (Hanson, 1996; Holland & Thayer, 2000; von Davier et al., 2004). These sources of lack of fit are addressed with extensive model searches and comparisons.

This study considers the use of indicator functions such as those in (2) as a data based and automatable approach to addressing unanticipated lack of fit. When an initial model such as (1) is fit, the lack of fit of this model could be addressed using indicator functions such as those in (2) to perfectly fit the score frequencies corresponding to initial model (1)'s largest residuals. Such a strategy would not require prior knowledge of model (1)'s inadequacies.

Several residuals for loglinear models could be used for directing the application of indicator functions to addressing unanticipated lack of fit. Three residuals considered in this study are the Freeman-Tukey residual, the standardized residual and the adjusted residual (Agresti, 2002; Cox, 1984; Freeman & Tukey, 1950; Haberman, 1973). The Freeman-Tukey residual is the square root of the *j*th part of the Freeman-Tukey chi-square statistic,

$$\sqrt{n_j} + \sqrt{n_j + 1} - \sqrt{4m_j + 1}\,. \tag{3}$$

The standardized residual is the square root of the *jth* part of the Pearson chi-square statistic,

$$\frac{n_j - m_j}{\sqrt{m_j}}\,. \tag{4}$$

Adjusted residuals adjust the standardized residuals in (4) to have asymptotic variances of 1, dividing each standardized residual by the square root of its estimated variance found in the following variance-covariance matrix,

$$\boldsymbol{I} - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}^t} - \boldsymbol{A}(\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t\,, \tag{5}$$

where $\boldsymbol{I}$ is a *J*-by-*J* identity matrix, $\boldsymbol{A} = N^{-1}\boldsymbol{D}_{\boldsymbol{p}}^{-1/2}(\boldsymbol{D}_{\boldsymbol{m}} - N^{-1}\boldsymbol{m}\boldsymbol{m}^{\boldsymbol{t}})\boldsymbol{X}$, the $\boldsymbol{D}_{\boldsymbol{m}}$ matrix is a diagonalized matrix of vector $\boldsymbol{m}$, the $\boldsymbol{D}_{\boldsymbol{p}}^{-1/2}$ matrix is a diagonalized matrix of vector of $\boldsymbol{p}^{-1/2}$, and $\boldsymbol{X}$ is the matrix containing the power functions of *X* shown in (1).

Another way to improve the fit of an initially chosen loglinear model like (1) may be to average the smoothed frequencies with the raw frequencies,

$$wn_j + (1-w)m_j = m_j + w(n_j - m_j), \quad 0 < w < 1. \tag{6}$$

Frequency averaging addresses a model's lack of fit by using a portion (*w*) of the residuals $(n_j - m_j)$ to rerough (Tukey, 1977) the smoothed frequencies (m_j) and, in this sense, utilizes residuals to improve model fit differently from the previously described indicator function approaches. The averaged frequencies from (6) preserve all of the observed moments that are preserved in the smoothed distribution. Frequency averaging may have the following benefits:

- the averaged distribution may not be as influenced as are indicator function strategies by sparse data and/or sampling variability because a model's lack of fit is addressed at an overall level that is more stable than score-specific levels,
- frequency averaging is a more easily implemented approach to addressing unanticipated lack of fit than the use of indicator functions for fitting large residuals, and
- because all of the m_j are greater than zero, all of the averaged frequencies are also greater than zero, avoiding the use of ad hoc rules needed to address difficulties the traditional equipercentile method has with scores where $n_j = 0$ (Kolen & Brennan, 2004).

Studying Loglinear Models and Equating Function Accuracy

To date, studies that evaluate loglinear models in equating have focused on the effects of fitting different numbers of distributions' overall moments on equipercentile equating accuracy (Hanson, 1991; Hanson et al., 1994; Livingston, 1993; Moses & Holland, 2008; Skaggs, 2004). While these studies are useful for describing the bias-variability tradeoff of loglinear models and equating function accuracy, they have not addressed situations where test distributions have

structures due to scoring practices and/or heterogeneous examinee groups. In addition, the research contexts where test distributions' complex structures have been modeled emphasized model searches that may be more time-intensive than equating timelines can realistically support. This study considers how loglinear models and equating functions can be improved by the use of alternative strategies that address the misfit of models like (1) through targeting large residuals with indicator functions or averaging the observed and smoothed distributions.

Method

A simulation study was designed to assess the effects of alternative loglinear modeling strategies on equating function accuracy. Population test score distributions and equating functions were defined by fitting loglinear models to test score data obtained in large-volume exam administrations. Several hundred datasets were drawn from the population distributions, the loglinear modeling strategies of interest were fit to the sample data, equipercentile equating functions were computed from the loglinear models' distributions, and the sample equating functions were compared to the population equating functions.

Loglinear Modeling Strategies

Six loglinear modeling strategies were assessed, including the four previously described alternative modeling strategies (averaging, Freeman-Tukey residuals, standardized residuals, and adjusted residuals) and two additional strategies used for comparative purposes (no alternative and population).

No alternative. This is the no alternative (no alt) model shown in (1), where D is selected from the sample data. The no alt strategy is most germane to equating practice, making it a useful baseline model for evaluating the potential improvements of the four alternative modeling strategies.

Residuals. Three strategies applied the indicator functions model in (2) to address large residuals in the no alt model. One strategy was based on Freeman-Tukey residuals (FT resid), one strategy was based on standardized residuals (std resid), and one strategy was based on adjusted residuals (adj resid). For these strategies, each large residual in the no alt model was fit with a single indicator function, $I_S(j)$. Because all three residuals tend to be interpreted as standardized deviates, residuals were defined as large enough to warrant indicator functions when they exceeded 2 in absolute value. For example, if the no alt model had three FT resids whose absolute

values were greater than 2 at scores j = 2, 5 and 9, then the FT resid strategy would use the following loglinear smoothing model,

$$\log_e(p_j) = \beta_0 + \sum_{d=1}^{D} \beta_d x_j^d + \beta_{D+1} I_S(j=2) + \beta_{D+2} I_S(j=5) + \beta_{D+3} I_S(j=9). \quad (7)$$

Averaging. For the averaging strategy, each observed frequency was averaged with the smoothed frequency in the no alt model by setting w equal to 0.5 in (6).

Population. The population strategy fit the population loglinear models to all of the sample datasets created for the replications of the study. The population models are described in the Population Distributions and Equating Functions section in this report. This strategy is the ideal scenario for using loglinear models in equating.

D Selection for the No Alt Model

To allow for the possibility that the alternative loglinear modeling strategies' performance might reflect how D was chosen, the no alt strategy and the four alternative strategies (averaging, Freeman-Tukey residuals [FT resid], standardized residuals [std resid], and adjusted residuals [adj resid]) were evaluated based on two methods for picking the no alt model's D from values ranging from 2 through 10. The two considered methods for picking D are the minimization of the Akaike information criterion (AIC) statistic (Akaike, 1981) and of the consistent AIC (CAIC) statistic (Bozdogan, 1987). Prior research has found that the minimization of the AIC statistic typically selects loglinear models with larger D values than the minimization of the CAIC statistic (Moses & Holland, 2008). Considering the performance of the alternative loglinear modeling strategies in terms of AIC and CAIC minimization allows for the possibility that differences in the alternative strategies may be larger when based on no alt loglinear models that fit the data less closely (i.e., loglinear models based on the CAIC's smaller D) than on initial loglinear models that fit the data more closely (i.e., loglinear models based on the AIC's larger D).

Population Distributions and Equating Functions

Two equivalent groups equating conditions from Moses and Holland (2008) were used in this study. One of the conditions was based on rights-scored tests, where most of the test distributions were obtained from large-volume teacher certification exams. The rights-scored X and Y relative frequency distributions used as populations in this study were obtained by fitting loglinear model (1)

with $D = 6$ to an observed X distribution and another loglinear model (1) with $D = 2$ to a different Y distribution. These modeled distributions are plotted in Figures 1 and 2 and the overall statistics are shown in Table 1. Another equating condition was obtained from large-volume aptitude exams where the tests were formula scored and rounded, with negative scores being rounded to score zero. The formula-scored X and Y frequency distributions used as populations in this study were obtained by fitting loglinear models that preserved the teeth distribution[1], the lump at zero, and four overall moments in observed X and Y data. These modeled distributions are plotted in Figures 3 and 4 and the overall statistics are shown in Table 1.

The population rights- and formula-scored X-to-Y equating functions were computed from the population relative frequency distributions in Figures 1–4. Both equating functions were complex equipercentile functions. The formula-scored equating condition presumably warranted the alternative loglinear modeling strategies due to structures in its X and Y distributions that were beyond what the no alt model would directly address. The rights-scored equating condition was of interest because it provided the opportunity to assess the performances of the alternative loglinear modeling strategies in a situation where the strategies were not actually needed and where the no alt model was expected to produce adequate results.

Table 1

Four Univariate Population Distributions and Two Population X-to-Y Equating Functions

	Rights-scored condition		Formula-scored condition	
	X	*Y*	*X*	*Y*
Score range	0–40	0–40	0–78	0–78
Mean	28.09	20.00	39.25	32.69
Standard deviation	7.44	6.88	17.23	16.73
Skew	-0.41	0.00	-0.11	0.24
Kurtosis	-0.63	-0.19	-0.77	-0.69

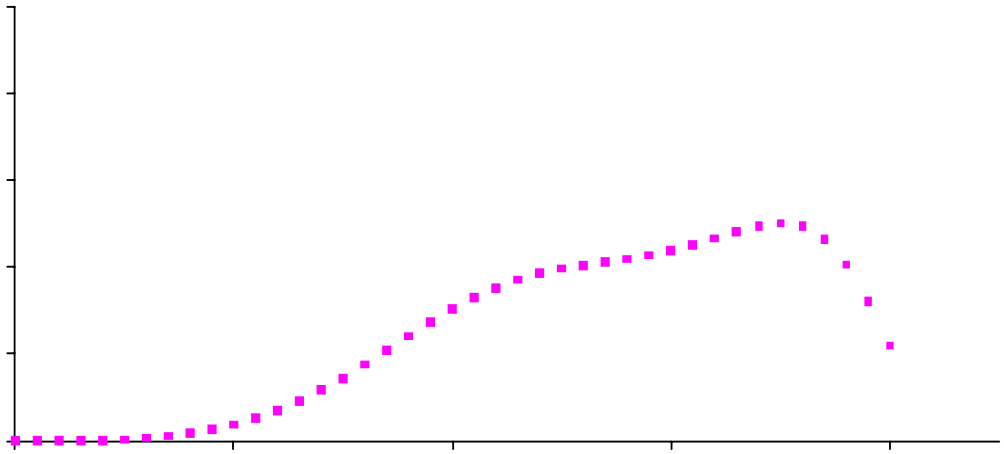

Figure 1. **Population *X* distribution for the rights-scored *X*-to-*Y* equating condition.**

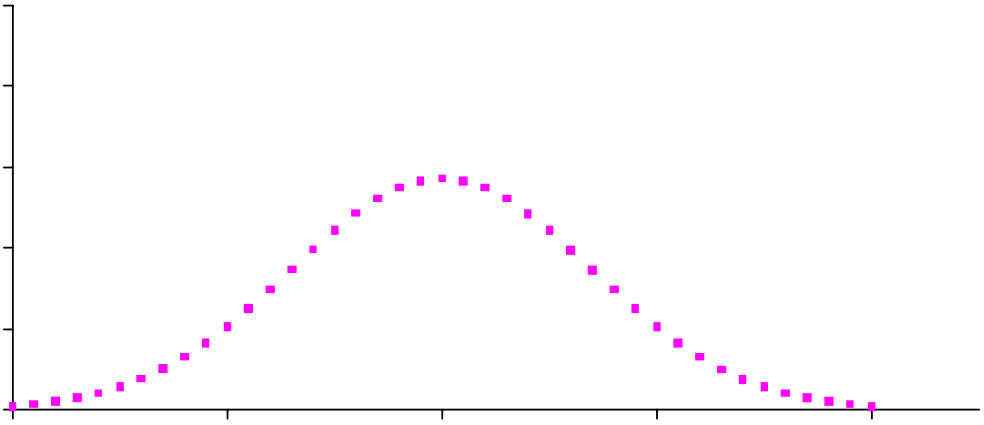

Figure 2. **Population *Y* distribution for the rights-scored *X*-to-*Y* equating condition.**

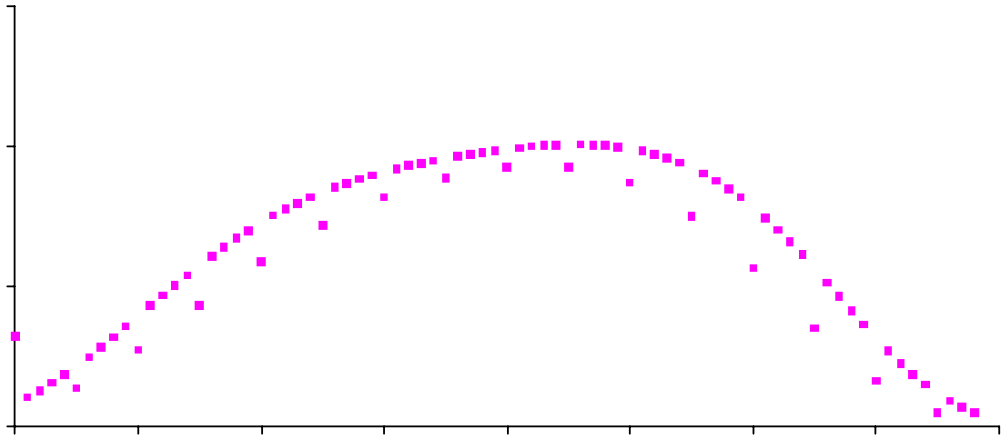

Figure 3. **Population *X* distribution for the formula-scored *X*-to-*Y* equating condition.**

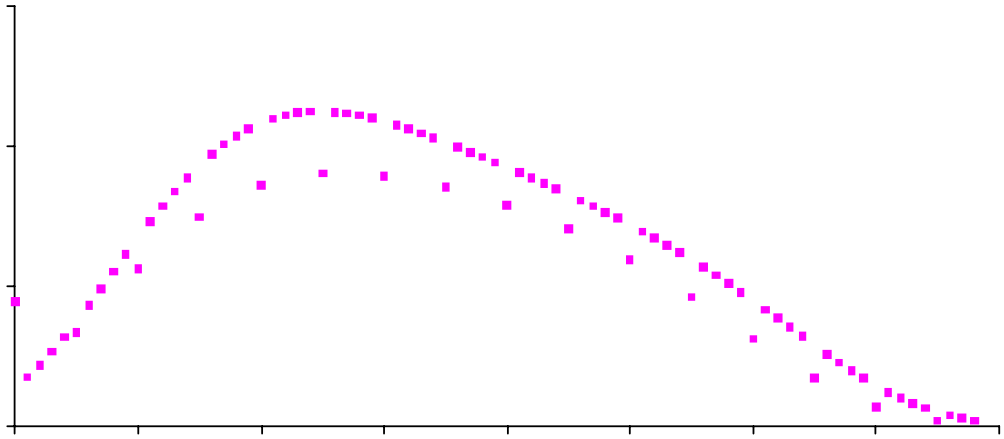

Figure 4. **Population Y distribution for the formula-scored X-to-Y equating condition.**

Sample Sizes and Random Datasets

Datasets were randomly drawn from each of the four population test score distributions, some with 1,000 examinees and others with 5,000 examinees. These datasets resembled the statistical characteristics and score ranges of their respective population distributions but also had random noise.

Evaluation

The four alternative loglinear modeling strategies (averaging, FT resid, std resid, and adj resid) were evaluated for two methods of selecting *D* in the no alt model, two sample sizes, and two population equating function conditions. For each of the *D*-by-sample size-by-equating function conditions, 200 sample datasets were drawn, the strategies were fit to the sample data, equipercentile *X*-to-*Y* functions (Kolen & Brennan, 2004) were computed from the loglinear models, and the 200 sample equating functions were compared to the population equating function.

Accuracy Measures

Accuracy measures for the *X*-to-*Y* equating functions ($e_y(x)$) based on different alternative loglinear model strategies were developed based on some of the measures used in prior equating studies (Hanson, 1991; Livingston, 1993; Moses & Holland, 2008). The measures are based on score-level mean squared error (MSE_j),

$$\begin{aligned} MSE_j &= \frac{1}{200}\sum_i \left(\hat{e}_{y,i}(x_j) - e_{y,Population}(x_j)\right)^2 \\ &= \frac{1}{200}\sum_i \left[\left(\bar{e}_y(x_j) - e_{y,Population}(x_j)\right)^2 + \left(\hat{e}_{y,i}(x_j) - \bar{e}_y(x_j)\right)^2\right] \\ &= Bias_j^2 + Variance_j \end{aligned} \tag{8}$$

where *j* indicates the *j*th score in the observed score range, *j* = 1 to *J*, *i* indicates one of the 200 random datasets drawn from one of the population distributions at one of the sample sizes, $\hat{e}_{y,i}(x_j)$ is the estimated *X*-to-*Y* equated score at the *j*th score in one of the 200 datasets, $\bar{e}_y(x_j)$ is the average of the 200 sample datasets' equated scores at the *j*th score, and $e_{y,Population}(x_j)$ is the population equated score at the *j*th score.

The measures of interest were averages of the root-squared bias and root-variance of the MSE_j. Square roots of these quantities were taken to obtain measures in the scale of the equated scores being estimated. The root-squared bias is referred to as a mean absolute deviation (MAD), averaged with respect to *X*'s population distribution, $P(X_{Population} = x_j)$,

$$\text{MAD} \quad \begin{aligned} &= \sum_j \sqrt{Bias_j^2}\, P(X_{Population} = x_j) \\ &= \sum_j \left|Bias_j\right| P(X_{Population} = x_j) \end{aligned} . \qquad (9)$$

The root-variance is referred to as a mean standard error of equating (MSEE), averaged with respect to *X*'s population distribution, $P(X_{Population} = x_j)$,

$$\text{MSEE} \quad \begin{aligned} &= \sum_j \sqrt{Variance_j}\, P(X_{Population} = x_j) \\ &= \sum_j SEE_j P(X_{Population} = x_j) \end{aligned} . \qquad (10)$$

To supplement the results of the overall MAD and MSEE values, score-level biases ($Bias_j$) and standard errors of equating (SEE_j) were also evaluated at specific combinations of *D* selections, sample size, and equating condition.

Results

Rights-Scored Equating Results

Tables 2 and 3 present the strategies' MAD and MSEE results for the rights-scored equating condition across the two considered sample sizes and *D* selection methods. The MAD results in Table 2 show that the average equating function produced by the no alt strategy is less accurate (i.e., larger MAD values) than the average equating functions of the four alternative strategies, which are in turn less accurate than the average equating function from using the population models. Of the four alternative strategies, the averaging strategy has the smallest MAD values. All strategies are more accurate when based on *D* values selected by minimizing the AIC statistic rather than the CAIC statistic and when based on sample sizes of 5,000 rather than sample sizes of 1,000.

Table 2

Mean Absolute Deviation (MAD) Values for the Rights-Scored Equating Situation

N	*D*	No alt	FT resid	Std resid	Adj resid	Averaging	Population
1,000	CAIC	0.137	0.103	0.091	0.080	0.075	0.029
1,000	AIC	0.035	0.037	0.032	0.032	0.030	
5,000	CAIC	0.044	0.033	0.033	0.029	0.024	0.006
5,000	AIC	0.007	0.007	0.007	0.007	0.007	

Note. Adj resid = adjusted residuals, AIC = Akaike information criterion, CAIC = consistent AIC, FT resid = Freeman-Tukey residuals, no alt = no alternative, std resid = standardized residuals.

Table 3

Mean Standard Error of Equating MSEE) Values for the Rights-Scored Equating Situation

N	*D*	No alt	FT resid	Std resid	Adj resid	Averaging	Population
1,000	CAIC	0.415	0.428	0.427	0.429	0.409	0.381
1,000	AIC	0.407	0.420	0.418	0.420	0.416	
5,000	CAIC	0.196	0.198	0.199	0.198	0.197	0.185
5,000	AIC	0.194	0.199	0.199	0.200	0.199	

Note. Adj resid = adjusted residuals, AIC = Akaike information criterion, CAIC = consistent AIC, FT resid = Freeman-Tukey residuals, no alt = no alternative, std resid = standardized residuals.

The MSEE results in Table 3 show that the least variable equating results were produced from using the population strategy. The no alt strategy was the second least variable strategy, except for when *D* values were selected with the CAIC statistic in sample sizes of 1,000. Of the four alternative strategies, the averaging strategy was the least variable. All strategies become less variable when based on sample sizes of 5,000 rather than sample sizes of 1,000.

Formula-Scored Equating Results

Tables 4 and 5 present strategies' MAD and MSEE results for the formula-scored equating condition across the two considered sample sizes and *D* selection methods. The MAD results in Table 4 show that the average equating function produced by the no alt strategy is less accurate (i.e., larger MAD values) than the average equating functions of the four alternative strategies. Results are inconsistent for identifying the best performing strategy in terms of MAD values, sometimes showing that the use of the population models is the most accurate strategy, other times showing that averaging is the most accurate strategy, and other times showing that adj resid is the most accurate strategy. Like the rights-scored results in Table 2, Table 4's results show that all strategies' MAD values are smallest when based on sample sizes of 5,000 rather than sample sizes of 1,000.

Table 5's MSEE results show that the least variable results were obtained from the population strategy. The second least variable strategy is sometimes the no alt strategy and other times the averaging strategy. All strategies become less variable when based on sample sizes of 5,000 rather than sample sizes of 1,000.

Score-Level Results

Figures 5–8 plot the score-level biases and SEEs for some of the rights- and formula-scored equating conditions. The plotted results are of the study conditions where the strategies are most easily differentiated, based on *D* selections based on the CAIC statistic and on sample sizes of 1,000. The score-level results in Figures 5–8 are generally consistent with the overall MAD and MSEE results in Tables 2–5, showing that for the majority of scores, the use of population models is the least biased and least variable strategy, the averaging strategy is the least biased and least variable of the four alternative loglinear modeling strategies, and the no alt strategy is the most biased strategy. As suggested in Tables 2–5, the score-level results for the nonplotted conditions are much smaller and, hence, would be less visible than what is shown in Figures 5–8.

Table 4

Mean Absolute Deviation (MAD) Values for the Formula-Scored Equating Situation

N	*D*	No alt	FT resid	Std resid	Adj resid	Averaging	Population
1,000	CAIC	0.321	0.228	0.216	0.207	0.158	0.061
1,000	AIC	0.071	0.061	0.064	0.063	0.057	
5,000	CAIC	0.085	0.042	0.043	0.042	0.054	0.051
5,000	AIC	0.074	0.047	0.049	0.041	0.047	

Note. Adj resid = adjusted residuals, AIC = Akaike information criterion, CAIC = consistent AIC, FT resid = Freeman-Tukey residuals, no alt = no alternative, std resid = standardized residuals.

Table 5

Mean Standard Error of Equating (MSEE) Values for the Formula-Scored Equating Situation

N	*D*	No alt	FT resid.	Std resid.	Adj resid.	Averaging	Population
1,000	CAIC	1.129	1.121	1.137	1.133	1.061	0.982
1,000	AIC	1.039	1.060	1.059	1.059	1.049	
5,000	CAIC	0.481	0.484	0.485	0.486	0.477	0.449
5,000	AIC	0.481	0.492	0.492	0.491	0.484	

Note. Adj resid = adjusted residuals, AIC = Akaike information criterion, CAIC = consistent AIC, FT resid = Freeman-Tukey residuals, no alt = no alternative, std resid = standardized residuals.

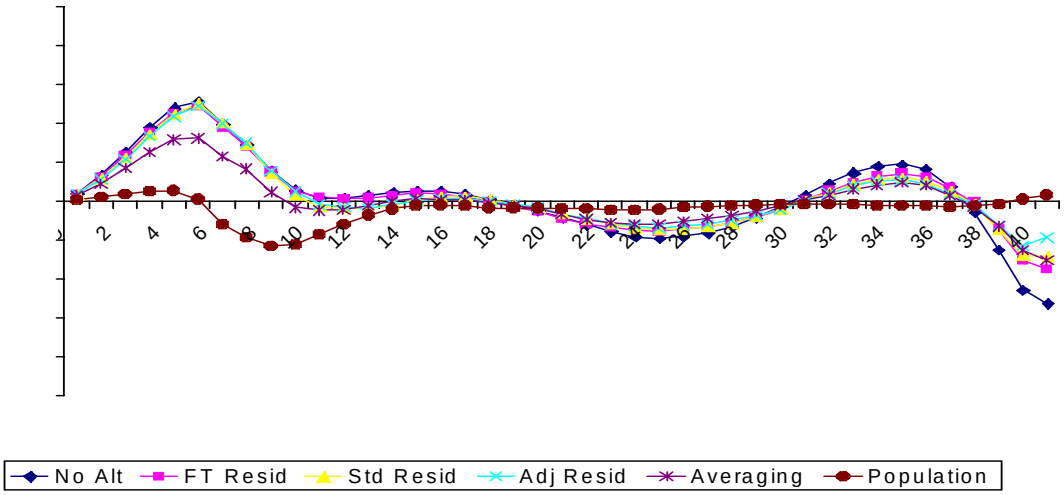

Figure 5. **Selection strategies' biases for the rights-scored equating condition, the consistent Akaike information criterion (CAIC) *D* selection, and sample sizes of 1,000.**

Note. Adj resid = adjusted residuals, FT resid = Freeman-Tukey residuals, no alt = no alternative, std resid = standardized residuals.

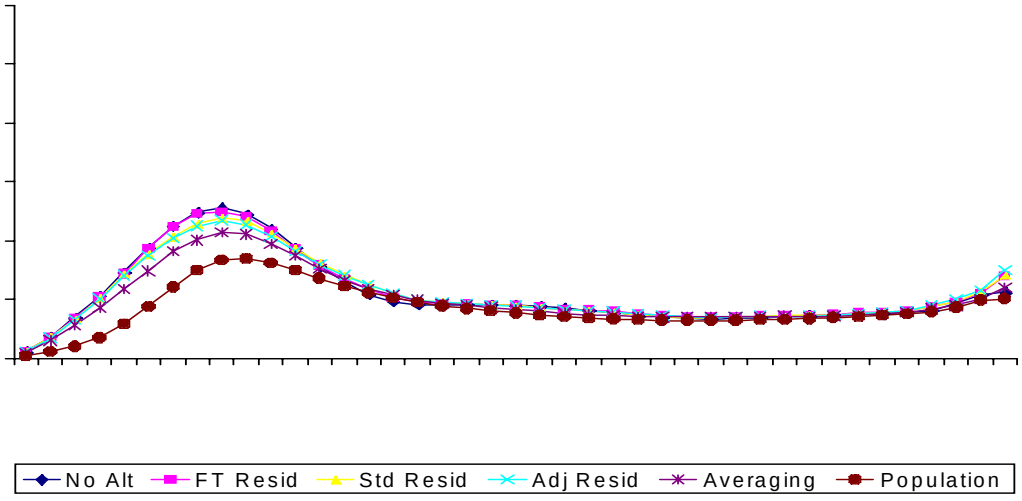

Figure 6. **Selection strategies' standard errors of equating (SEEs) for the rights-scored equating condition, the consistent Akaike information criterion (CAIC) *D* selection, and sample sizes of 1,000.**

Note. Adj resid = adjusted residuals, FT resid = Freeman-Tukey residuals, no alt = no alternative, std resid = standardized residuals.

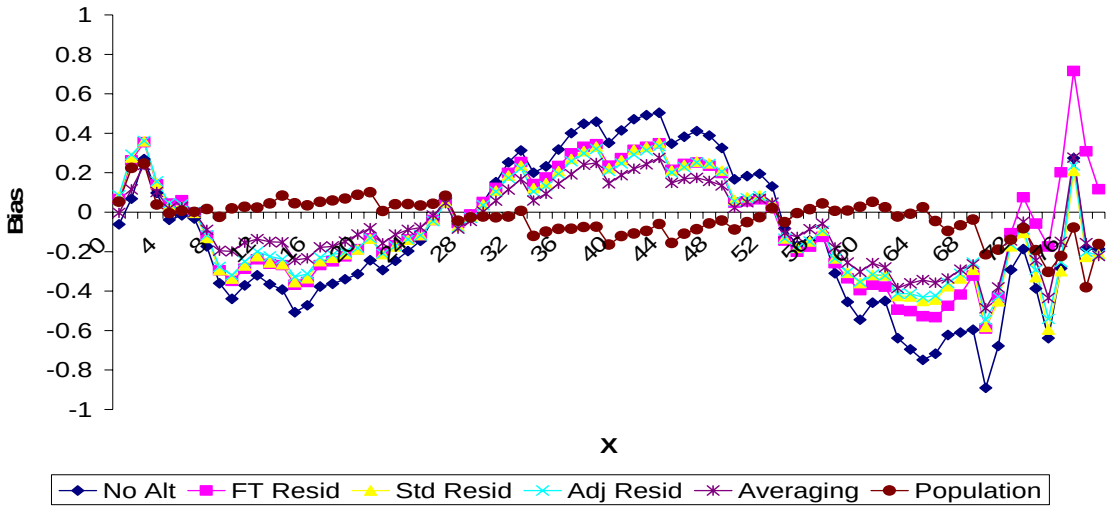

Figure 7. **Selection strategies' biases for the formula-scored equating condition, the CAIC D selection and sample sizes of 1,000.**

Note. Adj. resid. = adjusted residuals, FT resid.= Freeman-Tukey residuals, no alt = no alternative, std resid. = standardized residuals.

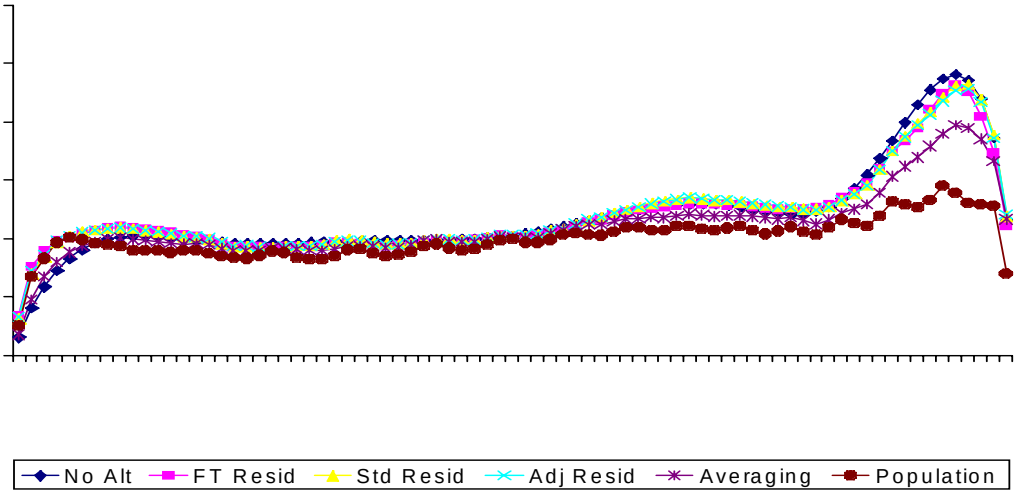

Figure 8. **Selection strategies' standard errors of equating (SEEs) for the formula-scored equating condition, the CAIC *D* selection, and sample sizes of 1,000.**

Note. Adj. resid. = adjusted residuals, FT resid.= Freeman-Tukey residuals, no alt = no alternative, std resid. = standardized residuals.

Discussion

The purpose of this study was to evaluate some data-driven and automatable loglinear smoothing strategies that are alternatives to the well-known strategies that predominantly focus of fitting distributions' moments. These alternative strategies use the misfit of a loglinear smoothing model as a basis for automatically improving the model. How well these alternative strategies appear to perform depends on what strategies they are compared to. One baseline model used in this study involved fitting the population models to all of the sample datasets in the simulation (population). Another baseline model used a statistical strategy to select the overall number of moments to fit in the test score distributions without an alternative strategy (no alt).

One evaluation of the alternative strategies was with respect to the ideal scenario where the population model is always known regardless of the sample data (population). The alternative strategies were always worse than the population strategy in terms of equating variability and were usually worse than the population strategy in terms of average absolute deviations from the population equating function. One apparent implication of comparisons with the population strategy is that population models should always be used when applying smoothing methods in research and practice. This implication is optimistic, as population models are always unknown in unsimulated data, at best are only closely approximated with extensive searches in sample data, and at worst are disregarded due to the constraints of time, tasks, and priorities that comprise equating practice. Another implication of the population results is that model search strategies that are structured to be consistent with the population model should perform well. For example, if separate model searches for *D* in the teeth and nonteeth distributions of formula-scored data eventually become feasible for equating practice and these separate searches were implemented using AIC minimization in sample sizes that were not too small, these separate searches would likely produce accurate equating functions.

When compared to the no alt strategy, most relevant to current equating practice, the alternative modeling strategies produced equating functions that more closely approximated the population equating function on average (Tables 2 and 4) but usually introduced more random variability (Tables 3 and 5). When the no alt strategy was based on a poor smoothing model, the improvements in accuracies of the alternative strategies' equating functions exceeded the increase in equating variability. When the no alt strategy was based on a well-chosen smoothing model, the improvements in the accuracies of the alternative strategies' equating functions were small relative

to the increase in equating variability. These results suggest that the alternative strategies are most useful in circumstances where the desired loglinear model appears to fit the data poorly and where time constraints and/or data conditions make traditional modeling approaches unrealistic.

Of the four alternative modeling strategies, the averaging strategy usually outperformed the other three strategies in terms of the accuracy of its average equating functions and its variability. The performance and simplicity of the averaging strategy suggest that it can be useful in improving equating practice at ETS, particularly when formula-scored test data are being equated and smoothing models known to be incorrect are used automatically and without review. Useful extensions of the averaging strategy would be to derive estimates of the variance-covariance matrices of the averaged smoothed and observed distributions and also to consider different approaches for choosing *w*.

The results suggest that the use of indicator functions to fit large residuals can be recommended when smoothing and equating formula-scored tests with sample sizes of at least 5,000. The residuals were most likely to reflect structures actually in the population distributions for the formula-scored tests rather than for the rights-scored tests. The residuals were most likely to be accurate indications of actual structures when based on sample sizes of 5,000 rather than 1,000. In comparing these residuals-based indicator function strategies to the averaging strategy, the important issue is how much of a large residual gets incorporated into the smoothing model being improved upon. With smaller sample sizes, the averaging strategy's incorporation of a portion of all residuals was more sensible than the residuals-based strategies' incorporation of all of the large residuals. With large sample sizes and formula-scored distributions, the residuals-based strategies' incorporation of all of the large residuals produced average equating functions that were more accurate than the averaging strategy, though the individual equating functions were also more variable. Of the three residuals-based strategies, the strategy which used adjusted residuals performed better than the strategies that used Freeman-Tukey and standardized residuals.

Throughout this study, the accuracy criteria used to evaluate the alternative strategies was with respect to a known population equating function. While population-based evaluations are commonly used in smoothing and equating studies (Hanson, 1991; Hanson et al., 1994; Livingston, 1993), they represent a limited number of possible criteria that could be used to evaluate equating (e.g., Harris & Crouse, 1993). Some smoothing and equating methods not addressed in this study have been developed to deliberately incorporate smoothness criteria in their

results (e.g., Kolen & Brennan, 2004, p. 86; von Davier et al., 2004, pp. 62–63). To the extent that smooth equating functions are valued, measures of an equating function's smoothness may be important supplements to the accuracy measures considered in this study. An important pursuit of future research and practice would be to clarify the relative importance of equating accuracy and equating smoothness, particularly in situations where test data have systematic irregularities that make accuracy and smoothness criteria inconsistent.

References

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

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics, 16,* 3–14.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–370.

Cox, C. (1984). An elementary introduction to maximum likelihood estimation for multinomial models: Birch's theorem and the delta method. *The American Statistician, 38*(4), 283–287.

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics, 21,* 607–611.

Haberman, S. J. (1973). The analysis of residuals in cross-classification tables. *Biometrics, 29,* 205–220.

Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement, 15*(4), 391–408.

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

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing methods in equipercentile equating* (ACT Technical Rep. No. 94-4). Iowa City, IA: ACT.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*(3), 195–240.

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.

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

Livingston, S. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30,* 23–39.

Moses, T., & Holland, P. W. (2008). *The influence of strategies for selecting loglinear smoothing models on equating functions* (ETS Research Rep. No. RR-08-25). Princeton, NJ: ETS.

Skaggs, G. (2004). Passing score stability when equating with very small samples. Presentation to the American Educational Research Association, San Diego, CA.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

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

Notes

[1.] Teeth in rounded formula-scored distributions arise because examinees vary in terms of their item omission patterns, and groups of examinees with different item omission patterns vary in their size. In rounded formula-scored distributions, item omission patterns define sets of total scores that are impossible to obtain. The teeth make up a set of impossible scores for the relatively large group of examinees that does not omit any items.