Research Report

The Influence of Strategies for Selecting Loglinear Smoothing Models on Equating Functions

Tim Moses

Paul Holland

May 2008

ETS RR-08-25

The Influence of Strategies for Selecting Loglinear Smoothing Models on Equating Functions

Tim Moses and Paul Holland

ETS, Princeton, NJ

May 2008

Abstract

This study addressed 2 issues of using loglinear models for smoothing univariate test score distributions and for enhancing the stability of equipercentile equating functions. One issue was a comparative assessment of several statistical strategies that have been proposed for selecting 1 from several competing model parameterizations. Another issue was an evaluation of the influence of the selection strategies on equating function accuracy. These issues were considered in a simulation study, where the accuracies of 17 selection strategies for loglinear models and their effects on equating function accuracies were assessed across a range of sample sizes, test score distributions, and population equating functions. The results differentiate the selection strategies in terms of their accuracies in selecting correct model parameterizations and define the situations where their use has the most important implications for equating function accuracy.

Key words: Selection strategies, loglinear smoothing, equipercentile equating

Table of Contents

List of Tables

List of Figures

Page

Introduction

The loglinear modeling used to smooth test score distributions (Holland & Thayer, 1987; Kolen, 1991; Livingston, 1993; Skaggs, 2004) is a psychometric procedure that is both flexible and complex. There are many possible parameterizations for loglinear models, ranging from simple models with few parameters to complex models with many parameters. There are also many ways to select models' parameterizations, including extensive analyses and comparative evaluations (e.g., Holland & Thayer, 2000; von Davier, Holland, & Thayer, 2004), and programmable selection strategies that are data driven (Agresti, 2002; Bishop, Feinburg, & Holland, 1975; Haberman, 1974a) and even sample-size driven. Although research has evaluated the use of smoothing in equating, not much is known about how selection strategies for loglinear models affect test equating results. The purpose of this study was to compare several selection strategies for loglinear models in terms of their accuracies and to evaluate the influence of selection strategies on test equating accuracy.

Univariate Loglinear Smoothing Models

The loglinear models considered in this study are those used to produce smooth versions of the frequency distribution for one test, *X*, with possible scores $x_1,\ldots,x_J$, or x_j, with $j = 1,\ldots,J$. The transposed row vector of observed score frequencies, $\boldsymbol{n} = (n_1,\ldots,n_J)^t$, sums to the total sample size, *N*. The loglinear model expresses the log of the expected (not actual) score probabilities in terms of a polynomial function of the test scores,

$$\log_e(p_j) = \beta_0 + \sum_{i=1}^{I} \beta_i x_j^i \,, \qquad (1)$$

where the x_j^i are score functions of the possible score values of test *X* (e.g., $x_j^1, x_j^2, x_j^3, \ldots, x_j^I$), β_0 is a normalizing constant that forces the sum of the expected probabilities (p_j) to equal 1, and the β_i are parameters to be estimated in the model-fitting process. The value of *I* determines the extent of smoothing and, when maximum likelihood estimation is used, the number of moments of the actual test score distribution that are preserved in the smoothed distribution. If *I* = 1, then the smoothed distribution preserves only the first moment (the mean) of the observed distribution. If *I* = 4, then the smoothed distribution preserves the first, second, third, and fourth moments (mean, variance, skewness, and kurtosis) of the observed distribution. The value of *I* also determines the extent to which the smoothed frequencies, $m_j = Np_j$, approximate the observed frequencies, n_j.

Model Selection Strategies

Selection strategies for selecting loglinear models like Equation 1 with different values of *I* can be categorized into distinct classes. The major classes considered in this paper are strategies based on significance tests of overall model fit statistics, on model fit relative to model parameterization (i.e., parsimonious fit), and on sample size.

Significance testing class. Statistical significance tests based on the extent to which a model's smoothed frequencies fit the observed frequencies in the total distribution can be useful in comparing and selecting loglinear smoothing models. Several asymptotically equivalent chi-square goodness-of-fit statistics have been developed, based on the assumption that the frequency data being modeled follow either a Poisson or a multinomial distribution (Bishop et al., 1975; Fisher, 1922; Haberman, 1974a; Read & Cressie, 1988). Four chi-square statistics are considered in the current study, including the likelihood ratio chi-square,

$$G^2 = 2\sum_j n_j \log_e\left(\frac{n_j}{m_j}\right), \qquad (2)$$

the Pearson chi-square,

$$\chi_P^2 = \sum_{\mathrm{j}} \frac{\left(n_j - m_j\right)^2}{m_j}, \qquad (3)$$

the Freeman-Tukey chi-square,

$$\chi_{FT}^2 = \sum_{\mathrm{j}}\left(\sqrt{n_j} + \sqrt{n_j + 1} - \sqrt{4m_j + 1}\right)^2, \qquad (4)$$

and the Cressie-Read chi-square,

$$CR = 1.8 \sum_{\mathrm{j}} n_j \left(\left(\frac{n_j}{m_j} \right)^{2/3} -1 \right). \qquad (5)$$

For a set of nested models that can be arranged in a sequence of simple to complex parameterizations (e.g., nine models from Equation 1, where models' *I* = 2, 3, 4,….,10), the significance testing of models' chi-square statistics could proceed in two directions. The testing could begin with the most complex models and make comparisons with simpler models (complex-to-simple strategy) or could begin with the simplest models and make comparisons with more complex models (simple-to-complex strategy).

The general complex-to-simple strategy was described in Haberman (1974b) and applied to test score distribution problems by Hanson (1990, 1996; Hanson & Feinstein, 1995). This strategy evaluates the improvement in fit of a complex model relative to a model that is one term simpler based on a chi-square significance test of the difference in the models' chi-square statistics and degrees of freedom. A nonsignificant chi-square test indicates that the term in the complex model and not in simpler model is fitting sampling noise (i.e., there is support for the null hypothesis that the simpler model's *I* is correct). A significant chi-square test indicates that the term in the complex model and not in the simpler model is in the population model so that all models with parameterizations less than the complex model are simultaneously rejected (i.e., there is support for the alternative hypothesis that the complex model's *I* is correct). While there is no theoretical basis for selecting among the remaining model parameterizations, Hanson recommended selecting the final model with the smallest of the remaining *I's*.

For choosing among possible *I* = 2 through 10 with a simple-to-complex strategy, the selection process begins by testing the improvement in model fit of the 3-parameter model relative to the 2-parameter model, based on the difference in the models' chi-square statistics and degrees of freedom. If the fit of the 3-parameter model is significantly better than that of the 2-parameter model, the 3-parameter model is selected. Then, the improvement in fit of the 4-parameter model relative to the three-parameter model is tested. If the fit of the 3-parameter model is not significantly better than that of the 2-parameter model, the 2-parameter model is selected. Then, the improvement in fit of the 4-parameter model relative to the 2-parameter model is tested. The improvements of fit for the 5- through 10-parameter models are similarly considered. If none of the models has significantly better fits than the 2-parameter model, then the 2-parameter model is selected. The simple-to-complex strategy considered here uses a Type I error level of $1-(1-\alpha)^{1/(\#Models-1)}$ for each significance test.

A third significance testing strategy considered in this study tests the overall fit of each individual model relative to the models' degrees of freedom. The selected model from this individual-models strategy is the simplest model (i.e., the model with the smallest *I*) that has an insignificant chi-square fit statistic. No adjustment is made for the overall Type I error level.

Parsimony class. The parsimony class of model selection strategies contains statistics that evaluate a model's G^2 with respect to the parameterization needed to achieve that G^2. Four parsimony class statistics considered in this study are the Akaike information criterion (AIC; Akaike, 1981),

$$AIC = G^2 + 2(I+1), \tag{6}$$

the Bayesian information criterion (BIC; Schwartz, 1978),

$$BIC = G^2 + \log_e(N)(I+1), \tag{7}$$

the consistent Akaike information criterion (CAIC; Bozdogan, 1987),

$$CAIC = G^2 + \left(1+\log_e(N)\right)(I+1), \tag{8}$$

and a statistic attributed to Goodman (Agresti, 2002),

$$Goodman = \left|\frac{G^2}{J-I-1} - 1\right|. \tag{9}$$

The strategy of model selection based on the parsimony class is to select the model from a set of competing models with the smallest statistic.

Sample size selection. Some statistical analysis teams at ETS base their model selections on sample size rather than on the fit of the observed and smoothed distributions (e.g., Table 1). An advantage of selecting models based on sample size relative to other strategies is increased efficiency for the smoothing and equating work that must be completed under increasingly tight time constraints.

Table 1

Sample Size Guidelines Commonly Used for Selecting the Parameterization of a Loglinear Model for a Score Distribution

Sample size	No. of moments to preserve (*I*)
Less than 40	2
40–199	3
200–299	4
300 or more	5

Studying Selection-Strategy Accuracy

One issue with the application of loglinear models and selection strategies to test equating is that in the situations where loglinear models are relied on to smooth test score distributions, sample sizes may be too small for the selection strategies to work (Fienburg, 1979; Haberman, 1988; Koehler & Larntz, 1980). Test distribution and equating studies have discouraged the use of selection strategies such as complex-to-simple likelihood ratio chi-square tests in favor of fixed-model strategies because the chi-square tests have displayed accuracy problems in small samples (Hanson, 1990, 1991). Another issue is that the accuracy of selection strategies may not be as closely related to equating function accuracy as equating practitioners might expect, because selection strategies evaluate model fit in terms of frequency distributions rather than in terms of the cumulative frequency distributions that are used by equipercentile equating methods. These issues are the basis of this study, in which the accuracies of the reviewed selection strategies were compared in terms of the test score distributions and sample sizes typically encountered in practice. The selection strategies also were evaluated with respect to equating function accuracy.

Method

This simulation study evaluated the accuracy of different model selection strategies on loglinear model parameterization accuracy and on equating function accuracy. Accuracy statistics for the model selection methods were computed based on 200 replications of each of the following combinations of distributions and sample sizes.

Eight Univariate Population Distributions

Three observed univariate test score distributions were used to create population distributions of interest. The distributions are most clearly distinguished in terms of their skewness, though they also differ in their other moments. The most skewed test distribution (skew = -.72) was estimated from 13,185 examinees. Another distribution (skew = -.41) was estimated from 8,746 examinees. The third distribution (skew = -.22) was estimated from 8,215 examinees. Loglinear models were selected for each of these distributions and used as the population distributions in the study. For the skew = -.72 distribution, a loglinear model preserving eight moments was selected. For the skew = -.41 and skew = -.22 distributions, loglinear models preserving six moments were selected. A final, simulated, and approximately normal distribution (skew = 0) was generated and modeled with a loglinear model that preserved two moments. The observed and smoothed probabilities for the four distributions are plotted in Figures 1–4, and the summary statistics are given in Table 2.

Four additional univariate distributions were also used to assess the accuracies of the model selection strategies with respect to the complicated score distributions that arise from the use of rounded formula scores (i.e., the test scoring practice that corrects for score inflation due to guessing by subtracting a proportion of the total number of incorrect answers from the total number of correct answers). The selected models of the two tests (X_P and Y_Q) and two external anchors (A_P and A_Q) were those used in the nonequivalent groups with anchor test (NEAT) design example of von Davier et al. (2004), but for this study these models are used simply as population univariate distributions (i.e., the bivariate aspects of von Davier et al.'s models are not considered in this study). The characteristics of the modeled X_P, A_P, A_Q, and Y_Q score distributions are plotted in Figures 5–8 and described in Table 2. In particular, the abnormally low frequencies that occur at every fifth score interval (i.e., the "teeth") and the abnormally high frequencies at the zero scores are structures that would not be modeled well by loglinear models that preserve only the overall moments in the distribution.

Simulating Sample Distributions From the Population Distributions

Datasets of a desired sample size were created based on the population distributions using the following procedure. First, cumulative probabilities were calculated from the score probabilities of a population distribution. Then, a desired sample size of (0, 1) uniform random

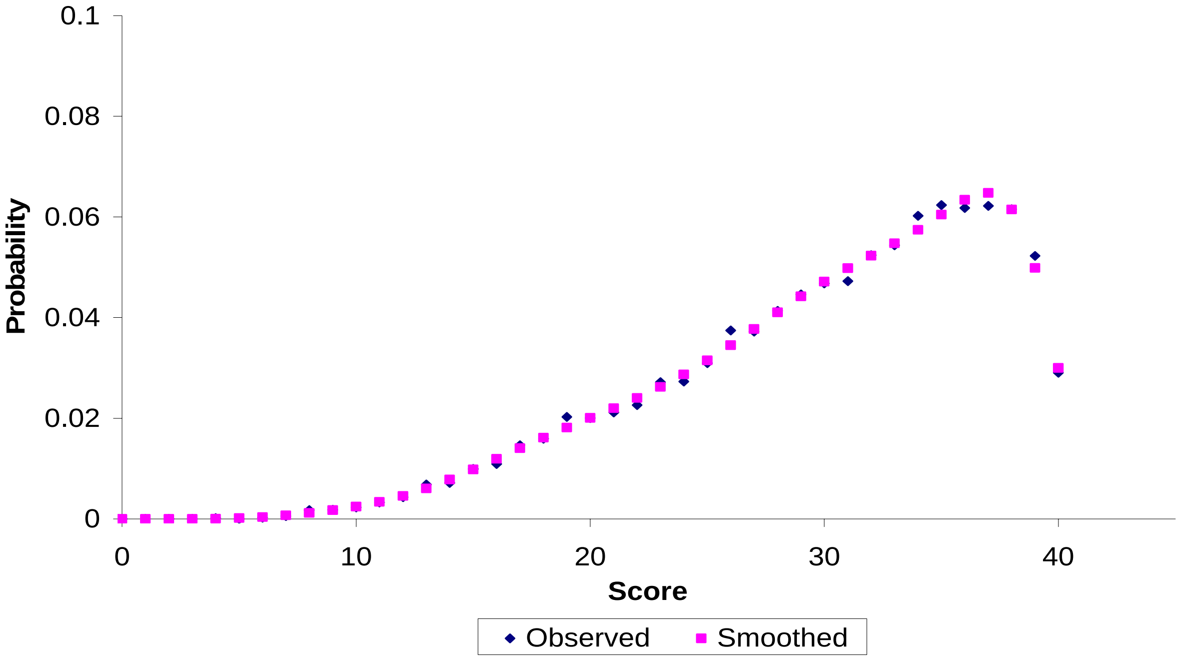

Figure 1. **Skew = -.72 and eight-parameter loglinear model.**

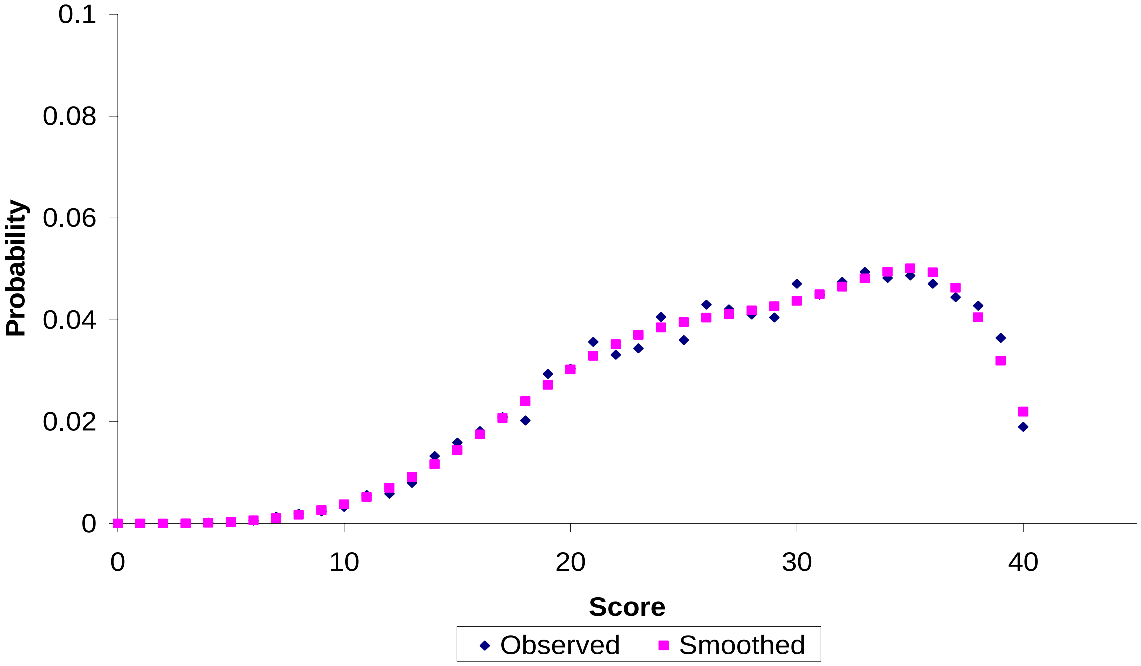

Figure 2. **Skew = -.41 and six-parameter loglinear model.**

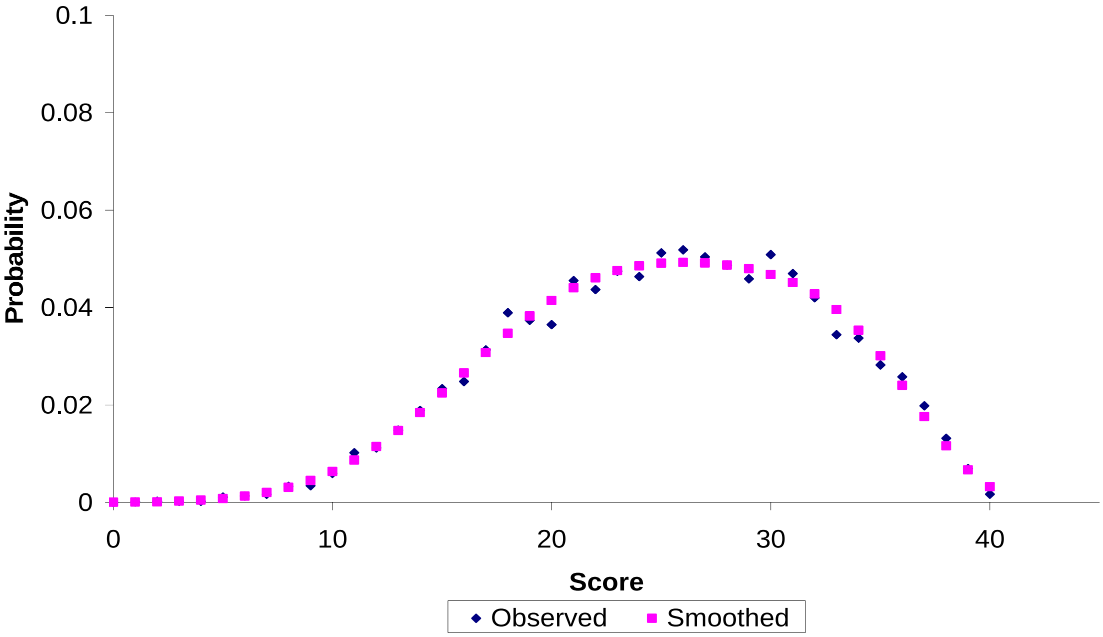

Figure 3. **Skew = -.22 and six-parameter loglinear model.**

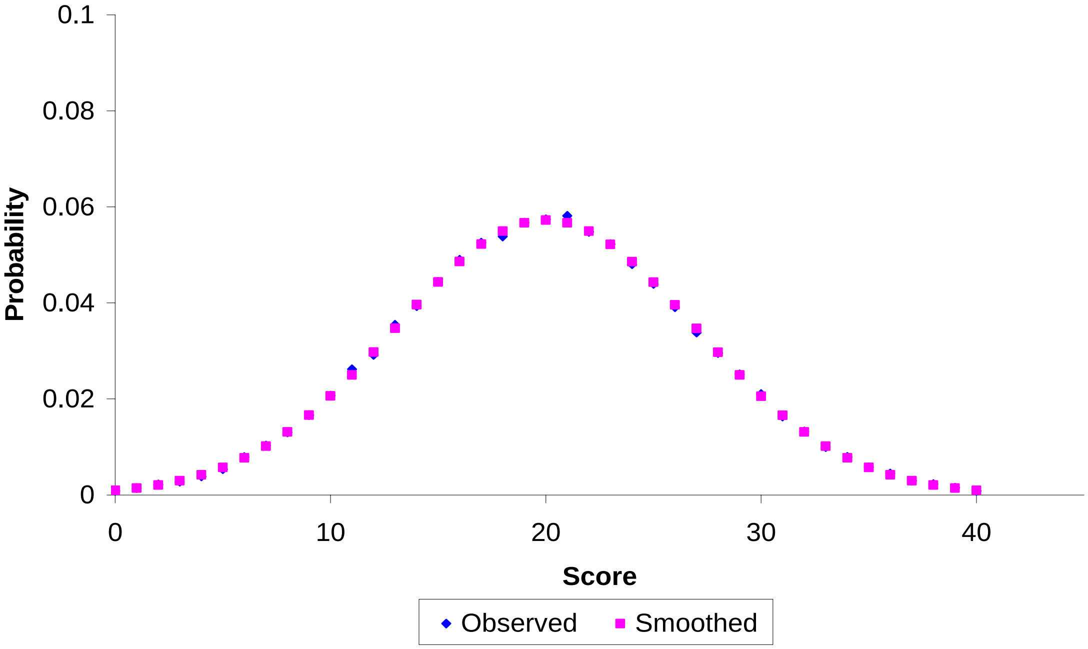

Figure 4. **Skew = .00 and two-parameter loglinear model.**

deviates was generated. The score with the largest cumulative probability that was less than the uniform deviate was assigned to each uniform deviate. The resulting datasets resembled the population distributions upon which they were based, but with a degree of random noise that corresponded to the sample size.

Table 2

Eight Univariate Population Distributions

	Skew							
Item	-.72	-.41	-.22	.00	X_P	A_P	Y_Q	A_Q
Score range	0–40	0–40	0–40	0–40	0–78	0–35	0–78	0–35
Population moments	8	6	6	2	9[a]			
Mean	30.04	28.09	25.18	20.00	39.25	17.05	32.69	14.39
SD	7.07	7.44	7.03	6.88	17.23	8.33	16.73	8.21
Skew	-0.72	-0.41	-0.22	0.00	-0.11	-0.01	0.24	0.26
Kurtosis	-0.17	-0.63	-0.55	-0.19	-0.77	-0.85	-0.69	-0.75

[a] Four overall moments, four moments for the teeth distribution, and one lump at score zero.

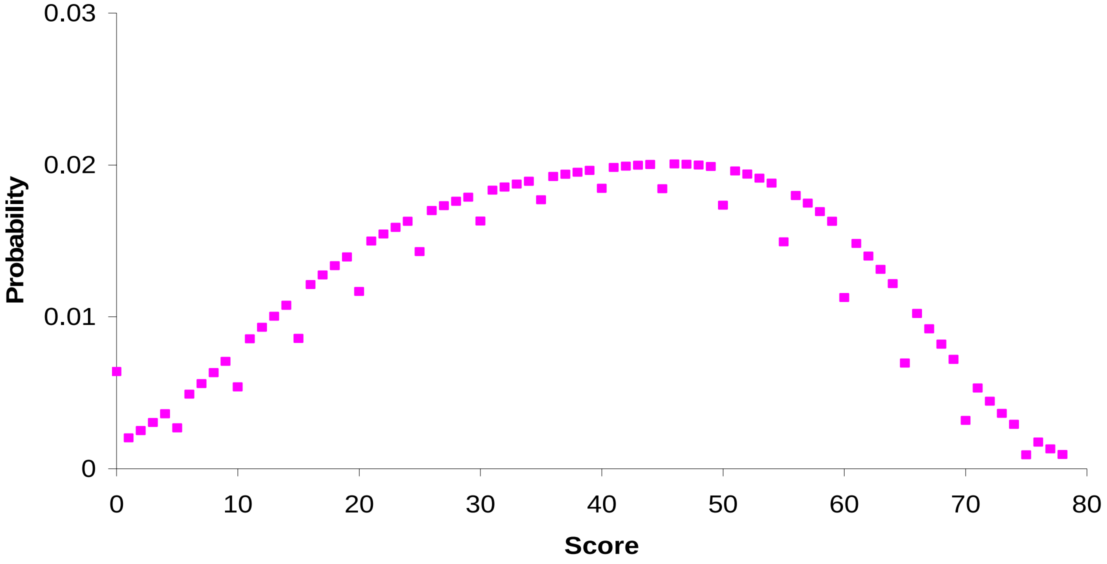

***Figure 5.* X_P score distribution from a nine-parameter loglinear model.**

Sample sizes. Three sample sizes were considered: (a) 100, (b) 1,000, and (c) 5,000.

Parameterization selection for the individual replications. For each individual sample, loglinear model parameterizations were selected based on 16 data-based selection strategies: the simple-to-complex, complex-to-simple, and individual-models significance testing strategies using each of the four overall chi-square statistics (3 x 4 = 12 strategies) and minimization strategies for the four parsimony class statistics (= 4 strategies). These selections were made out of nine possible parameterizations (*I* = 2–10). The individual-models significance testing strategy used a Type I error criterion of .05. The simple-to-complex and complex-to-simple significance testing strategies used a Type I error criterion of .00639 = 1 - (1 - .05)$^{1/8}$. The sample size guidelines shown in Table 1 were also considered.

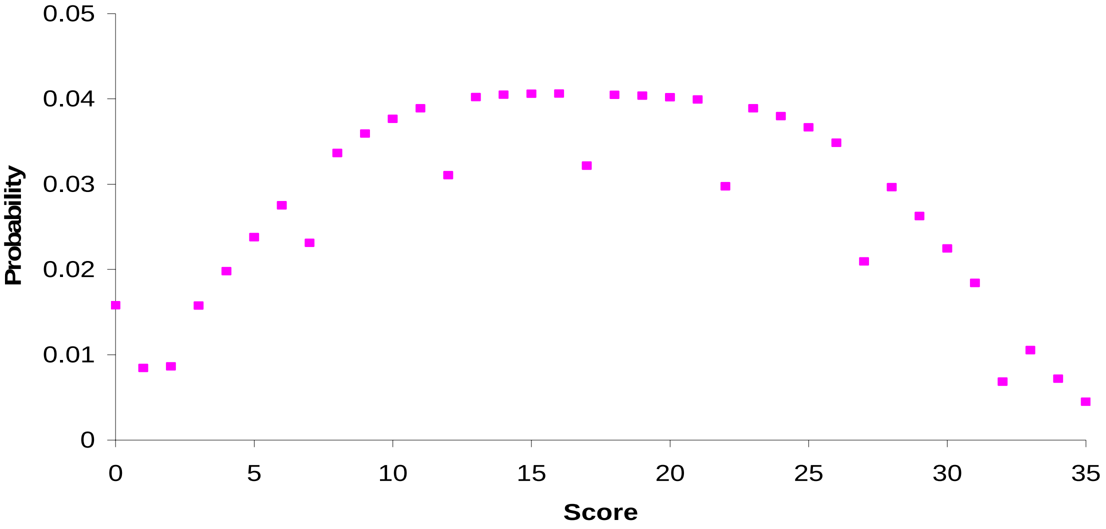

Figure 6. **A$_P$ score distribution from a nine-parameter loglinear model.**

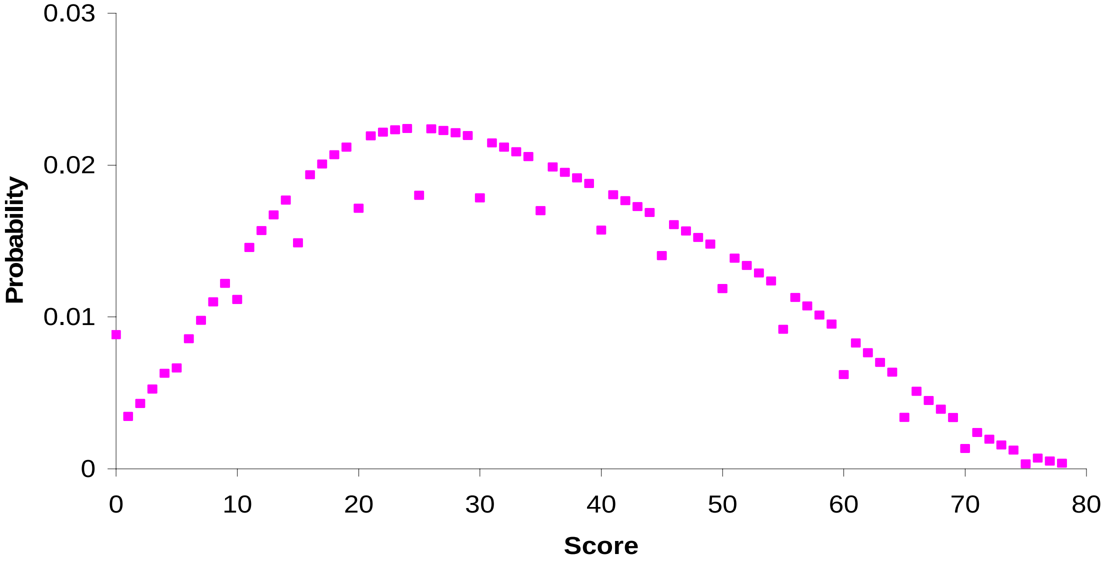

Figure 7. **Y$_Q$ score distribution from a nine-parameter loglinear model.**

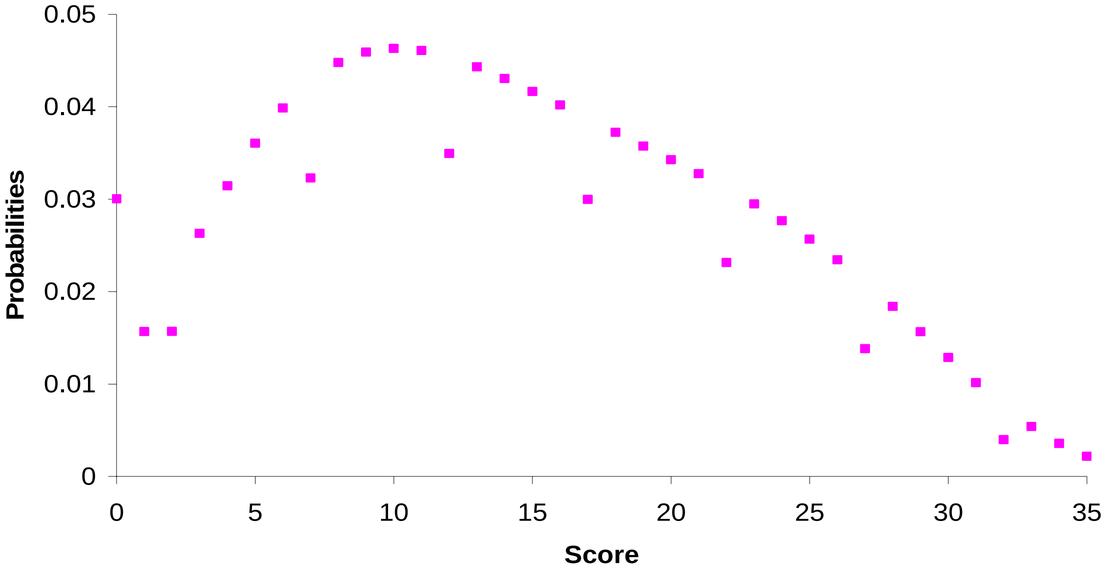

Figure 8. **A_Q score distribution from a nine-parameter loglinear model.**

Loglinear model estimation issues. To maximize convergence rates for model estimation, the loglinear models were fit using orthogonal polynomials of the test scores (degrees = 2–10) rather than the powers shown in Equation 1. All of the models for the sample sizes of 1,000 and 5,000 converged. A number of the models of degrees 9 or 10 did not converge for the sample sizes of 100. For these cases, the model selection procedures used the reduced range of converged models rather than the original range of 2–10 moments.

Traditional equipercentile and kernel equating or linking function evaluation. The implications of the model selection strategies on *X*-to-*Y* equating accuracy were assessed. For this assessment, pairs of the eight distributions were used to set up 14 equivalent groups equating or linking situations (Table 3). In some of the considered scenarios, the *X* and *Y* distributions were sampled from the same population distribution so that equating was not truly needed. In other scenarios, test *X* was sampled from a different population distribution than test *Y*, so that equipercentile equating was needed, the extent of which was based on how much the *X* and *Y* population distributions differed.

For each of 200 total replications, loglinear models for the two distributions were selected based on the model selection strategies. Then, equating was performed for the scores of the *X* distribution to the scores of the *Y* distribution. Results were evaluated with respect to the

Table 3

The Distributions Used for the 14 Evaluated X-to-Y Equating/Linking Situations

X distribution	*Y* distribution	Equating/linking needed?
Skew = -.72	Skew = -.72	No
Skew = -.41	Skew = -.41	No
Skew = -.22	Skew = -.22	No
Skew = .00	Skew = .00	No
X_P	X_P	No
Y_Q	Y_Q	No
Skew = -.72	Skew = -.41	Some
Skew = -.41	Skew = -.22	Some
Skew = -.22	Skew = 0	Some
Skew = -.72	Skew = -.22	Lots
Skew = -.41	Skew = 0	Lots
X_P	A_P	Lots
Y_Q	A_Q	Lots
X_P	Y_Q	Lots

equating functions computed in the population distributions. The traditional equipercentile method based on percentile ranks (Kolen & Brennan, 2004) and the kernel method based on cumulative density functions continuized by Gaussian kernel smoothing (as described in von Davier et al., 2004) were both evaluated.

For example, one considered situation from Table 3 involved equating *X* to *Y*, where *X* was the skew = -.41 distribution and Y was the skew = 0 distribution. Six hundred sample distributions were simulated based on the population skew = -.41 distribution and six hundred additional sample distributions were simulated based on the population skew = 0 distribution. The 600 sample distributions included 200 sample distributions of 100 observations each, 200 additional sample distributions of 1,000 observations each, and 200 additional sample distributions of 5,000 observations each. Two hundred kernel and traditional equipercentile equating functions were computed for 200 pairs of the *X* and *Y* sample distributions of a given sample size (the *X* and *Y* sample distributions were of equal sample size), and these sample equating functions were aggregated to assess equating function accuracy. This process was repeated for the remaining 13 equating conditions summarized in Table 3.

Equating function accuracy was assessed in terms of weighted-average absolute differences (WAD) and weighted-average variability (WAV). These indices were computed as,

$$WAD_{SampleSize,Selection}=\sum_j \left|\mu_{ey,SampleSize,Selection}(x_j)-e_{y,Populations}(x_j)\right| P(X_{Population}=x_j), \tag{10}$$

$$WAV_{SampleSize,Selection}=\sum_j \sigma_{ey,SampleSize,Selection}(x_j)P(X_{Population}=x_j), \tag{11}$$

where the μs and the σs denote the average and standard deviation of the 200 equated scores at x_j, $e_{y,Populations}(x_j)$ is the equated score at x_j based on the population distributions, and the *P*() terms are the population probabilities at score x_j. In preliminary analyses, versions of Equation 10 based on actual, squared, and absolute values were considered, and the absolute values were found to be most informative. Measures of actual equated score differences tended to underestimate the extent to which equating functions differed because the curvilinear equating functions weaved around each other so that the positive and negative equated score differences cancelled out. Measures of average squared differences, such as the squared bias part of mean squared error, are similar to the measure of absolute differences in Equation 10 but more directly focused on average squared differences (or root mean-squared differences) rather than on the average absolute differences (i.e., mean root-squared differences) that were of interest in this study.

Results

Model Selection Accuracy

The model selection accuracy results were tabulated to comparatively evaluate the 17 model selection strategies across the eight population distributions and three considered sample sizes. These results were organized into 8 X 3 = 24 total tables, from which 9 especially representative tables were selected for this section’s discussion. The accuracies of the model selection strategies are summarized in Tables 4–12 for the skew = -.41 distribution (Tables 4–6), the skew = 0 distribution (Tables 7–9), and the X_P distribution (Tables 10–12). Each table summarizes the selection strategies’ preferences in terms of the average parameters selected and the percentage of models with 2, 3, …., 10 parameters selected for 200 random samples of a particular size and drawn from a particular population distribution. For the skew = -.41 and skew = 0 distributions, the population parameterizations were included in the range of considered models (i.e., *I*’s of 6 and 2 were in the range of the considered *I* = 2, 3, …., 10), so that the results

Table 4

Model Selection Accuracy Percentage for the Skew = -.41 Distribution, N = 100, Six Parameters in the Population Model

	Avg. no. parameters	Selected no. of parameters (out of 200 replications): % accuracy								
Selection strategy	selected	2	3	4	5	6[a]	7	8	9	10
G^2										
Complex-to-simple	2.62	83	3	7	1	2	1	3	2	1
Simple-to-complex	2.31	88	3	6	1	1	0	1	1	1
Individual-models	2.06	99	0	1	0	1	0	1	0	0
χ_P^2										
Complex-to-simple	3.10	72	2	8	3	7	4	3	2	2
Simple-to-complex	2.57	83	2	8	2	3	1	1	1	1
Individual-models	2.04	99	0	1	0	0	0	1	0	0
χ_{FT}^2										
Complex-to-simple	2.40	89	1	6	0	1	0	3	1	1
Simple-to-complex	2.16	94	1	4	0	0	0	1	1	0
Individual-models	2.03	100	0	0	0	0	0	1	0	0
CR										
Complex-to-simple	2.68	81	2	8	2	3	2	2	1	1
Simple-to-complex	2.36	89	2	5	1	1	0	1	2	0
Individual-models	2.03	100	0	0	0	0	0	1	0	0
AIC	3.57	47	11	20	5	8	3	4	2	2
BIC	2.24	86	5	8	1	0	0	0	0	0
CAIC	2.14	91	4	5	0	0	0	0	0	0
Goodman	5.16	30	14	6	7	8	8	10	7	12
Sample size selection	3.00	0	100	0	0	0	0	0	0	0

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion.

[a] Population model.

Table 5

Model Selection Accuracy for the Skew = -.41 Distribution, N = 1,000, Six Parameters in the Population Model

	Average no.	Selected no. of parameters (out of 200 replications): % accuracy								
Selection strategy	parameters selected	2	3	4	5	6[a]	7	8	9	10
G^2										
Complex-to-simple	4.13	14	7	54	10	12	1	1	2	1
Simple-to-complex	4.43	7	3	57	14	16	2	1	1	1
Individual-models	2.97	49	15	31	4	2	0	0	0	1
χ_P^2										
Complex-to-simple	4.27	22	3	36	20	14	2	2	2	2
Simple-to-complex	4.40	17	3	37	22	17	3	1	1	1
Individual-models	2.82	61	15	15	8	1	1	0	1	1
χ_{FT}^2										
Complex-to-simple	3.96	18	6	56	13	7	1	0	1	1
Simple-to-complex	4.23	9	4	56	19	12	1	0	0	0
Individual-models	2.75	59	14	23	3	2	0	0	0	0
CR										
Complex-to-simple	4.20	20	3	43	17	14	1	1	2	1
Simple-to-complex	4.40	14	3	42	21	19	1	2	1	0
Individual-models	2.79	60	14	18	7	2	0	0	0	1
AIC	5.68	0	1	26	20	36	8	4	4	4
BIC	3.60	27	8	51	11	5	0	0	0	0
CAIC	3.24	41	8	42	8	2	0	0	0	0
Goodman	5.22	11	17	21	16	11	7	4	6	10
Sample size selection	5.00	0	0	0	100	0	0	0	0	0

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion.

[a] Population model.

Table 6

Model Selection Accuracy for the Skew = -.41 Distribution, N = 5,000, Six Parameters in the Population Model

	Avg. no. parameters	Selected no. of parameters (out of 200 replications): % accuracy								
Selection strategy	selected	2	3	4	5	6[a]	7	8	9	10
G^2										
Complex-to-simple	5.67	0	0	5	33	58	2	2	0	1
Simple-to-complex	5.69	0	0	3	33	61	3	1	0	1
Individual-models	5.06	0	0	32	49	15	1	1	0	4
χ_P^2										
Complex-to-simple	5.65	0	0	3	42	51	2	2	1	1
Simple-to-complex	5.68	0	0	2	40	53	3	2	1	1
Individual-models	5.06	0	0	25	60	11	2	0	0	4
χ_{FT}^2										
Complex-to-simple	5.59	0	0	5	39	53	1	2	0	1
Simple-to-complex	5.70	0	0	3	36	56	3	1	1	1
Individual-models	4.88	0	0	37	49	11	1	0	1	2
CR										
Complex-to-simple	5.63	0	0	3	41	53	2	2	0	1
Simple-to-complex	5.68	0	0	2	39	54	3	1	1	1
Individual-models	5.02	0	0	29	56	11	1	0	0	4
AIC	6.50	0	0	0	4	67	16	8	2	5
BIC	5.42	0	0	10	40	50	1	0	0	0
CAIC	5.31	0	0	13	45	42	1	0	0	0
Goodman	6.66	0	0	5	35	20	11	8	9	15
Sample size selection	5.00	0	0	0	100	0	0	0	0	0

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion.

[a] Population model.

Table 7

Model Selection Accuracy for the Skew = 0 Distribution, N = 100, Two Parameters in the Population Model

	Avg. no. parameters	Selected no. of parameters (out of 200 replications): % accuracy								
Selection strategy	selected	2[a]	3	4	5	6	7	8	9	10
G^2										
Complex-to-simple	2.40	92	1	1	2	1	1	1	1	2
Simple-to-complex	2.05	98	1	1	1	0	0	0	0	0
Individual-models	2.18	98	0	0	1	0	0	0	0	2
χ_P^2										
Complex-to-simple	3.16	75	4	3	4	3	1	9	1	3
Simple-to-complex	2.49	86	4	3	2	2	2	2	1	1
Individual-models	2.17	95	2	2	0	0	0	0	1	1
χ_{FT}^2										
Complex-to-simple	2.07	99	0	1	1	0	0	0	0	1
Simple-to-complex	2.01	100	0	1	0	0	0	0	0	0
Individual-models	2.00	100	0	0	0	0	0	0	0	0
CR										
Complex-to-simple	2.58	87	2	2	2	1	1	3	1	3
Simple-to-complex	2.14	96	2	1	1	0	1	0	0	1
Individual-models	2.11	98	1	0	0	0	1	0	0	1
AIC	2.94	71	9	5	3	5	2	3	1	2
BIC	2.06	96	3	1	1	0	0	0	0	0
CAIC	2.03	98	2	1	0	0	0	0	0	0
Goodman	5.41	25	14	8	9	5	11	6	10	14
Sample size selection	3.00	0	100	0	0	0	0	0	0	0

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion.

[a] Population model.

Table 8

Model Selection Accuracy for the Skew = 0 Distribution, N = 1,000, Two Parameters in the Population Model

	Avg. no. parameters	Selected no. of parameters (out of 200 replications): % accuracy								
Selection strategy	selected	2[a]	3	4	5	6	7	8	9	10
G^2										
Complex-to-simple	2.37	93	1	1	0	1	1	1	3	1
Simple-to-complex	2.09	98	1	1	0	0	0	1	1	0
Individual-models	2.37	95	1	0	0	0	0	0	0	5
χ_P^2										
Complex-to-simple	2.40	92	2	1	0	2	2	1	2	1
Simple-to-complex	2.12	97	2	0	0	0	0	1	1	1
Individual-models	2.22	97	1	0	1	0	0	0	0	3
χ_{FT}^2										
Complex-to-simple	2.29	94	1	1	0	1	1	1	2	1
Simple-to-complex	2.10	97	1	2	0	0	1	1	0	0
Individual-models	2.40	95	0	0	0	0	0	0	1	5
CR										
Complex-to-simple	2.25	96	1	0	0	1	1	1	2	1
Simple-to-complex	2.04	99	1	0	0	0	0	1	0	0
Individual-models	2.22	97	1	0	1	0	0	0	0	3
AIC	2.88	69	12	8	4	2	2	1	3	2
BIC	2.02	99	2	0	0	0	0	0	0	0
CAIC	2.01	99	1	0	0	0	0	0	0	0
Goodman	5.61	26	8	13	7	7	6	5	10	19
Sample size selection	5.00	0	0	0	100	0	0	0	0	0

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion.

[a] Population model.

Table 9

Model Selection Accuracy for the Skew = 0 Distribution, N = 5,000, Two Parameters in the Population Model

	Avg. no. parameters	Selected no. of parameters (out of 200 replications): % accuracy								
Selection strategy	selected	2^a	3	4	5	6	7	8	9	10
G^2										
Complex-to-simple	2.37	92	1	1	2	1	1	2	1	1
Simple-to-complex	2.08	98	1	0	0	0	0	1	0	1
Individual-models	2.23	97	0	0	1	0	0	0	1	2
χ_P^2										
Complex-to-simple	2.34	94	1	0	2	1	1	2	1	1
Simple-to-complex	2.07	98	1	1	1	0	0	0	0	1
Individual-models	2.16	97	1	1	1	0	0	0	1	1
χ_{FT}^2										
Complex-to-simple	2.36	92	2	0	2	1	1	2	1	1
Simple-to-complex	2.09	98	2	0	0	0	0	1	0	1
Individual-models	2.21	97	0	0	1	0	0	0	1	2
CR										
Complex-to-simple	2.36	93	1	0	2	1	1	2	1	1
Simple-to-complex	2.08	98	1	1	1	0	0	0	0	1
Individual-models	2.16	97	1	0	1	0	1	0	1	1
AIC	2.88	69	12	6	5	3	2	3	1	1
BIC	2.01	100	1	0	0	0	0	0	0	0
CAIC	2.01	100	1	0	0	0	0	0	0	0
Goodman	5.87	19	11	9	9	10	9	6	11	18
Sample size selection	5.00	0	0	0	100	0	0	0	0	0

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion.

[a] Population model.

Table 10

Model Selection Accuracy for the X_P Distribution, N = 100

	Avg. no. parameters	Selected no. of parameters (out of 200 replications): % accuracy								
Selection strategy	selected	2	3	4	5	6	7	8	9	10
G^2										
Complex-to-simple	2.55	83	2	8	2	3	2	1	1	1
Simple-to-complex	2.32	89	2	6	1	3	2	0	0	0
Individual-models	2.69	88	1	3	0	1	1	1	1	7
$\chi_P{}^2$										
Complex-to-simple	3.43	67	4	6	5	5	3	4	4	4
Simple-to-complex	2.62	82	5	4	3	2	1	2	1	2
Individual-models	2.28	93	2	2	0	1	0	0	0	3
$\chi_{FT}{}^2$										
Complex-to-simple	2.29	92	1	4	1	1	0	1	1	1
Simple-to-complex	2.08	97	1	1	1	1	0	0	0	0
Individual-models	2.05	99	0	1	0	0	0	0	0	1
CR										
Complex-to-simple	2.81	79	4	5	3	4	3	1	1	3
Simple-to-complex	2.41	86	4	4	2	2	1	2	0	1
Individual-models	2.18	97	0	1	0	1	1	0	0	2
AIC	3.69	43	10	21	9	9	3	3	2	2
BIC	2.22	87	6	7	1	0	0	0	0	0
CAIC	2.10	94	4	2	1	0	0	0	0	0
Goodman	5.53	19	14	11	12	8	8	9	8	14
Sample size selection	3.00	0	100	0	0	0	0	0	0	0

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion.

Table 11

Model Selection Accuracy for the X_P Distribution, N = 1,000

	Avg. no. parameters	Selected no. of parameters (out of 200 replications): % accuracy								
Selection strategy	selected	2	3	4	5	6	7	8	9	10
G^2										
Complex-to-simple	4.20	16	8	59	4	3	5	3	2	3
Simple-to-complex	4.36	9	7	64	6	4	4	5	1	2
Individual-models	5.01	26	10	30	4	3	3	4	1	21
χ_P^2										
Complex-to-simple	4.20	23	16	34	8	5	8	4	2	3
Simple-to-complex	4.51	16	12	38	11	5	9	7	2	2
Individual-models	4.30	37	17	16	4	5	4	4	1	13
χ_{FT}^2										
Complex-to-simple	4.14	14	8	64	3	1	5	3	1	3
Simple-to-complex	4.27	9	7	68	4	4	3	3	1	3
Individual-models	4.72	30	10	30	5	1	2	3	1	19
CR										
Complex-to-simple	4.09	24	12	41	6	4	7	4	1	3
Simple-to-complex	4.44	15	10	46	10	5	7	5	1	3
Individual-models	4.45	35	15	19	4	4	6	3	1	15
AIC	5.99	0	2	36	17	7	13	8	10	9
BIC	3.29	34	9	55	3	1	0	0	0	0
CAIC	3.10	42	10	46	2	1	0	0	0	0
Goodman	7.09	4	3	14	13	8	11	10	17	23
Sample size selection	5.00	0	0	0	100	0	0	0	0	0

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; CAIC = consistent Akaike information criterion.

Table 12

Model Selection Accuracy for the X_P Distribution, N = 5,000

	Avg. no. parameters	Selected no. of parameters (out of 200 replications): % accuracy								
Selection strategy	selected	2	3	4	5	6	7	8	9	10
G^2										
Complex-to-simple	7.18	0	0	22	7	5	15	16	20	16
Simple-to-complex	7.81	0	0	11	7	6	13	18	25	22
Individual-models	9.96	0	0	0	0	0	1	1	1	98
χ_P^2										
Complex-to-simple	7.51	0	0	9	10	8	18	20	22	14
Simple-to-complex	7.85	0	0	6	7	7	16	23	25	17
Individual-models	9.91	0	0	0	0	1	1	1	2	96
χ_{FT}^2										
Complex-to-simple	7.03	0	0	28	7	4	10	15	19	18
Simple-to-complex	7.64	0	0	19	6	5	9	14	26	23
Individual-models	9.95	0	0	0	0	0	1	2	1	97
CR										
Complex-to-simple	7.36	0	0	13	12	7	17	18	22	14
Simple-to-complex	7.80	0	0	8	7	6	16	20	27	17
Individual-models	9.93	0	0	0	0	1	1	2	1	97
AIC	9.04	0	0	2	2	2	8	12	22	54
BIC	4.56	0	0	72	17	4	2	3	2	1
CAIC	4.35	0	0	78	16	2	2	2	1	0
Goodman	9.26	0	0	1	1	1	7	12	20	60
Sample size selection	5.00	0	0	0	100	0	0	0	0	0

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion.

given in Tables 4–9 are indicative of selection strategies' accuracies. For the X_P distribution, the population parameterization and sample distributions included score-specific features (i.e., teeth) that were not included in the range of considered models (i.e., models with I = 2, 3, …., 10 do not directly fit the distributions of the teeth), so that the X_P results given in Tables 10–12 are indicative of selection strategies' preferences when considering a series of incorrect models.

One important result in Tables 4–12 is the influence of sample size on the model selection strategies across the population distributions. When the population model contained many parameters (e.g., the skew = -.41 population distribution contains 6 parameters, Tables 4–6), the selection strategies were least accurate for small sample sizes (N = 100, Table 4) and most accurate for large sample sizes (N = 5,000, Table 6). The accuracies of the model selection strategies for the skew = 0 distribution (two parameters in the population distribution, Tables 7–9) were relatively high and not strongly influenced by the three sample-size conditions. When the population model was not among the considered models, such as for the X_P distribution (Tables 10–12), then large sample sizes caused all of the selection strategies to select models with many parameters (e.g., N = 5,000, Table 12).

The AIC selection strategy tended to select models with more parameters than most of the other selection strategies, resulting in relatively high selection accuracy in selecting models for the skew = -.41 population distribution (Tables 4–6) and relatively low selection accuracy for selecting models for the skew = 0 population distribution (Tables 7-9). In terms of the other parsimony-class selection strategies, the BIC favored models with fewer parameters than the AIC, and the CAIC favored models with fewer parameters than the BIC (corresponding with the penalties with which these statistics were designed, Equations 6–8). The Goodman selection strategy favored models with many parameters and was so inconsistent in its selection that it cannot be recommended for general practice.

The selection strategies based on the likelihood ratio, Pearson, Freeman-Tukey and Cressie-Read chi-square statistics favored simple models with two or three parameters for the sample sizes of 100 (Tables 4, 7, and 10). The complex-to-simple selection strategies selected models with more parameters than the simple-to-complex strategies for sample sizes of 100. The differences between these two approaches were small and inconsistent for the sample sizes of 1,000 and 5,000. The individual-models selection strategy favored the simplest models out of all the selection strategies for the skew = -.41 and skew = 0 distributions but selected models with many parameters when samples of 5,000 were drawn from the X_P distribution (Table 12). No overwhelming winner emerged in terms of accuracy among the four chi-square statistics, though model selections based on the Freeman-Tukey statistic were typically the least accurate.

Equating Function Accuracy

To assess the impact of the 17 model-selection strategies on equating function accuracy, results were assessed across three sample sizes, 14 equating conditions, and two equating methods (kernel and traditional equipercentile), for a total of 28 results tables. This section summarizes the 28 tables' results by focusing on 5 representative tables involving the kernel equating results, which are very similar to the traditional equipercentile results; 3 of the 6 no-equating-needed result tables, and 2 of the 5 lots-of-equating needed result tables. The omitted results are similar to and within the range of the results that are presented. The presented results show three no-equating-needed situations, where the *X* and *Y* samples were drawn from the same population distribution, including (a) the skew = -.41 distribution (Table 13), (b) the skew = 0 distribution (Table 14), and (c) the X_P distribution (Table 15). Two lots-of-equating needed situations are also presented, one where the *X* samples were drawn from the skew = -.41 distribution and the *Y* samples were drawn from the skew = 0 distribution (Table 16), and another where the *X* samples were drawn from the X_P distribution and the *Y* samples were drawn from the Y_Q distribution (Table 17). All of the results tables present the WAD and WAV values for the 16 data-based, model-selection strategies; the sample-size selection strategy (Table 1); and, for reference, an additional set of WAD and WAV values for evaluating equatings based on using the population models for all replications (i.e., always-fit-the-population models). The results of the remaining nine equating situations considered in this study (Table 3) were similar to the results of the situations summarized in Tables 13–17.

A general result across Tables 13–17 is the influence of sample size on the sample equating functions' absolute deviations from the population equating function (WAD) and on the sample equating functions' variability (WAV). Large sample sizes reduced WAD values because they created situations where the model selection strategies were more accurate (Tables 4–6) or highly parameterized (Tables 10–12). In addition, large sample sizes produced more stable equating results and smaller WAV values for the equating functions based on all of the selection strategies.

Table 13

The X-to-Y Kernel Equating Situation, X and Y Sampled From the Skew = -.41 Distribution, No Equating Needed

	$N_X = N_Y = 100$		$N_X = N_Y = 1{,}000$		$N_X = N_Y = 5{,}000$	
Selection strategy	WAD	WAV	WAD	WAV	WAD	WAV
G^2						
Complex-to-simple	0.056	1.261	0.010	0.451	0.029	0.202
Simple-to-complex	0.047	1.254	0.013	0.444	0.028	0.199
Individual-models	0.045	1.197	0.023	0.478	0.028	0.206
χ_P^2						
Complex-to-simple	0.048	1.291	0.012	0.471	0.029	0.200
Simple-to-complex	0.047	1.259	0.007	0.470	0.029	0.199
Individual-models	0.045	1.203	0.031	0.469	0.029	0.202
χ_{FT}						
Complex-to-simple	0.050	1.239	0.012	0.458	0.029	0.202
Simple-to-complex	0.043	1.237	0.012	0.447	0.028	0.200
Individual-models	0.044	1.193	0.023	0.470	0.028	0.205
CR						
Complex-to-simple	0.047	1.269	0.008	0.470	0.029	0.200
Simple-to-complex	0.047	1.251	0.008	0.461	0.029	0.200
Individual-models	0.044	1.193	0.032	0.472	0.029	0.202
AIC	0.064	1.360	0.012	0.436	0.029	0.194
BIC	0.053	1.264	0.017	0.473	0.029	0.203
CAIC	0.049	1.242	0.016	0.486	0.029	0.206
Goodman	0.071	1.304	0.024	0.461	0.028	0.201
Sample size selection	0.048	1.279	0.010	0.412	0.028	0.186
Always-fit-the-population models	0.060	1.386	0.011	0.425	0.028	0.190

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion, WAD = weighted-average absolute differences, WAV = weighted-average variability.

Table 14

The X-to-Y Kernel Equating Situation, X and Y Sampled From the Skew = 0 Distribution, No Equating Needed

	$N_X = N_Y = 100$		$N_X = N_Y = 1,000$		$N_X = N_Y = 5,000$	
Selection strategy	WAD	WAV	WAD	WAV	WAD	WAV
G^2						
Complex-to-simple	0.016	1.125	0.020	0.372	0.018	0.165
Simple-to-complex	0.018	1.102	0.018	0.370	0.018	0.163
Individual-models	0.017	1.081	0.018	0.369	0.018	0.164
χ_P^2						
Complex-to-simple	0.033	1.177	0.022	0.374	0.018	0.165
Simple-to-complex	0.020	1.137	0.019	0.370	0.018	0.164
Individual-models	0.017	1.088	0.018	0.365	0.019	0.164
χ_{FT}^2						
Complex-to-simple	0.015	1.100	0.019	0.371	0.018	0.165
Simple-to-complex	0.016	1.077	0.018	0.370	0.018	0.163
Individual-models	0.016	1.076	0.018	0.366	0.018	0.164
CR						
Complex-to-simple	0.019	1.141	0.022	0.372	0.018	0.165
Simple-to-complex	0.019	1.111	0.019	0.368	0.018	0.163
Individual-models	0.017	1.079	0.019	0.365	0.018	0.164
AIC	0.037	1.196	0.023	0.396	0.018	0.179
BIC	0.021	1.113	0.017	0.364	0.018	0.162
CAIC	0.017	1.098	0.017	0.364	0.018	0.162
Goodman	0.025	1.198	0.021	0.418	0.018	0.185
Sample size selection	0.035	1.167	0.021	0.419	0.018	0.187
Always-fit-the-population models	0.016	1.075	0.018	0.357	0.018	0.161

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion, WAD = weighted-average absolute differences, WAV = weighted-average variability.

Table 15

The X-to-Y Kernel Equating Situation, X and Y Sampled From the X_P Distribution, No Equating Needed

	$N_X = N_Y = 100$		$N_X = N_Y = 1{,}000$		$N_X = N_Y = 5{,}000$	
Selection strategy	WAD	WAV	WAD	WAV	WAD	WAV
G^2						
Complex-to-simple	0.060	3.024	0.093	1.143	0.053	0.494
Simple-to-complex	0.082	3.035	0.096	1.107	0.053	0.490
Individual-models	0.057	2.966	0.097	1.174	0.053	0.481
$\chi_P{}^2$						
Complex-to-simple	0.058	3.100	0.090	1.188	0.052	0.492
Simple-to-complex	0.060	3.060	0.090	1.153	0.053	0.490
Individual-models	0.071	2.972	0.087	1.178	0.053	0.482
$\chi_{FT}{}^2$						
Complex-to-simple	0.063	2.962	0.094	1.138	0.053	0.497
Simple-to-complex	0.081	2.947	0.092	1.101	0.052	0.491
Individual-models	0.066	2.903	0.095	1.183	0.054	0.482
CR						
Complex-to-simple	0.052	3.036	0.090	1.180	0.053	0.494
Simple-to-complex	0.062	3.033	0.095	1.142	0.053	0.490
Individual-models	0.071	2.925	0.092	1.185	0.053	0.481
AIC	0.056	3.211	0.087	1.079	0.053	0.484
BIC	0.072	3.045	0.093	1.182	0.053	0.493
CAIC	0.074	3.008	0.087	1.199	0.053	0.494
Goodman	0.047	3.179	0.087	1.096	0.053	0.483
Sample size selection	0.058	3.116	0.086	1.042	0.053	0.462
Always-fit-the-population models	0.064	3.183	0.082	1.023	0.053	0.454

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion, WAD = weighted-average absolute differences, WAV = weighted-average variability.

Table 16

The X-to-Y Kernel Equating Situation, X sampled From the Skew = -.41 Distribution and Y Sampled From the Skew = 0 Distribution, Lots of Equating Needed

	$N_X = N_Y = 100$		$N_X = N_Y = 1{,}000$		$N_X = N_Y = 5{,}000$	
Selection strategy	WAD	WAV	WAD	WAV	WAD	WAV
G^2						
Complex-to-simple	0.307	1.277	0.102	0.409	0.026	0.193
Simple-to-complex	0.316	1.265	0.086	0.401	0.021	0.191
Individual-models	0.333	1.246	0.184	0.410	0.071	0.195
χ_P^2						
Complex-to-simple	0.274	1.311	0.091	0.417	0.027	0.191
Simple-to-complex	0.298	1.281	0.082	0.412	0.026	0.190
Individual-models	0.333	1.244	0.207	0.406	0.068	0.193
χ_{FT}^2						
Complex-to-simple	0.317	1.258	0.111	0.409	0.028	0.192
Simple-to-complex	0.328	1.250	0.094	0.403	0.025	0.191
Individual-models	0.339	1.239	0.210	0.404	0.075	0.194
CR						
Complex-to-simple	0.292	1.289	0.094	0.413	0.027	0.192
Simple-to-complex	0.318	1.264	0.079	0.406	0.025	0.191
Individual-models	0.339	1.241	0.204	0.405	0.071	0.194
AIC	0.211	1.352	0.044	0.411	0.008	0.194
BIC	0.311	1.271	0.114	0.407	0.034	0.193
CAIC	0.328	1.258	0.135	0.415	0.042	0.195
Goodman	0.219	1.340	0.087	0.428	0.037	0.201
Sample size selection	0.282	1.304	0.070	0.404	0.063	0.193
Always-fit-the-population models	0.149	1.339	0.018	0.387	0.006	0.185

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion, WAD = weighted-average absolute differences, WAV = weighted-average variability.

Table 17

The X-to-Y Kernel Equating Situation, X Sampled From the X_P Distribution and Y Sampled From the Y_Q Distribution, Lots of Equating Needed

	$N_X = N_Y = 100$		$N_X = N_Y = 1{,}000$		$N_X = N_Y = 5{,}000$	
Selection strategy	WAD	WAV	WAD	WAV	WAD	WAV
G^2,						
Complex-to-simple	0.657	2.979	0.101	1.067	0.046	0.487
Simple-to-complex	0.645	2.983	0.059	1.056	0.044	0.485
Individual-models	0.665	2.950	0.193	1.101	0.066	0.477
χ_P^2						
Complex-to-simple	0.490	3.089	0.150	1.107	0.040	0.487
Simple-to-complex	0.490	3.049	0.111	1.089	0.046	0.485
Individual-models	0.693	2.932	0.295	1.105	0.065	0.477
χ_{FT}^2						
Complex-to-simple	0.739	2.921	0.091	1.058	0.050	0.487
Simple-to-complex	0.746	2.905	0.064	1.053	0.045	0.487
Individual-models	0.756	2.894	0.221	1.096	0.065	0.476
CR						
Complex-to-simple	0.595	3.023	0.160	1.108	0.041	0.488
Simple-to-complex	0.617	2.997	0.101	1.081	0.045	0.484
Individual-models	0.722	2.908	0.292	1.107	0.065	0.477
AIC	0.281	3.158	0.035	1.029	0.058	0.480
BIC	0.579	2.989	0.210	1.114	0.056	0.481
CAIC	0.658	2.942	0.292	1.124	0.063	0.480
Goodman	0.273	3.178	0.033	1.049	0.059	0.479
Sample size selection	0.356	3.075	0.079	0.997	0.068	0.455
Always-fit-the-population models	0.349	3.428	0.055	0.974	0.050	0.447

Note. AIC = Akaike information criterion, BIC = Bayesian information criterion, CAIC = consistent Akaike information criterion, WAD = weighted-average absolute differences, WAV = weighted-average variability.

For the three no-equating-needed situations (Tables 13–15), the WAD values of the selection strategies from the criterion identity equating function were so small (< 0.1 raw score point) that they might be considered negligible in actual equating practice. Because WAD values were so small, no overwhelming winners or losers emerged from the selection strategies in Tables 13–15. One interesting finding is that the AIC, which selected models with more parameters than many of the other strategies, produced equating functions with slightly larger WAD values than other strategies for the sample sizes of 100 and 1,000. Always-fitting-the-population models did not have large advantages over the selection strategies in terms of accuracy in estimating the population identity equating function.

For the two lots-of-equating needed situations (Tables 16–17), the differences in selection strategies' WAD values were more visible than for the no-equating-needed situations. WAD values were above 0.3 and 0.6 raw score points for sample sizes of 100, and they differentiated the selection strategies according to strategies' tendencies to select more and fewer parameters. The AIC strategy (which usually selected models with the largest number of parameters) often produced the most accurate equating functions, with relatively small WAD values that were usually the closest out of all the selection strategies to the WAD values produced from always-fitting-the-population models. The chi-square-based individual-models selection strategies (which usually selected models with relatively few parameters) produced equating functions with larger WAD values than those of other selection strategies. The CAIC selection strategy produced equating functions with larger WAD values than the BIC and AIC selections, corresponding to the CAIC strategy's preference for models with few parameters. For the sample sizes of 100 and 1,000, the sample-size selection strategy selected models with more parameters than the chi-square selection strategies, resulting in equating functions with smaller WAD values than those from the chi-square strategies and, probably as a result of its consistency, smaller WAV values. For sample sizes of 5,000, the chi-square selection strategies selected more than the five parameters selected by the sample-size selection strategy, resulting in the chi-square strategies that produced equating functions with smaller WAD values than the sample-size selection strategy.

In the *X*-to-*Y* equating situation in Table 17, the models considered by the selection strategies did not include the *X* (X_P) and *Y* (Y_Q) population distributions used to generate the sample distributions. With large sample sizes, the selection strategies addressed the complex, score-specific features of the population distributions (Figures 5 and 7) by selecting models with large numbers of parameters (Tables 11 and 12). The result was that for large sample sizes

(1,000 and 5,000) the equating functions computed based on many of the selection strategies did not deviate much from the population equating functions in terms of the WAD values (Table 17), even though all of the selected models were incorrect. One apparent implication of these results is that the impact of score-specific features from rounded formula score distributions on equating functions can be adequately addressed in terms of loglinear models that include large numbers of distribution-level parameters (i.e., $I \geq 5$). The conditions that produce this implication are probably too complex to make the implication useful for practice, because interactions between sample size and the loglinear models make the performances for the selection strategies inconsistent (e.g., the AIC across the sample size conditions) and difficult to explain (e.g., some of the simple-to-complex selection strategies select models with relatively few parameters but have very good WAD values in Table 17). The approach to modeling rounded formula score distributions through the use of indicator functions (i.e., always-fit-the-population models) avoids some of the complex results produced by the considered selection strategies' use of highly parameterized but incorrect models.

Discussion

The purpose of this study was to compare several common strategies for selecting loglinear models in terms of their accuracy in selecting population models and their effect on equating function accuracy. The study considered a range of sample sizes, population distributions, and population equating functions. The results suggest that selection strategies for loglinear models are most accurate with large sample sizes, and that strategies that favor complex loglinear models over simpler models (i.e., minimizing the AIC statistic) result in the most accurate equating functions across a range of test score distributions. There is always a possibility that the selection process for loglinear models in sample data may add bias and variability to equating, but the added inaccuracy appears to be most serious when the selected models include too few parameters (i.e., fewer than three parameters or moments for most situations) rather than too many parameters.

Implications of Loglinear Model Selection on Equating Function Accuracy

The results of this study may be somewhat unexpected in terms of how small equating inaccuracy was for this study's models and selection strategies. Whereas the accuracies of many selection strategies were not all that high in samples of 1,000 (Table 5), the equating functions that used strategies' selected models in samples of 1,000 were often not problematic in terms of accuracy (Tables 13 and 16). Three issues that influence the association between selection strategies and equating function accuracy are (a) the differences in focus of traditional goodness-of-fit statistics and equated score differences, (b) the complexity of the population distributions and equating function being evaluated, and (c) the measures used to evaluate equating function accuracy.

The focus of loglinear model selection strategies is somewhat different from equipercentile equating, in which the selection strategies try to minimize the misfit in the score frequencies; equipercentile equating is based on continuized, cumulative versions of the score probabilities (e.g., percentile ranks or Gaussian kernel cumulative density functions). The extent of misfit that can occur in frequencies across the test score range is much greater than the misfit that can occur in cumulative probabilities. Frequencies at individual scores can vary somewhat independently of other frequencies, whereas cumulative probabilities vary in a much narrower range because, unlike frequencies, cumulative probabilities cannot decrease with increasing scores and are always forced to a final, maximum value of one. Even when a selection strategy selects a loglinear model that does not approximate the population frequency distribution as well as other models, the cumulative probabilities based on that model may fit the population cumulative probabilities very closely. Thus, the accuracy of the equating function that is based on the cumulative probabilities does not necessarily suffer from the inadequate loglinear model.

Another implication for how loglinear model selection influences equating function accuracy is the complexity of the equating function (i.e., the extent of difference in the distributions involved in the equating). When the score distributions were sampled from populations that did not differ, so that the identity equating function was appropriate, the entire complexity of the distributions did not need to be modeled in order to accurately produce the identity function. Only when the distributions and equating functions differed in complicated ways having to do with their shapes were the more complicated loglinear models needed. For situations where distribution differences are small and equipercentile equating is not needed, the equating function can be produced accurately from very simple loglinear models.

Finally, this study evaluated equating accuracy in terms of WAD, one of many possible indices with which equating accuracy could have been evaluated. The WAD measure was consistent with many of the loglinear model fit statistics in terms of weighting misfit based on where most of the population data were. Other accuracy indices could be of interest, especially those that give more weight to the misfit of equated scores at specific parts of a score range. These other indices are somewhat inconsistent with the focus of the fit of entire distributions that is the basis of many loglinear model fit statistics, and so evaluations of equating accuracy based on these alternative indices could produce results that differ from those reported in this study. Focusing on accuracy at specific score regions that may not necessarily be where most of the data are could be of interest to testing programs that pay great attention to the minimum or maximum scale scores, or to passing rates at particular cut scores.

Implications and Extensions

All of the selection strategies could be studied more thoroughly to specify the situations in which they function best. A broad guideline from the results of this study is that many of the selection strategies require sample sizes of at least 1,000 for selecting accurate loglinear models of univariate distributions. This guideline could be the focus of some extensions that study the strategies across wide ranges of sample sizes; considered models (e.g., *I's* from 2-–10 vs. *I's* from 3–8); and, for the significance tests, Type I error levels. Specific recommendations could be developed to define effective use for each of the selection strategies with respect to selection accuracy and equating function accuracy.

This study also could be extended to consider the use of other proposed statistics for selecting loglinear models. This study's results are broad enough to comment on some alternative measures. Bozdogan (1987) introduced a consistent AIC with Fisher information (CAICF) statistic along with his CAIC statistic. The CAICF is designed to select fewer parameters than the CAIC and therefore would produce equating functions that would not be as accurate as those from the strategies considered in this study. Gilula and Haberman (1994) introduced a modification to the AIC statistic that is theoretically appropriate for selecting among incorrect models. Preliminary investigations of Gilula and Haberman's statistic for this study showed that its performance is almost indistinguishable from that of the AIC. Bootstrapped versions of the goodness-of-fit statistics considered in this study have been developed and studied under sparse data situations that arise with item-level response data (von Davier, 1997), and the use of bootstrappng could address accuracy problems when modeling small-sample test score distributions.

A promising, alternative pursuit to additional comparisons of alternative goodness-of-fit statistics could be the development of a new class of measures that directly connect loglinear model fit to equating function accuracy. The development and evaluation of fit statistics for cumulative densities along the lines of the Kolmogorov-Smirnov (Smirnov, 1948) statistic and for inverse cumulative densities (i.e., equated scores) would avoid some of the difficulties of relating the fit of frequencies to the implications on equated scores. These alternative fit statistics would be especially useful for relating loglinear models to more complicated equating methods, such as the chained and poststratification or frequency estimation methods.

An extension to this study's focus on overall moments could include subset moments and indicator functions. The use of subset moments has been described in previous works (Holland & Thayer, 2000; von Davier et al., 2004) for modeling aspects of distributions that are known to cause systematic structures not attributable to sampling variability. The application of subset moments to modeling specific score regions or to abnormally large residuals may enhance equating function accuracy in specific situations. The small equating inaccuracy is produced from this study's always-fit-the-population models imply that data-driven applications of subset moments could improve model selection above what overall moments are able to accomplish for distributions that have complicated structures. In particular, combinations of strategies that first select parameters for the overall distribution and then try to improve model fit at specific regions or at scores where residuals are very large may be promising. Another potential strategy for reducing the influence of large residuals is to use a weighted average of raw frequencies and the smoothed frequencies in equating.

A final extension of this study would be the consideration of selection strategies to bivariate problems. Data sparseness in bivariate frequency tables typically makes chi-square significance testing based on the fits of one model unfeasible, because chi-square statistics are smaller than the degrees of freedom, even for models that do not fit the data well. Bivariate situations are likely to differentiate chi-square statistics more than the univariate situations considered in this study, as some chi-square statistics are known to respond differently than others in conditions of extreme data sparseness (e.g., the likelihood ratio and Pearson chi-square statistics in Holland & Thayer, 2000, p. 174). Suggestions for modeling bivariate distributions are to work from the outside in (Holland & Thayer), in which case the results of this study suggest that using an AIC minimization strategy for univariate distributions is an especially effective start to modeling bivariate distributions.

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.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis.* Cambridge, MA: MIT Press.

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

Fienberg, S. E. (1979). The use of chi-squared statistics for categorical data problems. *Journal of the Royal Statistical Society, Series B, 41*(1), 54–64.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of *p*. *Journal of the Royal Statistical Society, 85,* 87–94.

Gilula, Z., & Haberman, S. J. (1994). Conditional log-linear models for analyzing categorical panel data. *Annals of Mathematical Statistics, 21,* 607–611.

Haberman, S. J. (1974a). *The analysis of frequency data.* Chicago: University of Chicago Press.

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

Haberman, S. J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association, 83*(402), 555–560.

Hanson, B. A. (1990). *An investigation of methods for improving estimation of test score distributions* (Research Rep. No. 90-4). Iowa City, IA: American College Testing.

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., & Feinstein, Z. S. (1995, April). *A polynomial loglinear model for assessing differential item functioning.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED384629)

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.

Koehler, K. J., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association, 75,* 336–344.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28*(3), 257–282.

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

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

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.

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

Smirnov, N. V. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics, 19,* 279–281.

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

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research Online 2*(2). Retrieved from http://www.pabst-publishers.de/mpr/