Research Report

Pseudo Bayes Estimates for Test Score Distributions and Chained Equipercentile Equating

Tim Moses and Hyeonjoo J. Oh

December 2009

ETS RR-09-47

Pseudo Bayes Estimates for Test Score Distributions and Chained Equipercentile Equating

Tim Moses and Hyeonjoo J. Oh
ETS, Princeton, New Jersey

December 2009

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

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

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

Abstract

Pseudo Bayes probability estimates are weighted averages of raw and modeled probabilities; these estimates have been studied primarily in nonpsychometric contexts. The purpose of this study was to evaluate pseudo Bayes probability estimates as applied to the estimation of psychometric test score distributions and chained equipercentile equating functions. Population test score distributions were created from actual test data and random samples of varied size were drawn from the populations. Pseudo Bayes estimation was applied to the random samples, using ranges of loglinear models and weights in the pseudo Bayes estimates' weighted averages of the raw and modeled test score probabilities. Equipercentile equating functions based on the pseudo Bayes estimates were also evaluated. Results indicated that the pseudo Bayes estimates have the potential to improve estimation accuracy for test score distributions and chained equipercentile equating functions in situations where loglinear modeling is not ideal and where finding the population loglinear model selection is not likely.

Key words: equipercentile equating, loglinear smoothing, pseudo Bayes

Acknowledgments

The authors thank Dan Eignor, Skip Livingston, and Insu Paek for helpful reviews and Kim Fryer for the editorial work.

Table of Contents

List of Tables

Page

List of Figures

Page

Weighted averages of raw and modeled probabilities (i.e., pseudo Bayes probability estimates) have been rigorously studied and their use has been encouraged for the estimation of the cells of multiway frequency tables (Agresti, 1990; Bishop, Fienberg, & Holland, 1975; Fienberg & Holland, 1973). This work usually shows that pseudo Bayes probabilities are better estimates of population probabilities than raw probabilities. The pseudo Bayes probabilities may also be preferable to modeled probabilities, as in situations where model selection is difficult due to complex structures in the data. Presumably, pseudo Bayes probability estimates would be useful for the frequency tables that are commonly dealt with in psychometric contexts, such as the estimation of test score distributions and equipercentile equating functions. To date, the use of pseudo Bayes probability estimation for psychometric applications has been considered only in limited terms (described below).

The purpose of this study is to assess the viability of pseudo Bayes probability estimates for test score distributions and the chained equipercentile equating functions computed from these test score distributions. First, the frequency tables of test score distributions are described along with the approaches commonly used to estimate these frequency tables. Next, the proposals of pseudo Bayes probability estimates are described in terms of their potential to broaden and possibly improve on the more commonly used psychometric approaches. Finally, applications of pseudo Bayes methods to the estimation of test score distributions and chained equipercentile equating functions are evaluated in several simulations. Recommendations for practice are made based on the simulation results.

Psychometric Test Score Distributions and Common Estimation Approaches

One- and two-way frequency tables naturally arise with test score distributions. For example, the univariate distribution of a single test, X, with scores ranging from $x_j = x_1$ to x_J is contained in a one-way table of the raw frequencies, n_j. The bivariate distribution of two tests, X and A, with possible scores x_j and a_k is contained in a two-way table of the raw frequencies, n_{jk}. Several psychometric analyses work directly with the test scores' raw probabilities, $r_j = \frac{n_j}{N}$ and $r_{jk} = \frac{n_{jk}}{N}$. Testing programs may report the percentiles or norms corresponding to a test's scores (Kolen, 1991). Testing programs may also solve for test

scores with percentiles that match those of other test forms' scores, a process known as equipercentile equating (Kolen & Brennan, 2004). One difficulty encountered in the estimation of percentiles and equipercentile equating functions is that test scores' raw probabilities, r_j and r_{jk}, can exhibit considerable sampling instability. A related difficulty is that many of the raw sample probabilities can be zero when their population counterparts, p_j and p_{jk}, are assumed to be greater than zero.

Two common psychometric practices for addressing r_j and r_{jk} values that are unstable and/or implausibly zero are (a) averaging small constants with all test scores' raw probabilities (Hanson, 1990; Kolen & Brennan, 2004), or (b) employing a loglinear modeling strategy by modeling the log of r_j or r_{jk} as a polynomial function of the test scores (see this paper's Method section and Holland & Thayer, 1987, 2000). Both the small constants and the loglinear modeling strategies can produce versions of r_j and r_{jk}, s_j and s_{jk}, that are smooth and stable with values that are always greater than zero. However, both strategies can also produce biased estimates of the population values, p_j and p_{jk}. The use of small constants implies uniform and independent distributions which are usually not realistic for psychometric data. Loglinear models allow for more flexibility than the small constants strategy, in that they allow for possible parameterizations that range from very simple (e.g., uniform and independence models) to highly parameterized models that capture complex structures observed in test data (Hanson, 1996; Holland & Thayer, 2000). As with the use of small constants, though, loglinear models can produce biased estimates due to the complexities and inaccuracies of selection processes for the models' parameterizations (Agresti, 1990; Bishop et al., 1975; Hanson, 1990; Holland & Thayer, 2000; Moses & Holland, 2008, 2009b; von Davier, Holland, & Thayer, 2004).

Pseudo Bayes Estimates for Multiway Frequency Tables

Pseudo Bayes probability estimates may be useful for the estimation of frequency tables encountered with psychometric test data. Pseudo Bayes estimates can incorporate the smooth and stable features of the small constants and loglinear modeling strategies while reducing these strategies' potential for biased estimation. Pseudo Bayes estimates are weighted averages of raw (r) and modeled (s) probability estimates,

$$PB = wr + (1-w)s, \quad 0 \le w \le 1. \tag{1}$$

The Bayesian interpretation of the pseudo Bayes estimates in Equation 1 results from making standard multinomial assumptions for *r* and other assumptions for the prior distribution of population probability *p*, so that Equation 1 is said to estimate the posterior mean of *p* given *n* (Agresti, 1990; Bishop et al., 1975; Fienberg & Holland, 1973). The *pseudo* aspect of the pseudo Bayes estimates pertains to data-dependent choices of *w* and s in Equation 1. Previous work has shown that the *w* that minimizes mean squared error between the pseudo Bayes estimates and the *p*'s is

$$w = \frac{N}{N + \left(\frac{1 - \sum r^2}{\sum (r - s)^2} \right)} \tag{2}$$

(Fienberg & Holland). Analytic and simulation research has shown that the pseudo Bayes estimates can be better estimates of population probabilities than raw probabilities and modeled probabilities (Agresti; Bishop et al.; Fienberg & Holland).

This Study

The purpose of this study is to evaluate applications of pseudo Bayes estimates for the estimation of test score distributions and equipercentile equating functions. These applications extend previous considerations of pseudo Bayes estimation and of the smoothing methods used in psychometric practice. In earlier pseudo Bayes proposals (Agresti, 1990; Bishop et al., 1975; Fienberg & Holland, 1973), population probabilities have been primarily estimated by using s values from uniform and independence models when creating the pseudo Bayes estimates. These s values are probably unrealistic for the one- and two-way tables encountered in test data, where the distributions can be complex and joint distributions cannot be assumed to be uniformly and independently distributed. The current study expands on early pseudo Bayes studies by considering how s values from loglinear models that fit the observed data to varying degrees affect the accuracy of the pseudo Bayes estimates.

In terms of psychometric practice and equipercentile equating, the choices considered for test score distribution estimation have primarily included (a) the raw probabilities, (b) the raw probabilities averaged with small constants, and (c) the smoothed probabilities based on some

loglinear model (Hanson, 1990; Hanson, Zeng, & Colton, 1994; Moses & Holland, 2007). When these choices are interpreted as relatively limited applications of pseudo Bayes estimates (i.e., Equation 1 with $w = 0$, and s following a uniform distribution, or with $w = 1$), it becomes apparent that these choices are potentially improvable though the use of different w values. For example, a study by Moses and Holland (2009a) showed that pseudo Bayes estimates with w values of 0.5 improved the accuracy of equipercentile equating functions. This study broadens the Moses and Holland investigation and prior smoothing investigations by considering pseudo Bayes estimates with w values ranging between 0 and 1 along with a range of loglinear models for s.

Method

Pseudo Bayes estimates for test score distributions were studied in several simulations. Population distributions were obtained by fitting loglinear models to four large sample psychometric test score distributions. Random samples were then drawn from each of these population distributions and the pseudo Bayes estimation described in Equation 1 was used to estimate the population distributions from each random sample. Equipercentile equating functions were computed in the random samples based on the test scores probabilities produced from the pseudo Bayes estimates. Different applications of pseudo Bayes estimation were considered by varying the accuracy of the loglinear model used for s and also varying the choice of w. The accuracies of the probability estimates and the equipercentile equating functions based on the pseudo Bayes applications were assessed by comparing the sample estimates from each pseudo Bayes application to the corresponding values in the population distribution. The conditions of the simulation are described in more detail below.

Four Population Test Score Distributions and Their Population Loglinear Models

Four bivariate test score distributions were obtained from the operational data of a large-scale testing program. These bivariate distributions comprised two nonequivalent groups with anchor test equating situations where in each situation a new test form (X) was to be equated to an old test form (Y) and an anchor test (A) internal to X and Y was used to account for the nonequivalence of the examinee groups taking X and Y. The (X,A) and (Y,A) bivariate distributions of the first equating situation are summarized in Table 1. The (X,A) and (Y,A) bivariate distributions of the second equating situation are

summarized in Table 2. The tests and anchors were both scored as rounded formula scores, meaning that each test and anchor score was computed by subtracting a portion of examinees' incorrect responses from their total number of correct responses and then rounded. The rounded formula-scoring produced distributions with complex structures (described below).

Table 1

Descriptive Statistics for Two of the Bivariate (Test, Anchor) Distributions Comprising the First Equating Situation

	New form population		Old form population	
	X	*A*	*Y*	*A*
Test score range	-10 to 39	-7 to 26	-12 to 49	-7 to 26
Mean	13.633	8.382	24.402	12.438
Standard deviation	9.622	6.530	10.163	-0.088
Skewness	0.311	0.370	-0.533	-0.602
Kurtosis	-0.672	-0.544		
Correlation	0.968		0.955	
Original sample sizes	440,102		198,094	

Table 2

Descriptive Statistics for Two of the Bivariate (Test, Anchor) Distributions Comprising the Second Equating Situation

	New form population		Old form population	
	X	*A*	*Y*	*A*
Test score range	-10 to 39	-7 to 26	-12 to 47	-7 to 26
Mean	18.838	11.983	23.747	13.039
Standard deviation	8.908	5.847	9.850	5.703
Skewness	0.044	0.110	-0.027	-0.016
Kurtosis	-0.697	-0.583	-0.617	-0.582
Correlation	0.964		0.953	
Original sample sizes	169,333		193,605	

Loglinear models that closely fit the raw bivariate distributions were used as population distributions for the study. The loglinear models allowed for the study of the pseudo Bayes estimation with respect to s values based on the "true" population model and also with respect to varying degrees of inadequacy (to be described). The close fits of the loglinear models to the original datasets meant that the estimation results from the simulation would be applicable to realistic test score distributions and equating situations. The parameterization of the loglinear model used as the bivariate (test, anchor) population distributions was the same for both of the (*X*,*A*) distributions and for both of the (*Y*,*A*) distributions whose summary statistics are shown in Tables 1 and 2. For an (*X*,*A*) bivariate distribution, this loglinear model can be expressed as

$$\log_e(s_{jk}) = \sum_{d=0}^{D=6}\beta_{x,d}x_j^d + \sum_{e=0}^{E=6}\beta_{a,e}a_k^e + \sum_{f=0}^{F=6}\beta_{xf,f}T(j)x_j^f + \sum_{g=0}^{G=6}\beta_{ag,g}T(k)a_k^g + \sum_{h=0}^{H=2}\sum_{i=0}^{I=2}\beta_{xa,hi}x_j^h a_k^i . \quad (3)$$

In Equation 3, s_{jk} is the modeled relative frequency of examinees obtaining scores x_j and a_k, where the *D* and *E* are the numbers of parameters used to model the univariate distributions of *X* and *A* (= 6) and *H* and *I* are the numbers of parameters used to model the joint (*X*,*A*) distribution (= 2). *D* and *E* values of 6 produce a modeled distribution where the first six moments of *X* and *A* match those of the raw distribution. *H* and *I* values of 2 result in a modeled distribution where the conditional means and variances of *X* given *A* and of *A* given *X* match those of the raw distribution.

The $\sum_{f=0}^{F=6}\beta_{xf,f}T(j)x_j^f$ and $\sum_{g=0}^{G=6}\beta_{ag,g}T(k)a_k^g$ terms in Equation 3 model score-specific structures in the marginal distributions of *X* and *A* that are due to the rounded formula-scoring of *X* and *A*. Specifically, when *X* and *A* are formula-scored such that portions of incorrect answers on *X* and *A* are subtracted from the totals of correct responses, the marginal distributions of *X* and *A* contain abnormally low frequencies that occur at fixed score intervals. The $T(j)$ and $T(k)$ terms in Equation 3 are indicator functions set to 1 for the *X* and *A* scores with abnormally low frequencies and set to 0

otherwise. The result of using product terms $T(j)x_j^f$ and $T(k)a_k^g$ is that the total sample size and the first 6 moments of the distributions of abnormally-low frequencies of *X* and *A* in the modeled distribution will match those of the raw distribution.

Figures 1–5 plot the marginal and bivariate population distributions for the modeled (*X*,*A*) distribution used in the first equating example. The marginal distributions of *X* and *A* from this bivariate distribution are shown in Figures 1 and 2. To illustrate the joint (*X*,*A*) distribution, the conditional means, standard deviations and skews of *X* given *A* are plotted in Figures 3–5. These conditional distributions illustrate the complexity of bivariate data typically encountered in psychometric testing, where the conditional means of *X* increase nonlinearly with *A* (Figure 3) and where the conditional standard deviations and skews *X* decrease nonlinearly with *A* (Figures 4–5). Figures 1–5 are representative of the other bivariate distributions, (*Y*,*A*) for the first equating example, and (*X*,*A*) and (*Y*,*A*) for the second equating example.

Sample Sizes

One thousand sample datasets were randomly drawn from each of the four population distributions, the (*X*,*A*) and (*Y*,*A*) bivariate distributions for the first and second equating examples. Five hundred datasets were drawn for a sample size of 1,000 and five hundred datasets were drawn for a sample size of 10,000.

Applications of Pseudo Bayes Estimates

The pseudo Bayes estimation in Equation 1 was used to estimate the population distributions from each of the randomly-drawn sample datasets. Different applications of Equation 1 were considered based on varying both the loglinear model (s) and the weight (w) given to *r*. The s and w values were crossed, such that every considered loglinear model for s was paired with every considered weight (w).

Loglinear models for the test score distribution(s). Four bivariate loglinear models based on Equation 3 were considered for the s values in the pseudo Bayes estimation Equation 1. (These four models are defined in increasing complexity and summarized in Table 3.)

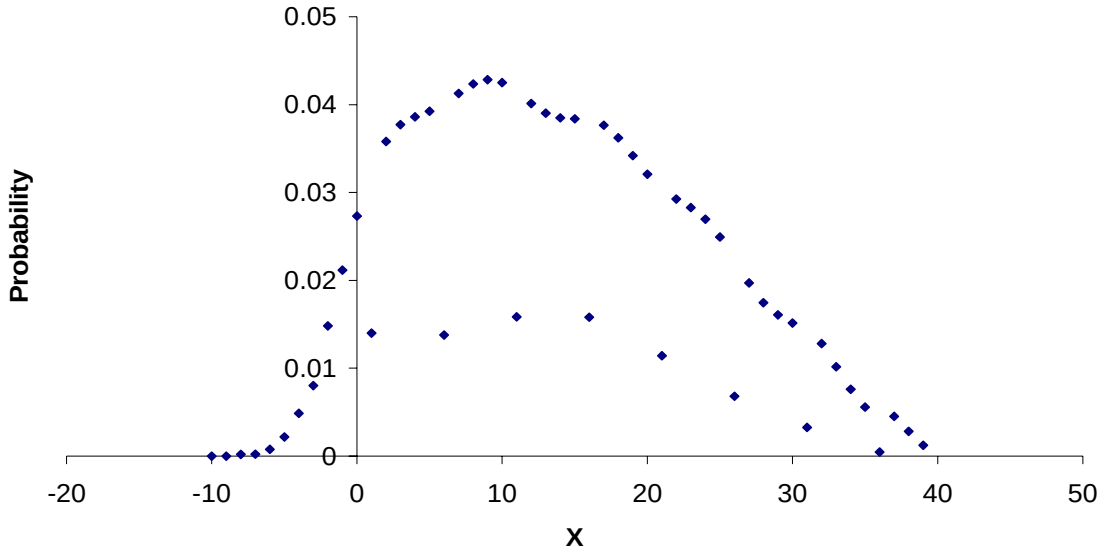

Figure 1. **Population distribution of *X*: First equating situation.**

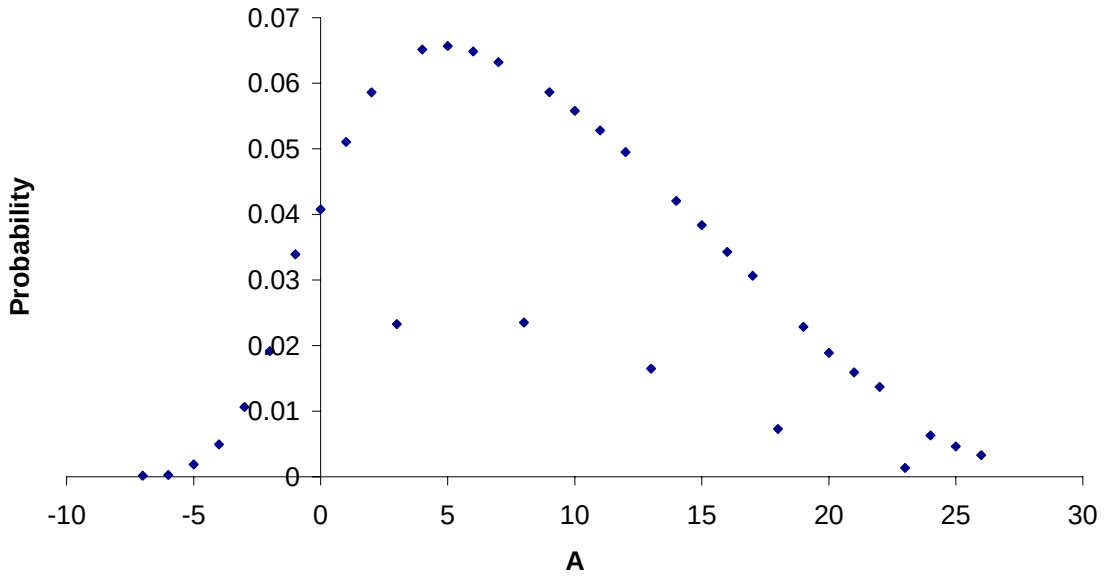

Figure 2. **Population distribution of A: First equating situation.**

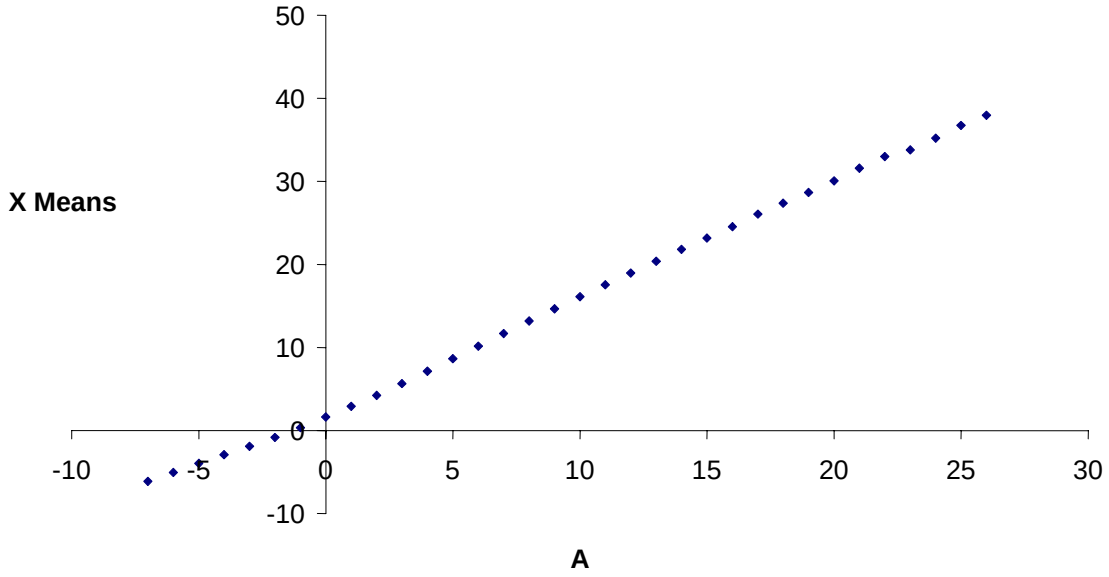

***Figure 3*. Population means of *X* given *A*: First equating situation.**

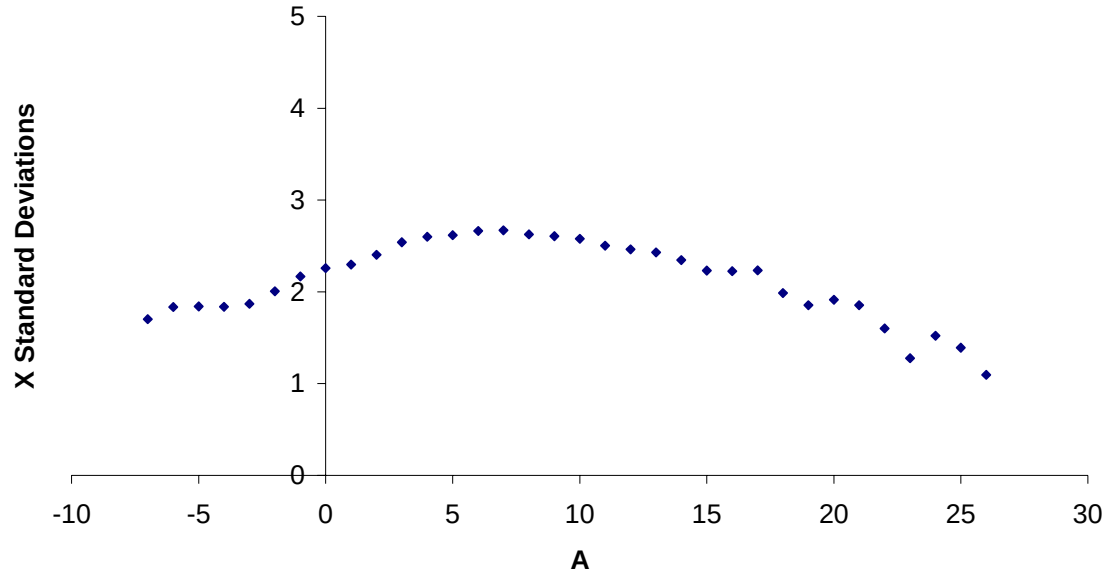

***Figure 4*. Population standard deviation of *X* given *A*: First equating situation.**

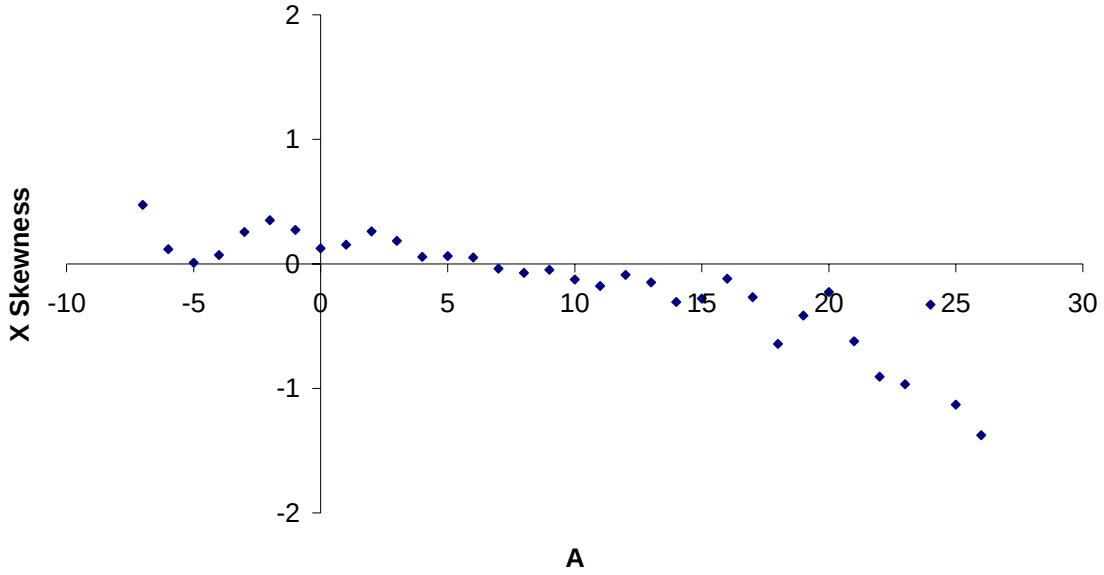

***Figure 5*. Population skewness of *X* given *A*: First equating situation.**

Table 3

Summary of the Four Loglinear Models Used for s in the Pseudo Bayes Estimates

	Marginal moments preserved (*D* & *E*)	Fit the score-specific structures (*F* & *G* = 6)?	Conditional moments preserved (*H* & *I*)
Model 1	0	No	0
Model 2	3	No	1
Model 3	6	No	1
Model 4[a]	6	Yes	2

[a] Population model.

- *Model 1*: A revised Equation 3 with $D = E = F = G = H = I = 0$, the loglinear model that corresponds to the uniform and independence distributions used in the early proposals for pseudo Bayes estimation (Bishop et al., 1975; Fienberg & Holland, 1973) and the psychometric practice of averaging data with small constants (Hanson, 1990; Kolen & Brennan, 2004).
- *Model 2*: A revised Equation 3 with $D = E = 3$, $F = G = 0$ and $H = I = 1$, the loglinear model prior research has suggested contains the minimal univariate and bivariate parameterizations to produce acceptably accurate equipercentile equating functions (Moses & Holland, 2007, 2008).
- *Model 3*: A revised Equation 3 with $D = E = 6$, $F = G = 0$ and $H = I = 1$, the loglinear model used operationally by the testing program providing the data.
- *Model 4*: The actual Equation 3 with $D = E = F = G = 6$ and $H = I = 2$, the population loglinear model.

Weights for the raw probabilities (w). The following w values were considered for the pseudo Bayes estimates from each random sample fit with each considered loglinear model, $w =$ 1, 0.75, 0.50, 0.25, and 0. The data-adaptive w value suggested in Fienberg and Holland (1973) and shown in Equation 2 was also considered, so there were six w values in total that were considered.

Simulation

To review the simulation, 500 random samples of size 1,000 and 10,000 were drawn from each of the four bivariate population distributions. The population probabilities were estimated in each random sample by using pseudo Bayes estimation Equation 1 with a range of loglinear models for s and a range of weights (w). Equipercentile equating of samples' new test to the old test was also conducted across the new (X,A) and old (Y,A) bivariate distribution samples from the two equation situations shown in Tables 1 and 2, based on the pseudo Bayes probability estimates. The accuracies of the probability estimates and equipercentile equating functions were assessed by averaging the squared deviations of the sample estimates from the population values for all of the random samples drawn in the simulation.

Accuracy measure for probability estimates. The extent to which the pseudo Bayes estimates approximated the population distributions was evaluated in terms of root mean squared error,

$$RMSE_{Distribution,N,w,s} = \sqrt{\sum p_{Distribution} \frac{\sum_{v=1}^{V=500} \left(PB_{Distribution,N,w,s,v} - p_{Distribution} \right)^2}{500}} . \quad (4)$$

The subscripts of the $RMSE_{Distribution,N,w,s}$ in Equation 4 indicate that this criterion was computed for each of the four bivariate population distributions (*Distribution*), for each of the two sample sizes (*N*), and for each combination of w value and loglinear model (s). In Equation 4, the inaccuracies of the pseudo Bayes estimates of each of the *p*'s was summarized as the sum of the squared deviations of all *V* = 500 samples' pseudo Bayes estimates from *p* ($\sum_{v=1}^{V=500} \left(PB_{Distribution,N,w,s,v} - p_{Distribution} \right)^2$). To obtain a single summary measure for the entire bivariate distribution, the squared deviations for each probability were averaged with respect to *p* ($\sum p_{Distribution}$ ….) and a final square root was taken to produce a measure that was on the scale of the *p* values.

Accuracy measure for equipercentile equating. The extent to which equipercentile equating functions based on the pseudo Bayes estimates approximated the first (Table 1) and second (Table 2) population equating functions computed from the population (*X*,*A*) and (Y,A) bivariate distributions was evaluated in terms of root mean squared error,

$$RMSE(x)_{Equating,N,w,s} = \sqrt{\frac{\sum_{v=1}^{V=500} \left(\phi(x)_{Equating,N,w,s,v} - \phi(x)_{Population\ Equating} \right)^2}{500}} . \quad (5)$$

In Equation 5, the subscripts denote one of two equating situations (*Equating*), one of two sample sizes (*N*), and the pseudo Bayes' combination of *w* value (*w*) and smoothing model (*s*). The $\phi(x)$ values denote *X*-to-*Y* chained equipercentile equating functions at *X* score = *x*, as in equipercentile equating functions from the *X* scores to the *Y* scores across

nonequivalent examinee populations using the *A* scores in the samples to adjust for the nonequivalence of the examinee samples (Kolen & Brennan, 2004). Equation 5 shows that equating inaccuracy was summarized as the sum of the squared deviations of all $V = 500$ samples' pseudo Bayes chained equating functions from the population chained equating function

$\phi(x)_{PopulationEquating}$ ($\sum_{v=1}^{V=500} \left(\phi(x)_{Equating,N,w,s,v} - \phi(x)_{Population\ Equating}\right)^2$). A final square root was taken to produce a measure that was on the scale of the $\phi(x)$ values.

A single summary measure of Equation 5 was obtained by averaging the squared deviations of the equated scores with respect to the marginal distribution of *X* in the population ($\sum p(x)_{Distribution}$ ….). A final square root was taken to produce a measure that was on the scale of the $\phi(x)$ values

$$RMSE_{Equating,N,w,s} = \sqrt{\sum p(x)_{Distribution} RMSE(x)^2_{Equating,N,w,s}} . \quad (6)$$

Results

Probability Estimation Results

The probability estimation results are presented in Tables 4–7. Each table shows the root mean squared error (*RMSE)* accuracy values for estimating one of the four bivariate population distributions at the two considered sample sizes (1,000 and 10,000), using pseudo Bayes estimates based on the six considered w values (1, 0.75, 0.50, 0.25, 0 and Equation 2's data-adaptive w value), and the four considered loglinear models. The implications of sample size, loglinear model and w value can be observed on the accuracy (*RMSE*) of the pseudo Bayes estimates for all four population distributions.

Both sample size and the loglinear models had strong overall effects on the *RMSE* values in Tables 4–7. Probability estimation was most accurate (smallest *RMSE* values) for pseudo Bayes estimates based on sample sizes of 10,000 and the population loglinear model (Model 4). Probability estimation was least accurate (largest *RMSE* values) for pseudo Bayes estimates based on for sample sizes of 1,000 and the simplest loglinear model (Model 1).

Table 4

Root Mean Squared Errors (RMSEs) for Estimating the Bivariate (X,A) Distribution (500 Replications): First Equating Situation

Sample size	Loglinear model	$w = 1$	$w = 0.75$	$w = 0.50$	$w = 0.25$	$w = 0$	Data-adaptive w
1,000	Model 1	0.0024	0.0022	0.0029	0.0039	0.0052	0.0022
	Model 2	0.0024	0.0019	0.0016	0.0016	0.0020	0.0017
	Model 3	0.0024	0.0019	0.0016	0.0016	0.0019	0.0017
	Model 4[a]	0.0024	0.0018	0.0013	0.0008	0.0006	0.0012
10,000	Model 1	0.0008	0.0014	0.0026	0.0039	0.0052	0.0007
	Model 2	0.0008	0.0008	0.0011	0.0015	0.0020	0.0007
	Model 3	0.0008	0.0007	0.0010	0.0014	0.0019	0.0007
	Model 4[a]	0.0008	0.0006	0.0004	0.0003	0.0002	0.0004

[a] Population model.

Table 5

Root Mean Squared Errors (RMSEs) for Estimating the Bivariate (Y,A) Distribution (500 Replications): First Equating Situation

Sample size	Loglinear model	$w = 1$	$w = 0.75$	$w = 0.50$	$w = 0.25$	$w = 0$	Data-adaptive w
1,000	Model 1	0.0021	0.0020	0.0025	0.0034	0.0045	0.0019
	Model 2	0.0021	0.0016	0.0013	0.0012	0.0014	0.0014
	Model 3	0.0021	0.0016	0.0013	0.0012	0.0014	0.0014
	Model 4[a]	0.0021	0.0016	0.0011	0.0007	0.0005	0.0011
10,000	Model 1	0.0007	0.0012	0.0023	0.0034	0.0045	0.0007
	Model 2	0.0007	0.0006	0.0008	0.0011	0.0014	0.0006
	Model 3	0.0007	0.0006	0.0008	0.0010	0.0014	0.0006
	Model 4 [a]	0.0007	0.0005	0.0004	0.0002	0.0002	0.0004

[a] Population model.

Table 6

Root Mean Squared Errors (RMSEs) for Estimating the Bivariate (X,A) Distribution (500 Replications): Second Equating Situation

Sample size	Loglinear model	$w = 1$	$w = 0.75$	$w = 0.50$	$w = 0.25$	$w = 0$	Data-adaptive w
1,000	Model 1	0.0025	0.0024	0.0033	0.0046	0.0061	0.0023
	Model 2	0.0025	0.0020	0.0017	0.0018	0.0023	0.0019
	Model 3	0.0025	0.0020	0.0017	0.0018	0.0023	0.0019
	Model 4 [a]	0.0025	0.0019	0.0014	0.0009	0.0007	0.0013
10,000	Model 1	0.0008	0.0016	0.0031	0.0046	0.0061	0.0008
	Model 2	0.0008	0.0008	0.0012	0.0017	0.0022	0.0008
	Model 3	0.0008	0.0008	0.0012	0.0017	0.0022	0.0008
	Model 4 [a]	0.0008	0.0006	0.0004	0.0003	0.0002	0.0004

[a] Population model.

Table 7

Root Mean Squared Errors (RMSEs) for Estimating the Bivariate (Y,A) Distribution (500 Replications): Second Equating Situation

Sample size	Loglinear model	$w = 1$	$w = 0.75$	$w = 0.50$	$w = 0.25$	$w = 0$	Data-adaptive w
1,000	Model 1	0.0022	0.0020	0.0026	0.0035	0.0047	0.0020
	Model 2	0.0022	0.0017	0.0013	0.0012	0.0014	0.0014
	Model 3	0.0022	0.0017	0.0013	0.0012	0.0014	0.0014
	Model 4 [a]	0.0022	0.0017	0.0012	0.0007	0.0005	0.0011
10,000	Model 1	0.0007	0.0013	0.0024	0.0035	0.0047	0.0007
	Model 2	0.0007	0.0006	0.0008	0.0011	0.0014	0.0006
	Model 3	0.0007	0.0006	0.0008	0.0010	0.0013	0.0006
	Model 4 [a]	0.0007	0.0005	0.0004	0.0002	0.0002	0.0004

[a] Population model.

The effect of the value of w on the accuracy (*RMSE*) of the pseudo Bayes estimates can be understood in terms of the interaction of w with the loglinear models and sample sizes. For Model 1, large w values produced more accurate pseudo Bayes estimates. In contrast, when Model 4 was used, small w values produced more accurate pseudo Bayes estimates. The most accurate probability estimation was achieved when using Model 4 with w values of 0. For loglinear models other than Model 4, sample size altered the impact of w values on probability estimation accuracy, making larger w values most useful for large sample sizes and smaller w values most useful for small sample sizes. For example, Tables 4–7 show that with the relatively poor-fitting Model 2, w values of 1 and 0.75 produced the most accurate probability estimates for sample sizes of 10,000 while w values of 0.50 and 0.25 produced the most accurate probability estimates for sample sizes of 1,000.

The rightmost columns of Tables 4–7 allow for the evaluation of data-adaptive w values Equation 2 on probability estimation accuracy. For overly simple loglinear models (i.e., Models 1 and 2) the data-adaptive w values produce relatively accurate probability estimates compared to other *w* values. For Model 4 the data-adaptive w values produce probability estimates that were not as accurate as those based on w values of 0.

Overall Chained Equipercentile Estimation Results (RMSEs)

The chained equipercentile estimation results for the first and second equating situations are summarized in Tables 8–9. Within each table, the *RMSE* accuracy values are shown for the two considered sample sizes (1,000 and 10,000), the six considered w values (1, 0.75, 0.50, 0.25, 0 and Equation 2's data-adaptive w value), and the four considered loglinear models.

The equating estimation accuracy results shown in Tables 8 and 9 are similar to the probability estimation accuracy results of Tables 4–7 with respect to sample sizes and loglinear models. Equating accuracy is higher (smaller *RMSE* values) when equating is conducted with large sample sizes and with the population loglinear models (Model 4). Equating accuracy is lower (larger *RMSE* values) when equating is conducted with small sample sizes and with overly simple loglinear models (Model 1). Practically large *RMSE* values (0.5 or greater) indicate overall equating inaccuracy that is large enough to affect

Table 8

Root Mean Squared Errors (RMSEs) for Estimating the X-to-Y Chained Equipercentile Equating Function (500 Replications): First Equating Situation

Sample size	Loglinear model	$w = 1$	$w = 0.75$	$w = 0.50$	$w = 0.25$	$w = 0$	Data-adaptive w
1,000	Model 1	0.7394	0.8672	0.9097	0.6954	0.4548	0.7955
	Model 2	0.7394	0.4318	0.3581	0.3222	0.3329	0.3837
	Model 3	0.7394	0.4744	0.4279	0.4069	0.4157	0.4429
	Model 4[a]	0.7394	0.4939	0.4462	0.4141	0.4037	0.4432
10,000	Model 1	0.2400	0.8185	0.8915	0.6876	0.4548	0.3052
	Model 2	0.2400	0.1548	0.1621	0.1955	0.2406	0.1636
	Model 3	0.2400	0.1713	0.1751	0.2026	0.2441	0.1777
	Model 4[a]	0.2400	0.1713	0.1520	0.1392	0.1342	0.1509

[a] Population model.

Table 9

Root Mean Squared Errors (RMSEs) for Estimating the X-to-Y Chained Equipercentile Equating Function (500 Replications): Second Equating Situation.

Sample size	Loglinear model	$w = 1$	$w = 0.75$	$w = 0.50$	$w = 0.25$	$w = 0$	Data-adaptive w
1,000	Model 1	0.5289	0.4163	0.5478	0.6631	0.7321	0.3856
	Model 2	0.5289	0.3389	0.2963	0.2887	0.3167	0.3101
	Model 3	0.5289	0.3628	0.3351	0.3330	0.3556	0.3429
	Model 4[a]	0.5289	0.3714	0.3329	0.3079	0.2983	0.3306
10,000	Model 1	0.1575	0.3356	0.5231	0.6582	0.7321	0.1442
	Model 2	0.1575	0.1287	0.1570	0.2020	0.2557	0.1237
	Model 3	0.1575	0.1322	0.1577	0.1988	0.2484	0.1269
	Model 4[a]	0.1575	0.1189	0.1061	0.0974	0.0940	0.1053

[a] Population model.

rounded equated scores. The practically large *RMSEs* in Tables 8 and 9 correspond to pseudo Bayes estimates using Model 1, using *w* values of 1, and applied to sample sizes of 1,000.

The effect of *w* values on equating accuracy depends on the sample size and the loglinear model. For Model 4, *w* values of 0 produce the most accurate equating functions. When the loglinear model is simpler than the population model (Models 2 and 3), *w* values in between 0 and 1 produce the most accurate equating functions, with larger *w* values being preferable for large sample sizes and smaller *w* values being preferable for small sample sizes. For the simplest loglinear model (Model 1) the effect of *w* value differed for the first and second equating situations. For the first equating situation (Table 8), the use of Model 1 with a *w* value of 0 produced relatively accurate equating functions. For the second equating situation (Table 9), the use of Model 1 with data-adaptive *w* values and *w* values of 0.75 and 1 produced relatively accurate equating functions.

The data-adaptive *w* values had influences on equating accuracy similar to their influences on probability estimation accuracy for every loglinear model except Model 1. When used with the population loglinear model (Model 4) the data-adaptive *w* values produced equating functions that were not as accurate as the equating functions based on *w* values of 0. When used with Models 2 and 3, the data-adaptive *w* values produced equating functions that were more accurate than those based on most other *w* values. For Model 1, the data-adaptive *w* value produced equating functions that were optimally accurate for the second situation (Table 9) but not as accurate as other *w* values for the first equating situation (Table 8).

Score-Level Chained Equipercentile Estimation Results (RMSE(x)'s)

Assessments of the score-level equating accuracies for some of the chained equipercentile equating functions were conducted to evaluate the extent to which *X*-to-*Y* equating accuracy varied across the *X* scores. Plots of the score-level root mean squared error (*RMSE(x)*) were used to compare various pseudo Bayes estimates. All plots used vertical axes on scales of 0 to 2 *RMSE(x)* units, a range large enough to show most of the *RMSE(x)* results and narrow enough to differentiate the more and less accurate equating

results. *RMSE(x)* values so large that they are not visible in the plots correspond to score ranges of *X* where there is expected to be essentially no data in the samples (Figure 1). *RMSE(x)* values of 0.5 or greater may be considered practically large, as these values indicate equating inaccuracy that is large enough to affect rounded equated scores.

Figures 6–9 plot the *RMSE(x)* values for the four loglinear models based on the five w values (1, 0.75, 0.50, 0.25 and 0) for the first equating situation and sample sizes of 1,000. Figure 6 plots the *RMSE(x)* values for Model 1 and the five *w* values, suggesting as in Tables 8 and 9 that there is no w value that produces the most accurate equating function across all of the *X* scores. For Models 2 and 3 (Figures 7 and 8), small w values are generally better for equating accuracy, though the w value that results in the most accurate equating results varies across the *X* scores. Figure 9 shows that for the population loglinear model (Model 4), small w values (0 and 0.25) consistently produce the most accurate equating functions and large w values (1 and 0.75) consistently produce the least accurate equating functions across most of the *X* scores.

To assess the data-adaptive w values on score-level equating accuracy, Figures 10 and 11 compare the score-level *RMSE(x)* values from using data-adaptive w values with the four loglinear models for the first and second equating situations. The results are generally consistent across the two equating situations, showing small differences in the *RMSE(x)* values based on all loglinear models except for Model 1. The *RMSE(x)* values based on Model 1 are visibly worse than those of the other models for almost all of the *X* scores. Although the *RMSE(x)* values based on Model 1 with data-adaptive w values seem to be better for the second equating situation (Figure 11) than the first (Figure 10), they are worse than those based on other loglinear models and data-adaptive w values across most of the *X* scores for both equating situations.

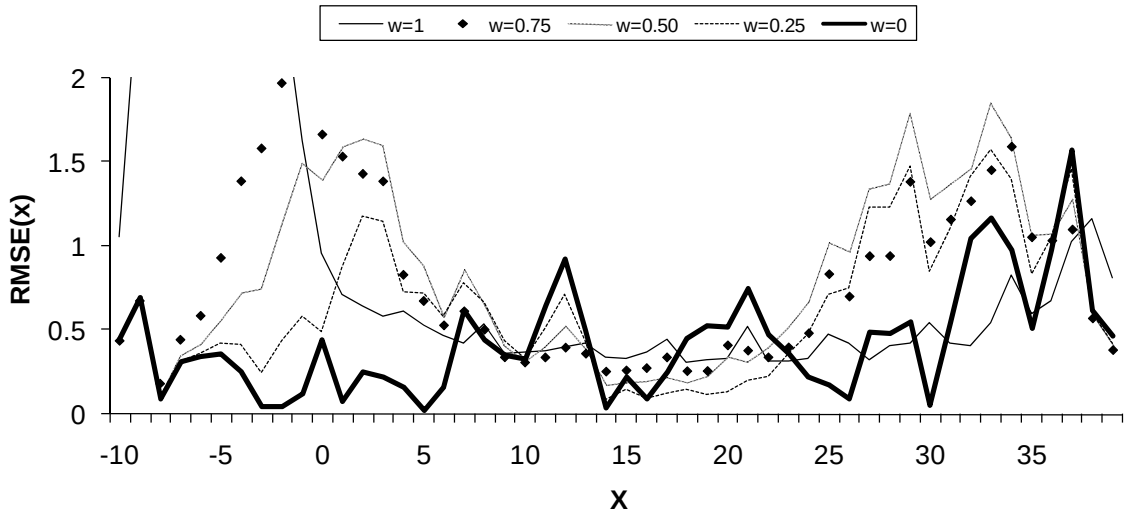

***Figure 6*. Score-level accuracies of equating (RMSE(*x*)): First equating situation. Sample sizes of 1,000. Model 1.**

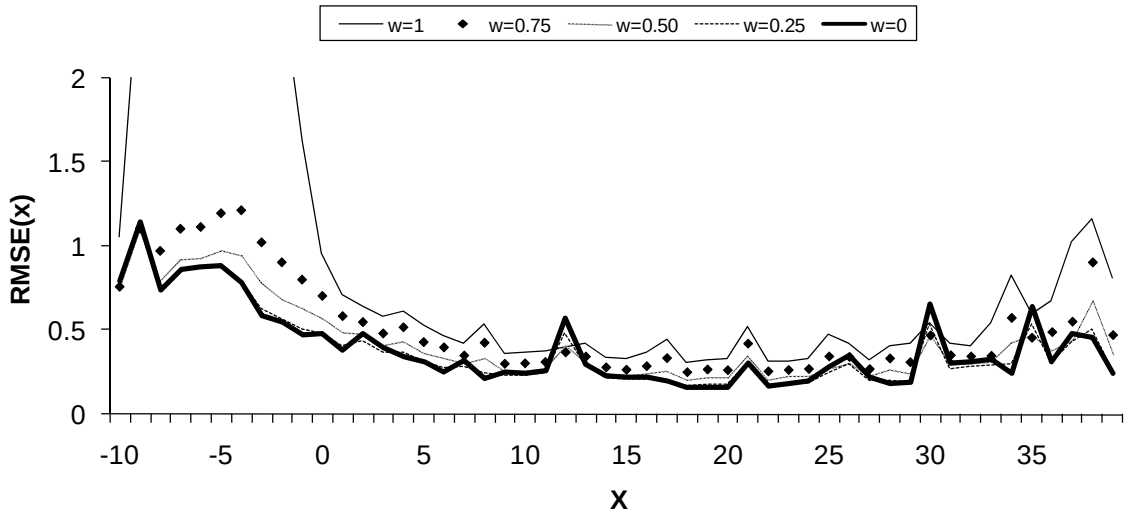

***Figure 7*. Score-level accuracies of equating (RMSE(*x*)): First equating situation. Sample sizes of 1,000. Model 2.**

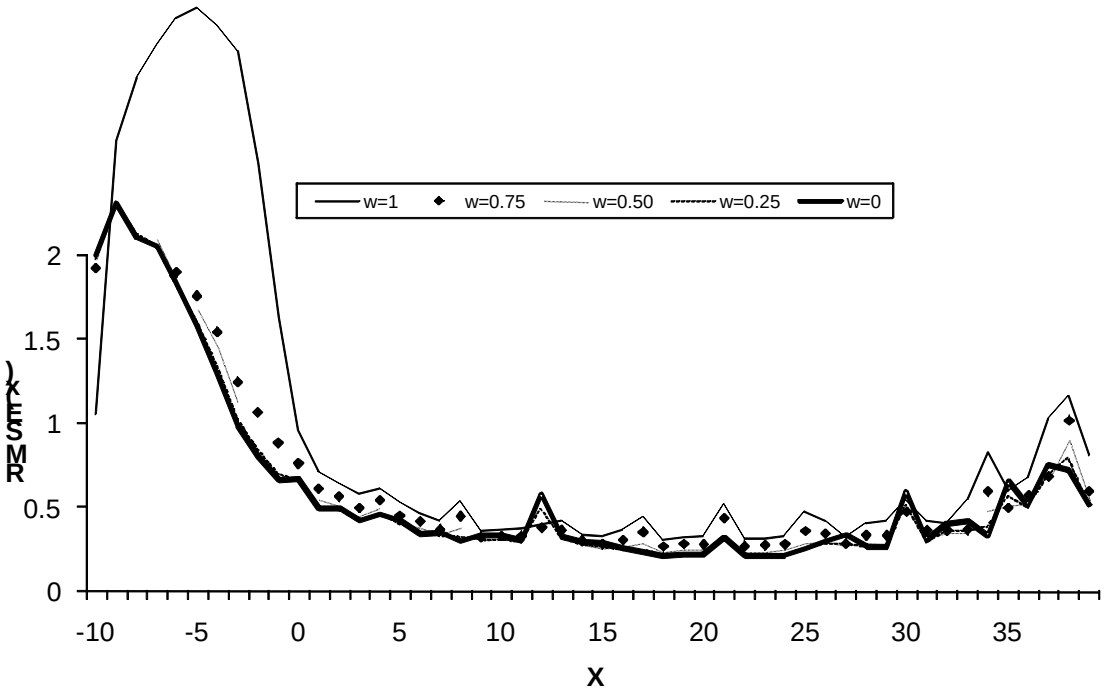

***Figure 8*. Score-level accuracies of equating (RMSE(*x*)): First equating situation. Sample sizes of 1,000. Model 3.**

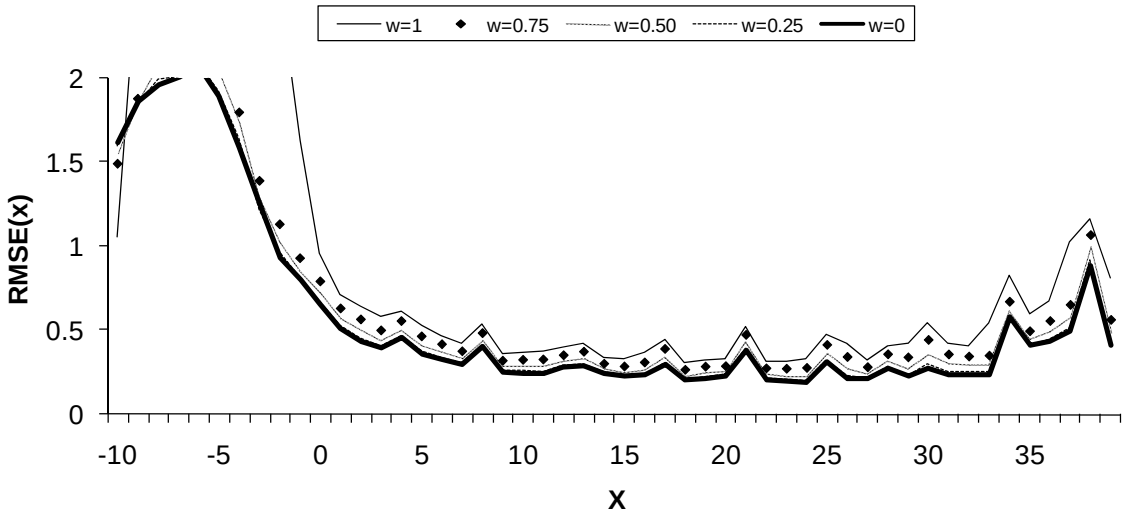

***Figure 9.* Score-level accuracies of equating (RMSE(x)): First equating situation. Sample sizes of 1,000. Model 4.**

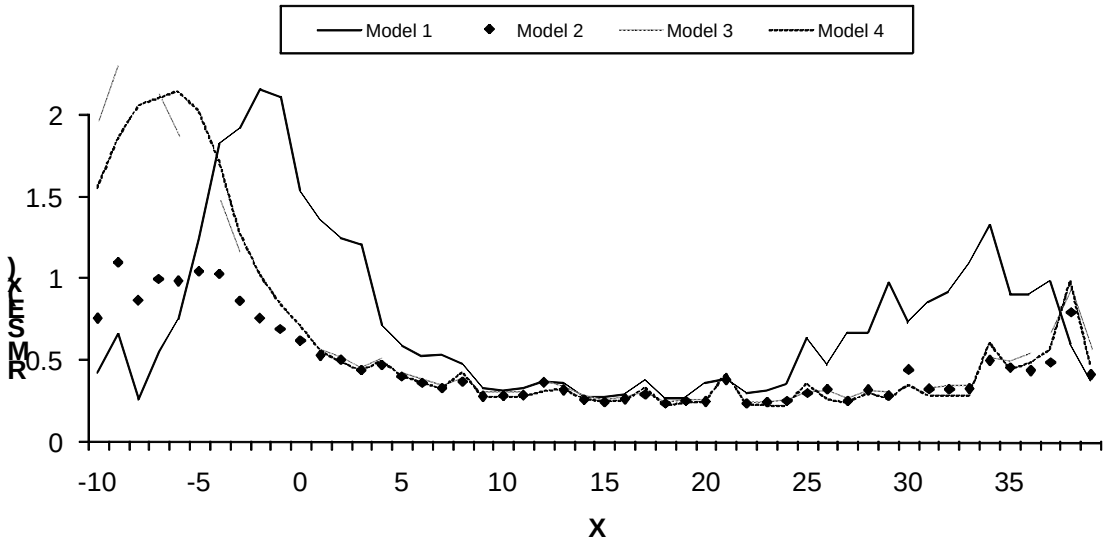

***Figure 10*. Score-level accuracies of equating (RMSE(*x*)): First equating situation. Sample sizes of 1,000. All loglinear models with data-adaptive *w* values.**

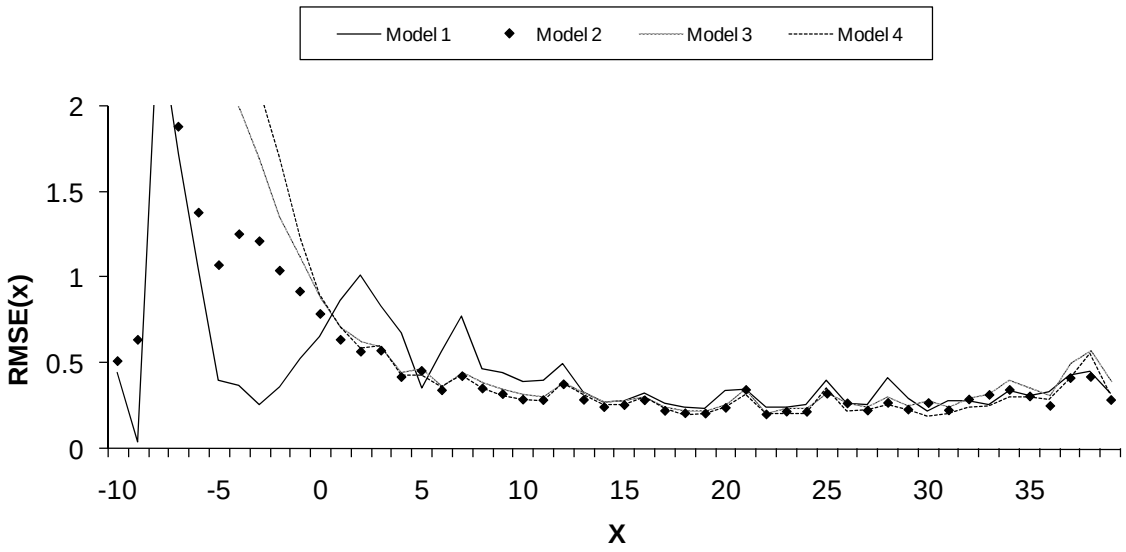

***Figure 11*. Score-level accuracies of equating (RMSE(*x*)): Second equating situation. Sample sizes of 1,000. All loglinear models with data-adaptive *w* values.**

Discussion

Difficulties in the estimation of test score distributions and equipercentile equating functions occur when raw cell probabilities exhibit sampling fluctuation or are implausibly zero. The traditional psychometric practices for addressing these issues are to smooth the raw probabilities or to average the raw probabilities with small constants (Hanson, 1990; Hanson et al., 1994; Kolen & Brennan, 2004). Pseudo Bayes methods developed in nonpsychometric contexts have been shown to estimate population probabilities more accurately than raw and modeled probabilities (Agresti, 1990; Bishop et al., 1975; Fienberg & Holland, 1973). The accuracy potential of pseudo Bayes estimates for test score distributions and equipercentile equating functions was evaluated in this study. The implications of choices for the "pseudo" aspects of pseudo Bayes probability estimates were considered, including the implications of different weights for producing weighted averages of the raw and modeled probabilities and the implications of different loglinear models for the modeled probabilities.

The overall results of this study showed that the pseudo Bayes applications produced estimates of test score probabilities and chained equipercentile equating functions that were usually not as accurate as the corresponding estimates from the population loglinear models. This study's results suggest that the practical use of Fienberg and Holland's (1973) data-adaptive weights in the pseudo Bayes estimates is limited, in that the data-adaptive weights appeared to be optimal weights of the raw data when the loglinear model was overly simple, but gave too much weight to the raw data when the population loglinear model was used. A major implication for practice is that under ideal conditions such as large sample sizes and easily modeled test score distributions, the most preferable approach to the estimation of test score probabilities and chained equipercentile equating functions is to find and use the best possible loglinear model(s) rather than other pseudo Bayes applications based on averaging raw and modeled probabilities.

It is under less than ideal equating situations where pseudo Bayes applications have some utility. Specifically, finding the most accurate loglinear model for observed test data is not always feasible in practice. Accurate loglinear model selection can require more time than typical equating timelines accommodate and larger sample sizes than those that are encountered in testing programs' equating work. Model selection depends not only on adequate sample sizes, but also on infrastructure that can support the fit and comparison of complex loglinear models such as Equation 3. To the extent that circumstances of equating practice result in less than

perfect loglinear models, pseudo Bayes applications can improve estimation accuracy. Pseudo Bayes estimates based on data-adaptive weights and the "6-6-1" loglinear model routinely used in equating practice can produce probability estimates and chained equipercentile equating functions that are more accurate than those obtained from using only the 6-6-1 loglinear models when those 6-6-1 loglinear models are known to under-fit the observed data. The pseudo Bayes estimates based on data-adaptive weights and the 3-3-1 or 6-6-1 loglinear models have accuracies that are similar to the accuracies from using population loglinear models.

Some concluding statements are warranted for the implications of pseudo Bayes estimates based on the simplest uniform and independence models. As suggested in prior pseudo Bayes studies (Agresti, 1990; Bishop et al., 1975; Fienberg & Holland, 1973), this study found that the pseudo Bayes estimates' use of uniform and independence models with data-adaptive weights results in more accurate probability estimation than the exclusive use of either the raw probabilities or the modeled probabilities from the uniform and independence models. In equating applications the use of small constants is one common strategy for eliminating raw score probabilities of zero when estimating raw equipercentile equating functions (Hanson, 1990; Kolen & Brennan, 2004). To the extent that there is interest in evaluating a raw equipercentile equating function while also avoiding the difficulties of equipercentile calculations when test score probabilities are zero, this study's results suggest two ways which pseudo Bayes applications can improve on the small constants strategy. First, the data-adaptive weights of pseudo Bayes estimation can result in relatively accurate pseudo Bayes estimates when small constants are averaged with raw data. More importantly, this study's results suggest that a raw equipercentile equating function can be most accurately estimated by averaging probability estimates from loglinear models that fit more of the raw data than the uniform and independence models that correspond to the small constants strategy.

References

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

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

Fienberg, S. E., & Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association, 68*(343), 683–691.

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. (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.

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (ETS Technical Rep. No. 87-79). 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. (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.

Moses, T., & Holland, P. (2007). *Kernel and traditional equipercentile equating with degrees of presmoothing* (ETS Research Rep. No. RR-07-15). Princeton, NJ: ETS.

Moses, T., & Holland, P. (2008). *Selection strategies for univariate loglinear smoothing models and their effect on equating functions* (ETS Research Rep. No. RR-08-25). Princeton, NJ: ETS.

Moses, T., & Holland, P. (2009a). *Alternative loglinear smoothing models and their effect on equating function accuracy* (ETS Research Rep. RR-09-48). Princeton, NJ: ETS.

Moses, T., & Holland, P. (2009b). *Selection strategies for bivariate loglinear smoothing models and their effects on NEAT equating functions* (ETS Research Rep. No. RR-09-04). Princeton, NJ: ETS.

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