Statistica Sinica **8**(1998), 669-690

ESTABLISHING SCORE COMPARABILITY IN HETEROGENEOUS POPULATIONS

Michelle Liou

Academia Sinica

Abstract: In educational testing contexts, the relative comparability of scores on two tests is commonly established using the equipercentile method, which equates scores based on the corresponding percentile ranks in test score distributions. Because of security or disclosure considerations, data collection for a comparability study is often conducted using an incomplete-data design, that is, the two tests are given to two non-random groups at slightly different time points, and a set of common items is included in the test administration to allow some statistical adjustments for possible sample-selection bias. In the literature, researchers have made the missing-at-random assumption when estimating population score distributions using the common-item scores. This assumption can be violated in various ways, especially when the groups differ in ages or when the tests are administered a few months apart. In this study a general model is proposed for estimating score distributions using incomplete data; the model considers background information (e.g., gender, ethnicity) together with common-item scores as possible predictors of sample-selection bias, and allows nonresponse to depend on missing scores. The model parameters are estimated using the maximum-likelihood method and a Bayesian procedure. The standard errors of comparable scores are also derived under the proposed model. The use of the model is illustrated in two applications.

Key words and phrases: Bayesian methods, categorical data, data-imputation, equipercentile equating, EM algorithm, log-linear smoothing.

1. Introduction

Two educational tests measuring similar content are typically highly correlated and it has become a common practice to address the question "What score on Test-X is comparable to what score on Test-Y?" For instance, concordance tables were built for scores on the American College Testing (ACT) Assessment and the College Board's Scholastic Aptitude Test (SAT) (Marco, Abdel-Fattah and Baron (1992)) such that relative competence could be determined between applicants taking different tests for college admission; similarly, equivalent scores were found for the 1988 and standard versions of the Armed Services Vocational Aptitude Battery (ASVAB) such that the scores could be compared between takers of the new version and those of the older versions (Little and Rubin (1994),

Thomasson, Bloxom and Wise (1994)). In testing practice, the term "test equating" has been commonly refered to as the scaling of two equivalent forms of the same test (e.g., different versions of the ASVAB) to achieve score comparability. This study will adopt the general term "comparability studies" to include the scalings of target tests that measure similar content but are not necessarily equivalent forms (e.g., the ACT assessment and the SAT).

The equipercentile method (Angoff (1984)) defines a function e, such that $\mathrm{G}[e(x)] = \mathrm{F}(x)$ and $\mathrm{F}[e^{-1}(y)] = \mathrm{G}(y)$ for all x and y scores on the respective target tests X and Y, where F and G denote the distribution functions of x and y in the reference population; by definition, $e(x)$ and x are comparable scores on Y and X, respectively. In many testing programs, distributing tests to randomly equivalent groups before estimating F and G is not feasible because of security or disclosure considerations. A conventional data-collection design is to administer separate tests to two non-random groups at different time points. Because the two groups represent possibly different populations, a set of common items measuring content similar to the target tests is also included in the test administration to allow for some adjustments. Conventionally, the comparability study using the common-item scores relies on the missing-at-random (MAR; see Rubin (1976)) or ignorable missing assumption. In other words, the subpopulations the two groups represent are assumed to have the same target-score distribution when the common-item score is held constant (Braun and Holland (1982), Holland and Thayer (1989), Liou and Cheng (1995a), Rubin (1976)). This assumption is likely to fail when the two groups differ substantially in ages or abilities. In a few testing programs, information such as ethnicity, educational levels, and grades in school are also collected for takers of target tests; these background variables are useful for estimating missing scores. In the literature, certain background information has been recommended for matching subpopulations in comparability studies when common-item scores have small correlation with target-test scores (Wright and Dorans (1993)).

This research considers a generalized log-linear model to estimate missing distribution functions for takers of target tests; the model considers background information (e.g., gender, ethnicity, school grades, etc.) together with common-item scores as possible predictors of sample-selection bias, and allows nonresponse to depend on missing scores. The approach of treating nonignorable missingness considered herein builds on work by Little (1985), Little and Rubin (1987), Fay (1988), and Baker and Laird (1988) by adding a grouping variable for takers of either one of the tests to the model. In the next section, a brief review of the equipercentile method and its recent development is given, followed by an introduction to the generalized log-linear model for estimating missing score distributions. The model parameters are estimated using the maximum-likelihood (ML) method and a Bayesian procedure. Finally, the standard errors of estimating comparable scores are derived and the effectiveness of the model is examined in two applications.

2. The Equipercentile Method

In the equipercentile method, a score on Test-Y is comparable to a score on Test-X if

$$\hat{\xi} \equiv \hat{e}(x) = \hat{\mathrm{G}}^{-1}[\hat{\mathrm{F}}(x)], \tag{1}$$

where $\hat{\xi}$ or $\hat{e}(x)$ denotes the score on Test-Y corresponding to a score x on Test-X, and $\hat{\mathrm{F}}$ and $\hat{\mathrm{G}}$ are sample estimates of the population distribution functions. Because test scores are integers, a monotone transformation of $\hat{\mathrm{F}}$ and $\hat{\mathrm{G}}$ is required so that the function in (1) be well defined. This study follows the convention of assuming that the frequency attached to an integer score i is uniformly distributed in the interval $[i - 0.5, i + 0.5)$. Then the asymptotic standard error of $\hat{\xi}$ in (1) can be expressed as

$$Se(\hat{\xi}) \cong \{\mathrm{Var}\,[\hat{\mathrm{F}}(x)] + \mathrm{Var}\,[\hat{\mathrm{G}}(\xi)] - \mathrm{Cov}\,[\hat{\mathrm{F}}(x), \hat{\mathrm{G}}(\xi)]\}^{\frac{1}{2}} / \mathrm{g}(\xi)|_{\xi=\hat{\xi}} \tag{2}$$

(Liou and Cheng (1995b)), where $\mathrm{Var}\,[\hat{\mathrm{F}}(x)]$ and $\mathrm{Var}\,[\hat{\mathrm{G}}(\xi)]$ are the sample variances of the distribution estimates; $\mathrm{Cov}\,[\hat{\mathrm{F}}(x), \hat{\mathrm{G}}(\xi)]$ is the covariance; $\mathrm{g}(\xi)$ is the density at ξ. The use of (2) has been recommended for samples containing at least 1,000 examinees.

For small samples, estimates of population distributions are subject to substantial sampling errors. Therefore, a parametric smoothing on the sample distribution functions has been recommended to reduce noise in comparability studies (Rosenbaum and Thayer (1987), Hanson, (1991), Little and Rubin (1994)). A common smoothing model for score distributions is an extension of the log-linear model for scored multiway tables described by Haberman (1974). Let $\mathrm{f}(x)$ denote the density at x; the log-linear model for smoothing the x score distribution assumes

$$\mathrm{f}(x) = (\eta)(\tau_1)^x(\tau_2)^{x^2}\cdots(\tau_{\mathrm{q}})^{x^{\mathrm{q}}}, \tag{3}$$

where $\tau_1, \ldots, \tau_{\mathrm{q}}$ are the parameters to be estimated and η normalizes the sum of all $\mathrm{f}(x)$ to 1 (Rosenbaum and Thayer (1987)). For numerical stability, the original raw scores can be centered by replacing x by $x - \mathrm{I}/2$, for $x = 0, \ldots, \mathrm{I}$ before estimating the parameters. The ML estimators have the property that the fitted and observed moments of the X distribution are still identical after smoothing, to order q, so the smoothed distribution preserves some important features of the original sample. In the study by Liou and Cheng (1995b), the asymptotic formula in (2) worked reasonably well for sample sizes of 100 or more after the sample estimates of F and G were smoothed using (3).

As mentioned earlier, the sample estimates of population distributions can be biased through the selection of target tests takers. However, bias in comparable scores might still be minor, provided selection of samples depends on one of the target tests (X or Y), and all scores found in the reference population are also found in the selected population (Little and Rubin (1994)). In testing practice,

if selection is not symmetric in the two test taker groups (e.g., the ability of one group is substantially higher than the other), a set of common items must be administered to allow for possible adjustments. Let α and β denote groups of Test-X and Test-Y respectively, and let v be a common-item score which is not counted toward target test scores, with distribution $K(v)$. The MAR assumption posits that $\mathrm{F}_\alpha(x|v) = \mathrm{F}_\beta(x|v)$ (Holland and Thayer (1989), Liou and Cheng (1995a)), so $\mathrm{F}(x)$ can be estimated by computing

$$\hat{\mathrm{F}}(x) = \omega_\alpha \hat{\mathrm{F}}_\alpha(x) + \omega_\beta \sum_v \hat{\mathrm{F}}_\alpha(x|v)\hat{\mathrm{k}}_\beta(v), \tag{4}$$

where $\omega_\alpha = (1-\omega_\beta) = \mathrm{N}_\alpha/(\mathrm{N}_\alpha + \mathrm{N}_\beta)$, and N_α and N_β are the sample sizes. One can estimate $\mathrm{G}(y)$ in the same way. After $\hat{\mathrm{F}}(x)$ and $\hat{\mathrm{G}}(y)$ are found for all x and y scores, comparable scores on the two tests can be established using (1).

In small samples, the joint distribution of (x, v) and of (y, v) can also be smoothed using a model similar to (3) before estimating F and G in the reference population; some important features of the empirical distributions can be preserved in the model, say the major moments in the marginal x, y, and v distributions, and the cross-product moments in the bivariate (x, v) and (y, v) distributions. The use of smoothing bivariate tables for comparability studies has been empirically supported by Hanson (1991), Livingston (1993) and Allen, Holland and Thayer (1994). Alternatively, the estimation of the distribution in (4) and parameters in the smoothing model can be combined into an EM algorithm–the E-step computes the sufficient statistics for estimating τ's based on the MAR assumption and the M-step solves for τ's in the likelihood equations. It was shown (Liou and Cheng (1995a)) that the EM algorithm yields more efficient estimates of F and G and gives more stable estimates of comparable scores when the MAR assumption holds.

Conventionally, comparability studies using common-item scores have estimated the (x, v) and (y, v) distributions in the reference population; for example, the two distributions were estimated separately by implementing the EM algorithm twice in the study by Liou and Cheng (1995a). This study will focus on a general model for imputing the joint (x, y, v) distribution based on incomplete data and we allow nonresponse to depend on missing scores. The general model makes an examination of the XY interaction tenable and simplifies the estimation of $\mathrm{Cov}\,(\hat{F}, \hat{G})$ in (2).

3. Approaches to Estimating Population Distributions

Two missing-data patterns emerge from comparability studies: (i) Group α has observed $\hat{\mathrm{F}}_\alpha(x)$, $\hat{\mathrm{K}}_\alpha(v)$, possible background variables, and missing $\hat{\mathrm{G}}_\alpha(y)$; (ii) Group β has observed $\hat{\mathrm{G}}_\beta(y)$, $\hat{\mathrm{K}}_\beta(v)$, possible background variables, and missing $\hat{\mathrm{F}}_\beta(x)$. This section elaborates on different approaches to estimating $\mathrm{F}(x)$ and $\mathrm{G}(y)$ based on the observed data.

3.1. A general model

In a comparability study, there are at least three test variables (X, Y, and common item V scores), and one grouping variable C (α and β). Let $\mathrm{f}(i,j,k,c) = \mathrm{f}(x=i, y=j, v=k, \mathrm{C}=c)$ for $i = 0, \ldots, \mathrm{I}$, $j = 0, \ldots, \mathrm{J}$, $k = 0, \ldots, \mathrm{K}$, and $c = \alpha$ and β. The incomplete-data likelihood can be expressed as follows:

$$\mathrm{L} = \Big\{ \prod_i \prod_k [\sum_j \mathrm{f}(i,j,k,\alpha)]^{\mathrm{n}(i,.,k,\alpha)} \Big\} \Big\{ \prod_j \prod_k [\sum_i \mathrm{f}(i,j,k,\beta)]^{\mathrm{n}(.,j,k,\beta)} \Big\}, \qquad (5)$$

where $\mathrm{n}(i,.,k,\alpha)$ and $\mathrm{n}(.,j,k,\beta)$ are the observed marginal counts of the incompletely classified data in Groups α and β , respectively. For simplicity, let $\gamma^{\bar{C}} = (\tau_\alpha^C)/(\tau_\beta^C)$, $\gamma_i^{X\bar{C}} = (\tau_{i\alpha}^{XC})/(\tau_{i\beta}^{XC})$ and so on, where τ's are parameters to be estimated. Under the saturated model, the odds pertaining to C can be expressed as

$$\begin{aligned} \Omega_{ijk\cdot}^{\bar{\mathrm{C}}} &= \gamma^{\bar{\mathrm{C}}} \gamma_i^{\mathrm{X}\bar{\mathrm{C}}} \gamma_j^{\mathrm{Y}\bar{\mathrm{C}}} \gamma_k^{\mathrm{V}\bar{\mathrm{C}}} \gamma_{ij}^{\mathrm{XY}\bar{\mathrm{C}}} \gamma_{ik}^{\mathrm{XV}\bar{\mathrm{C}}} \gamma_{jk}^{\mathrm{YV}\bar{\mathrm{C}}} \gamma_{ijk}^{\mathrm{XYV}\bar{\mathrm{C}}}, \quad \text{where} \\ \Omega_{ijk\cdot}^{\bar{\mathrm{C}}} &= \mathrm{f}(i,j,k,\alpha)/\mathrm{f}(i,j,k,\beta), \quad \text{and} \end{aligned} \qquad (6)$$

$\prod_i \gamma_i^{\mathrm{X}\bar{\mathrm{C}}} = \prod_j \gamma_j^{\mathrm{Y}\bar{\mathrm{C}}} = \prod_k \gamma_k^{\mathrm{V}\bar{\mathrm{C}}} = \cdots = 1$. In the model, there are totally $2\mathrm{IJK} - 1$ parameters, and the incomplete data have $\mathrm{IK} + \mathrm{JK} - 1$ degrees of freedom for estimating them. Obviously, the model cannot be identified without more constraints. Table 1 contains five different models and their corresponding numbers of parameters.

Table 1. Hierarchical models for the 2×I×J×K table

Nonresponse Models	Fitted Marginals	Number of Parameters
1.	{XV,YV,VC}	IK+JK−1
2.	{YV,XVC}	IK+JK−1+(I−1)K
3.	{XV,YV,XC,VC}	IK+JK−1+(I−1)
4.	{XVC,YVC}	IK+JK−1+(I+J−2)K
5.	{XV,YV,XC,YC,VC}	IK+JK−1+(I+J−2)

The models in Table 1 contain at least IK+JK−1 parameters. The XY margin is tentatively removed from these models because X and Y were never given to the same takers. But the XY interaction is not inestimable in the strict sense; for example, the data contain information pertaining to the XY margin indirectly via the V variable (Rubin and Thayer (1978), Rubin (1995)). If the XY margin is included in these models, the number of parameters will exceed the size of data to a greater degree. We return to the problem of identification later in the section. In Table 1, Models 2 and 4 contain inestimable parameters because the XV margin is not observed in Group β nor is the YV margin observed in

Group α. In practice, if the size of the XV or YV interaction differs in Groups α and β, information on other background variables must be collected along with test data (Little and Rubin (1987)).

If the MAR assumption holds, (6) reduces to Model 1 as follows:

$$\Omega^{\bar{\mathrm{C}}}_{ijk\cdot} = \gamma^{\bar{\mathrm{C}}}\gamma^{\mathrm{V}\bar{\mathrm{C}}}_{k}, \tag{7}$$

where $\prod_k \gamma^{\mathrm{V}\bar{\mathrm{C}}}_k = 1$. This model contains IK + JK − 1 parameters and was recommended by Liou and Cheng (1995a) for imputing $\mathrm{F}(x)$ and $\mathrm{G}(y)$ when missing data patterns were generated from a double-sampling scheme. Alternatively, (7) can be described in the following equivalent form:

$$\begin{aligned} \mathrm{f}(i,j,k,c) &= \eta^{\mathrm{XYV}}_{ijk}\tau^{\mathrm{C}}_{c}\tau^{\mathrm{VC}}_{kc}, \quad \text{where} \\ \eta^{\mathrm{XYV}}_{ijk} &\equiv \eta\tau^{\mathrm{X}}_{i}\tau^{\mathrm{Y}}_{j}\tau^{\mathrm{V}}_{k}\tau^{\mathrm{XV}}_{ik}\tau^{\mathrm{YV}}_{jk}. \end{aligned} \tag{8}$$

In the regular log-linear model, a complete marginal distribution must be preserved whenever it is specified in the model. For instance, the X marginal probability is constrained to be identical in the fitted and observed distributions if the X main effect is included in the model. However, if Test-X contains 100 items, then $i = 0, \ldots, 100$; a complete specification of those 100 parameters for fitting the X marginal probability seems too tedious. Because the XYV margin is a scored three-way table (i.e., raw scores on the three tests), we may apply less stringent constraints (Haberman (1974), Rosenbaum and Thayer (1987)) as the smoothing model at (3), for instance,

$$\tau^{\mathrm{X}}_{i} = (\tau^{*}_{1})^{i}(\tau^{*}_{2})^{i^2}(\tau^{*}_{3})^{i^3}(\tau^{*}_{4})^{i^4}, \tag{9}$$

$$\tau^{\mathrm{XV}}_{ik} = (\tau^{**}_{1})^{ik}(\tau^{**}_{2})^{i^2k}(\tau^{**}_{3})^{ik^2}, \quad \text{and} \tag{10}$$

$$\tau^{\mathrm{VC}}_{kc} = (\tau^{*\mathrm{C}}_{1c})^{k}(\tau^{*\mathrm{C}}_{2c})^{k^2}, \tag{11}$$

The X marginal probability is constrained by (9), that is, the first, second, third, and fourth moments of the fitted and observed distributions are constrained to be identical; furthermore, the $\tau^{*}_{1}, \ldots, \tau^{*}_{4}$ parameters in (9) remain the same with respect to i. Similarly, the XV marginal probability is constrained by (10) so that the cross-product moments XV, $\mathrm{X}^2\mathrm{V}$, and XV^2 are identical in the fitted and observed distributions; the VC marginal probability is constrained by (11) which specifies that the group differences on the first and second moments of the V variable be identical in the fitted and observed distributions. In subsequent discussion, model (9) is denoted by 1st/2nd/3rd/4th −X, model (10) by linear/quadratic −XV, and model (11) by 1st/2nd −VC.

If missing data are truly MAR, then the ignorable model in (7) will work well for imputing and estimating the F and G distributions. When the missing data do not follow the MAR assumption, some XC and YC interactions may

be included in the model. If Groups α and β differ substantially in ability, for instance, their X and Y scores distributions are likely different and the odds can be expressed as

$$\Omega^{\bar{\mathrm{C}}}_{ijk\cdot} = \gamma^{\bar{\mathrm{C}}}\gamma_i^{\mathrm{X}\bar{\mathrm{C}}}\gamma_j^{\mathrm{Y}\bar{\mathrm{C}}}\gamma_k^{\mathrm{V}\bar{\mathrm{C}}}, \tag{12}$$

where $\prod_i \gamma_i^{\mathrm{X}\bar{\mathrm{C}}} = \prod_j \gamma_j^{\mathrm{Y}\bar{\mathrm{C}}} = \prod_k \gamma_k^{\mathrm{V}\bar{\mathrm{C}}} = 1$. The hypothesis in (12) corresponds to Model 5 in Table 1 and has more parameters than degrees of freedom in the data. Therefore, some properties of scored multiway tables must be considered as was done at (9) through (11). The model in (12) is only one example of nonignorable missingness. Other cases arise in which the XV (likewise YV) interactions differ in the two groups. For instance, a subset of common items might have been given to a group in the past and scores on these items do not count toward total test scores. If the identification of these common items is known to some takers at the second time points, these takers tend to do worse on the common items, and the correlation between X and V in the second group tends to be lower. As was mentioned, the parameters pertaining to the XVC and YVC margins are inestimable. In applications, the cross-classifications of samples on some background variables (e.g., grades in school and socioeconomic status) might also be available. Let B be the cross-classification of samples on the background variables. If the XVB and YVB margins are included in the hierarchical response models, the XVC and YVC interactions can possibly be estimated (Little and Rubin (1987), Baker and Laird (1988)). The background variables might also increase the efficiency of estimating $\mathrm{F}(x)$ and $\mathrm{G}(y)$ if they are highly correlated with test scores.

3.2. The ML estimates of parameters

In comparability studies, the principal interest is in making inference about the X and Y margins based on the combined data from Groups α and β. The ML estimates of model parameters in (5) can be evaluated via the EM algorithm. The complete-data sufficient statistic for estimating these τ parameters can be expressed as follows:

$$S = \sum_{ijkc} i^{\mathrm{s}} j^{\mathrm{t}} k^{\mathrm{u}} g^{\mathrm{v}} \tilde{\mathrm{n}}(i,j,k,c), \tag{13}$$

where g is the code assigned to the c-th group (e.g., $\alpha = -1$ and $\beta = 1$), and s, t, u and v take on integer values as they are specified in the marginal model (for instance, the sufficient statistic for estimating τ_2^{**} in (10) has $\mathrm{s} = 2$, $\mathrm{t} = 0$, $\mathrm{u} = 1$, and $\mathrm{v} = 0$); $\tilde{\mathrm{n}}(i,j,k,c)$ denotes the estimated counts for the completely classified data. Specifically, the EM cycle in the q-th iteration consists of the following steps:

E-Step.

$$S^{(\mathrm{q})} = \sum_{ijkc} i^{\mathrm{s}} j^{\mathrm{t}} k^{\mathrm{u}} g^{\mathrm{v}} \tilde{\mathrm{n}}^{(\mathrm{q})}(i,j,k,c), \quad \text{where}$$
$$\tilde{\mathrm{n}}^{(\mathrm{q})}(i,j,k,\alpha) = \{\tilde{\mathrm{f}}^{(\mathrm{q})}(i,j,k,\alpha)/\tilde{\mathrm{f}}^{(\mathrm{q})}(i,.,k,\alpha)\}\mathrm{n}(i,.,k,\alpha) \quad \text{and}$$
$$\tilde{\mathrm{n}}^{(\mathrm{q})}(i,j,k,\beta) = \{\tilde{\mathrm{f}}^{(\mathrm{q})}(i,j,k,\beta)/\tilde{\mathrm{f}}^{(\mathrm{q})}(.,j,k,\beta)\}\mathrm{n}(.,j,k,\beta).$$

This step adjusts the values of the sufficient statistics to the provisional $\mathrm{n}(i,\cdot,k,\alpha)$, $\mathrm{n}(\cdot,j,k,\beta)$, and the current parameter $\tilde{\tau}^{(q)}$ estimates, and in these equations $\tilde{\mathrm{f}}^{(\mathrm{q})}(i,.,k,\alpha)$ and $\tilde{\mathrm{f}}^{(\mathrm{q})}(.,j,k,\beta)$ are the estimates of the XV and YV marginal probabilities from the q-th iteration in Groups α and β, respectively.

M-Step.

$$\sum_{ijkc} i^{\mathrm{s}} j^{\mathrm{t}} k^{\mathrm{u}} g^{\mathrm{v}} \tilde{\mathrm{f}}^{(\mathrm{q}+1)}(i,j,k,c) = S^{(\mathrm{q})}/\mathrm{N}.$$

This step solves for the parameter $\tilde{\tau}^{(q+1)}$ estimates using the adjusted values of the sufficient statistics from the latest E-step. The EM cycles are repeated until the sequence of iterates $\tilde{\tau}^{(\mathrm{q})}$ becomes stable. The ML estimates of $\mathrm{F}(x)$ and $\mathrm{G}(y)$ can be calculated using $\tilde{\mathrm{n}}(i,j,k,c)$ from the stopped EM cycle.

A model's goodness-of-fit is normally measured by the likelihood ratio statistic, that is,

$$\mathrm{LR} = 2\Big\{ \sum_{ik} \mathrm{n}(i,.,k,\alpha) \log[\mathrm{n}(i,.,k,\alpha)/\tilde{\mathrm{n}}(i,.,k,\alpha)] + \sum_{jk} \mathrm{n}(.,j,k,\beta) \log[\mathrm{n}(.,j,k,\beta)/\tilde{\mathrm{n}}(.,j,k,\beta)] \Big\}. \tag{14}$$

The degrees of freedom for LR are IK + JK − (number of estimable τ parameters fit). When data are incomplete and some τ parameters are not strictly estimable (e.g., the correlation between X and Y), the LR statistic may not work well for selecting between ignorable and nonignorable models. Experience suggests that a model with a small LR can produce unsatisfactory F and G estimates. In practice, other criteria are also useful for selecting between models, such as bivariate plottings of x scores vs. their comparable scores on Y, or of x scores vs. the standard error of their comparable scores. The issue of using LR to select between models will be discussed later in the context of applications.

3.3. A Bayesian procedure

Selecting between ignorable or nonignorable models can cause controversy. Rather than selecting one model for data analyses, a compromise on all possible solutions would assume a smoother nature for the missing-data mechanism. Rubin and Schafer (1988) suggest assigning a low a priori probability to the presence

of higher-order interactions in the saturated log-linear model. This prior distribution pulls the τ estimates toward a parsimonious model rather than toward an ignorable or a nonignorable model (Rubin (1995)). A simple family of prior distributions that accomplish this aim can be illustrated via a few examples as follows:

$$\tau_i^{\mathrm{X}} \sim \mathrm{N}(0, \sigma^2) \tag{15}$$

$$\tau_{ik}^{\mathrm{XV}} \sim \mathrm{N}(0, \sigma^2/\lambda) \tag{16}$$

$$\tau_{ikc}^{\mathrm{XVC}} \sim \mathrm{N}(0, \sigma^2/\lambda^2) \quad \text{for } \sigma^2 > 0, \quad \text{and } \lambda > 1 \tag{17}$$

where the τ's represent the first-, second-, and higher-order interactions in the saturated log-linear models, and are assumed to be independently distributed.

This Bayesian procedure applied to the saturated model will be more robust than any ignorable model when the MAR assumption is seriously violated (Rubin and Schafer (1988)). As mentioned earlier, a complete specification of the saturated model is unnecessarily tedious for scored tables. For the use of comparability studies, this family of prior distributions can be applied to a reduced model which includes all useful parameters for estimating the X and Y margins. The EM algorithm can also be applied to the log-posterior distribution of τ's as easily as to the log-likelihood function at (5).

4. Asymptotic Standard Error of Comparable Scores

For notational simplicity let $p \equiv \mathrm{F}(x)$ and

$$\tilde{e}(x) = \tilde{\mathrm{G}}^{-1}(\tilde{p}), \tag{18}$$

where $\tilde{p}$ and $\tilde{\mathrm{G}}$ denote the F and G estimates computed using $\tilde{\mathrm{n}}(i, j, k, c)$ from the stopped EM cycle. Based on the uniform assumption, the first derivatives of F and G exist almost everywhere. Then by the Bahadur theorem (1966), $\tilde{\mathrm{G}}$ can be expressed as follows:

$$\tilde{\mathrm{G}}^{-1}(p) = \xi + \frac{p - \tilde{\mathrm{G}}(\xi)}{\mathrm{g}(\xi)} + \mathrm{R_N}, \tag{19}$$

where $0 < p < 1$, and N is the sample size. Moreover, it can be shown that $\mathrm{R_N} = \mathrm{o}_p(\mathrm{N}^{-\frac{1}{2}})$ (e.g., Ghosh (1971)). Thus (18) can be written as

$$\tilde{\xi} \equiv \tilde{\mathrm{G}}^{-1}(\tilde{p}) = \mathrm{G}^{-1}(\tilde{p}) + \frac{\tilde{p} - \tilde{\mathrm{G}}[\mathrm{G}^{-1}(\tilde{p})]}{\mathrm{g}[\mathrm{G}^{-1}(\tilde{p})]} + \mathrm{R_N}. \tag{20}$$

which yields

$$\begin{aligned} \tilde{\mathrm{G}}^{-1}(\tilde{p}) &\cong \mathrm{G}^{-1}(p) + \frac{\tilde{p} - p}{\mathrm{g}(\xi)} + \Big\{\tilde{p} - \tilde{\mathrm{G}}[\mathrm{G}^{-1}(p) + \frac{\tilde{p} - p}{\mathrm{g}(\xi)} + \cdots]\Big\}/\mathrm{g}(\xi) \\ &\cong \mathrm{G}^{-1}(p) + \frac{\tilde{p} - p}{\mathrm{g}(\xi)} + \Big\{\tilde{p} - \tilde{\mathrm{G}}(\xi) - \frac{\tilde{p} - p}{\mathrm{g}(\xi)}[\frac{\partial \tilde{\mathrm{G}}(t)}{\partial t}|_{t=\xi}]\Big\}/\mathrm{g}(\xi). \end{aligned} \tag{21}$$

With some simplification, (21) can be reduced to

$$\begin{aligned}\tilde{\mathrm{G}}^{-1}(\tilde{p}) &= \xi + \frac{\tilde{p}-p}{\mathrm{g}(\xi)} + \frac{p-\tilde{\mathrm{G}}(\xi)}{\mathrm{g}(\xi)}[1+\mathrm{o}(1)] \\ &\cong \xi + \frac{\tilde{p}-\tilde{\mathrm{G}}(\xi)}{\mathrm{g}(\xi)}. \end{aligned} \tag{22}$$

According to (22), the asymptotic standard error of $\tilde{\xi}$ can be estimated by

$$Se(\tilde{\xi}) = \left\{\mathrm{Var}\,[\tilde{\mathrm{F}}(x)] + \mathrm{Var}\,[\tilde{\mathrm{G}}(\xi)] - 2\,\mathrm{Cov}\,[\tilde{\mathrm{F}}(x), \tilde{\mathrm{G}}(y)]\right\}^{\frac{1}{2}} / \mathrm{g}(\xi)|_{\xi=\tilde{\xi}}, \tag{23}$$

where $\mathrm{g}(\tilde{\xi})$ is the ML estimate. In (23), both $\tilde{\mathrm{F}}(x)$ and $\tilde{\mathrm{G}}(y)$ are functions of τ's; the variances of these functions of random variables can be estimated by

$$\mathrm{Var}\,[\tilde{\mathrm{F}}(x)] = [\frac{\partial \mathrm{F}(x)}{\partial \underline{\tau}}]\,\mathrm{Cov}\,(\underline{\tilde{\tau}})[\frac{\partial \mathrm{F}(x)}{\partial \underline{\tau}}]^{\mathrm{T}}|_{\underline{\tau}=\underline{\tilde{\tau}}}, \tag{24}$$

$$\mathrm{Var}\,[\tilde{\mathrm{G}}(\xi)] = [\frac{\partial \mathrm{G}(\xi)}{\partial \underline{\tau}}]\,\mathrm{Cov}\,(\underline{\tilde{\tau}})[\frac{\partial \mathrm{G}(\xi)}{\partial \underline{\tau}}]^{\mathrm{T}}|_{\underline{\tau}=\underline{\tilde{\tau}}}, \text{ and} \tag{25}$$

$$\mathrm{Cov}\,[\tilde{\mathrm{F}}(x), \tilde{\mathrm{G}}(\xi)] = [\frac{\partial \mathrm{F}(x)}{\partial \underline{\tau}}]\,\mathrm{Cov}\,(\underline{\tilde{\tau}})[\frac{\partial \mathrm{G}(\xi)}{\partial \underline{\tau}}]^{\mathrm{T}}|_{\underline{\tau}=\underline{\tilde{\tau}}}, \tag{26}$$

where T denotes the transposition of a matrix; $\underline{\tau}$ denotes the vector of parameters and $\mathrm{Cov}\,(\underline{\tilde{\tau}})$ is the variance-covariance matrix of the estimates. The last matrix can be estimated by the inverse of the observed information matrix. The observed information matrix can itself be derived either by using the supplemented EM algorithm (Meng and Rubin (1991)) or by taking the second derivative of the incomplete-data log-likelihood in (5). If $\mathrm{F}(x)$ and $\mathrm{G}(y)$ are estimated via the Bayesian procedure, $\mathrm{Cov}\,(\tilde{\tau})$ can be approximated by taking the inverse of the posterior information.

5. Examples

5.1. The NAEP reading test

The reading test scores were collected for the 1990 National Assessment of Educational Progress (NAEP) across different age samples. The Age 13/Grade 8 Form contained 34 multiple-choice (MC) items and was given to 1,248 takers; on the other hand, the Age 17/Grade 12 Form contained 33 MC items and was given to 1,180 takers; an additional fourteen common items were given to the two age groups. We denote the Age 13/Grade 8 Form as X, the Age 17/Grade 12 Form as Y, and the common-items as V. In order to make a comparison of reading performance between age groups, comparable scores had to be found for the two forms (the scaling of different age forms to achieve score comparability has been termed vertical equating in educational testing). For each test taker, the study

computed two scores: one on the common items and one on the appropriate target test (scores were simply the number of correct answers). Let x, y and v be scores on the target forms and common items, respectively. The basic statistics for these scores were (i) mean $x = 19.93$, mean $y = 23.55$, mean $v_\alpha = 5.83$ (Age 13/Grade 8 samples), mean $v_\beta = 8.22$ (Age 17/Grade 12 samples); (ii) standard deviation $x = 6.87$, standard deviation $y = 6.27$, standard deviation $v_\alpha = 2.71$, standard deviation $v_\beta = 3.10$. The correlation between x and v was .708 in the Age 13/Grade 8 sample, and that between y and v was .725 in the Age 17/Grade 12 sample. Because of the difference in age, common-item score distributions had different shapes and locations.

Table 2. Model comparisions for the NAEP reading data

	Response Models	No. of Par.	LR
N1.	{XV,YV,XY,C}		
	Nla linear −XY,−YV,−XY	18	1237.550
	N2b linear/quadratic −XV,−YV,−XY	24	1184.612
N2.	{XV,YV,XY,VC}		
	linear/quadratic −XV,−YV,−XY+		
	N2a. 1st/2nd/3rd −VC	27	793.336
	N2b. 1st/2nd −VC	26	793.526
	N2c. 1st/3rd −VC	26	793.444
N3.	{XV,YV,XY,XC}		
	linear/quadratic −XV,−YV,−XY+		
	N3a. 1st/2nd −XC	26	808.140
	N3b. 1st/3rd −XC	26	806.719
N4.	{XY,YV,XY,YC}		
	linear/quadratic −XV,−YV,−XY+		
	N4a. 1st/2nd −YC	26	793.855
	N4b. 1st/3rd −YC	26	794.022
N5.	{XV,YV,XY,YC,VC}		
	linear/quadratic XV,YV,XY+		
	N5a. 1st −YC,−VC	26	803.489
	N5b. 1st −YC+3rd−VC	26	792.842
N6.	{XV,YV,XY,XC,YC,VC}		
	linear/quadratic −XV, −YV,−XY+		
	N6a. 3rd −XC,−VC+1st/2nd−YC	28	789.810
	N6b. 1st/3rd −XC,−VC+1st/2nd−YC	31	775.335

Note: All these models fit the first five moments of X and Y, and the first four moments of V

Table 2 contains the LR statistics for fitting different response models to the NAEP reading data. Model N1 fits all possible two-way interactions between X,

Y and V and assumes that the three variables have no interaction with C. This model contains two submodels; in Model N1a, linear relationships are preserved for those two-way interactions, and in Model N1b, both the linear and quadratic relationships are preserved (see section 3.1 for the specification of parameters in the model). The results suggest that the linear/quadratic −XV, −YV and −XY model has smaller LR value than the linear model (1184.612 vs. 1237.550). Models N2 through N6 are extensions of Model N1b. In Table 2, N1 and N2 are ignorable models in which the XC and YC interactions are not considered. The XC and YC interactions are successively included in Models N3 through N6. The marginal X, Y, V and C contains 15 parameters to be estimated for all the models in Table 2 (except for N1a). The total number of parameters equals 15 plus parameters for estimating the two-way interactions; for instance, the number of parameters in Model N6a equals 15 + 9 (for XV, YV, XY) + 2 (for XC, VC) + 2 (for YC).

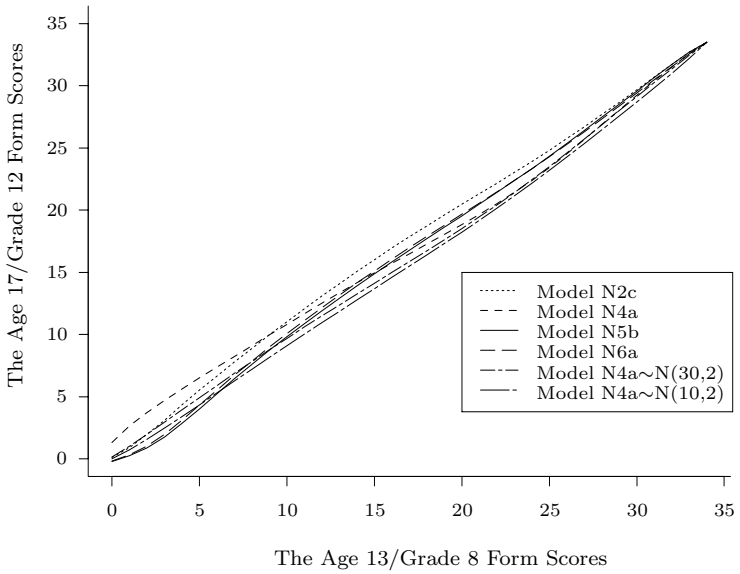

Figure 1. The equipercentile functions between scores on the Age 13/Grade 8 and Age 17/Grade 12 Forms for different models in Table 2.

Figure 1 plots the equipercentile functions based on the different models in Table 2. The parameters in Model N4a were also estimated by the Bayesian procedure with $\sigma^2 = 30, 10$ and $\lambda = 2$, respectively. The values of σ^2 and λ were selected such that the Bayesian estimates would not drastically change the shapes of the score distributions derived from those of their ML counterparts. In other applications, different criteria could be used for specifying the σ^2 and

λ values. Figure 1 also presents the equipercentile functions based on these Bayesian estimates. Interestingly, Model N2c (i.e., the ignorable model) tends to have larger comparable scores on Y across the x-score range as compared with other models. The equipercentile functions for Models N5b and N6a coincide with each other even though the two models fit the data differently (LR = 792.842 and 789.810). Model N4a tends to have larger comparable scores at the lower tail. The Bayesian procedure pulls the Model N4a function downward at the lower end. Figure 2 plots the standard errors of comparable scores based on (23), for different models. The ignorable model has the smallest standard errors across the score range. Model N6a has larger standard errors as compared with other models. The Bayesian estimates for Model N4a have smaller standard errors than their ML counterparts, particularly for those comparable scores at the lower and upper tails.

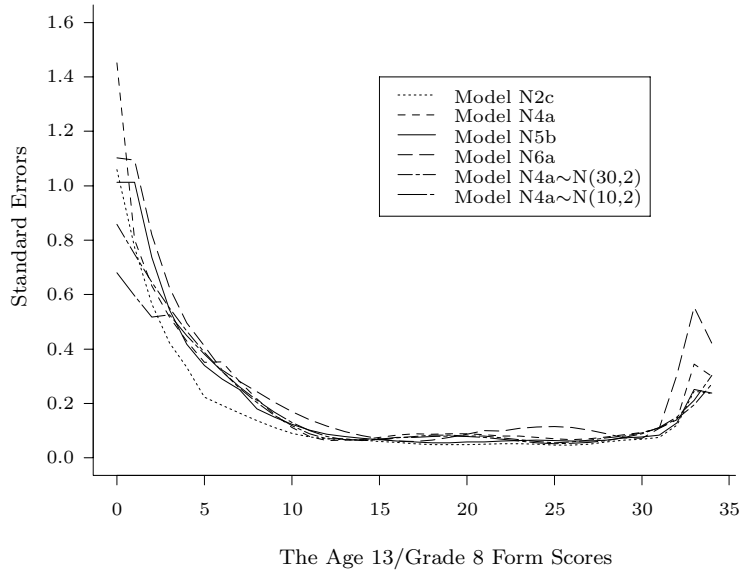

Figure 2. The standard errors of comparable scores for different models in Table 2.

The standard error plots suggest that the ignorable model yields smaller standard errors across the score range than do other nonignorable models. Figure A1 in Appendix plots the observed and fitted X and Y distributions from Model N2c. According to Figure A1 − (a), the observed distribution in the Age 13/Grade 8 sample was mildly right-skewed; the fitted distribution after including imputed scores from the Age 17/Grade 12 sample becomes slightly left-skewed. In Figure A1 − (b), however, the fitted distribution after including adjustments with scores from younger students (the larger sample) still possesses a similar shape as the

observed distribution. Intuitively, we would expect a greater portion of samples having lower y scores. Figure A2 plots the fitted distribution based on Model N4a − the model yields a similar shape of the fitted X distribution as that of Model N2c; however, it shifts the mode of the Y distribution to a lower score. For this particular example, the nonignorable model yields score distributions that are intuitively more appealing as compared with those from the ignorable model, even though the likelihood ratio statistics suggest that these models fit the data equally well (i.e., 793.444 and 793.855). Figure A3 presents the Bayesian estimates of the X and Y distributions with $\sigma^2 = 30$ and $\lambda = 2$ for Model N4a. The Bayesian procedure provides smoother estimates than does its ML counterpart for the two marginal distributions.

If the two forms were equally difficult, the equipercentile function would have been closer to a standard line with a slope of 1 and an intercept of 0. Another criterion for selecting between the models in Table 2 would come from evaluating the equipercentile functions in Figure 1 against this standard line. Notably, Model N2c yields comparable scores that all lie above the standard line across the x-score range. Accordingly, an x score on the Age 13/Grade 8 Form is comparable to a larger y score on the harder form. Intuitively, we would expect the opposite to occur. The N4a function lies above the standard line at the lower tail, but its Bayesian counterparts lie below the standard line across the x-score range. The two other nonignorable models also yield functions which lie below the standard line at the lower tail and closer to the standard line at the upper tail. A choice between these models must also take into account other issues; for instance, an analysis can be done to compare the relative difficulty between the two forms. In general, we would prefer the nonignorable models in Figure 1 to the ignorable model.

5.2. The Spanish language tests

Comparable scores were found for two Spanish language tests developed for a placement service. The tests were distinct and each contained 39 MC items and three constructed-response (CR) questions. Comparable scores on the MC items were well-established for the two tests from earlier studies. Therefore, the MC tests were assumed to be interchangeable for test takers once their raw scores on the tests were rescaled into comparable scores. Rescaled scores on the MC items are denoted as V. The CR questions scored from 0 to 52 and were considered the target tests in the comparability study. Sample scores on the Spanish language tests were collected from 1993 through 1994 and there were 142 takers for Test-X together with 103 takers for Test-Y available for the study. The basic statistics for test scores were: (i) mean $x = 34.51$, mean $y = 31.51$, mean $v_\alpha = 24.88$

(Test-X Group), and mean $v_\beta = 23.95$ (Test-Y Group); (ii) standard deviation $x = 7.85$, standard deviation $y = 10.32$, standard deviation $v_\alpha = 7.51$, and standard deviation $v_\beta = 7.47$. The correlation between x and v was .698 in the Test-X Group, and that between y and v was .797 in the Test-Y Group.

Table 3. Model comparisions for the Spanish language tests.

	Response Models	No. of Par.	LR
S1.	{XV,YV,XY,L,C}		
	linear −XV,−YV,−XY	14	1247.126
S2.	{XVL,YVL,XYL,C}		
	1st −VL, −XL,−YL+		
	linear −XV,−YV,−XY+		
	linear −XVL, −YVL, −XYL	20	1201.319
S3.	{XVL,YVL,XYL,VC}		
	1st −VL, −XL, −YL+		
	linear −XV,−YV,−XY+		
	linear −XVL,−YVL,−XYL+		
	1st −VC	21	1200.250
S4.	{XVL,YVL,XYL,XC,YC}		
	1st −VL, −XL, −YL+		
	linear −XV,−YV,−XY+		
	linear −XVL,−YVL,−XYL+		
	1st −XC,−YC	22	1200.004
S5.	{XVL,YVL,XYL,XVC,YVC}		
	1st −VL, −XL, −YL+		
	linear −XV,−YV,−XY+		
	linear −XVL,−YVL,−XYL+		
	S5a. linear −XVC,−YVC	22	1198.590
	S5b. 1st −XC, −YC+linear −XVC,−YVC	24	1186.658

Note: All these models fit the first three moments of X, Y and V.

Additionally, the background information on "language learned as child (Spanish vs. others)" was also available for all samples. Those test takers who spoke Spanish as children tended to score higher on the tests than those who did not. Let L be a classification of test takers on language groups (Spanish = −1, others = 1). Table 3 contains the LR statistics for fitting different response models to the Spanish language data. Because the incomplete-data distributions for both tests were sparse due to the small sample sizes, this study simply fitted the X, Y, and V marginal probabilities by preserving the first three univariate moments of these margins in order to compromise between the biases and standard errors in $\tilde{\mathrm{F}}(x)$ and $\tilde{\mathrm{G}}(y)$ (it was found that the model preserving the first three

sample moments yielded the least mean-squared difference between sample and population comparable scores in small samples; see Livingston (1993)). According to the LR statistics, the interactions between L and test variables significantly reduce the fit statistics (Model S1 vs. Model S2). After the difference between the two test groups on X, Y and V have been accounted for by L, the interactions between C and other variables do not further reduce the fit statistic. Figure 3 plots the equipercentile functions based on different models. The parameters in Model S5a were estimated using the Bayesian procedure with $\sigma^2 = 30, 10$ and $\lambda = 2$, respectively. Figure 3 also plots the equipercentile functions based on the Bayesian estimates.

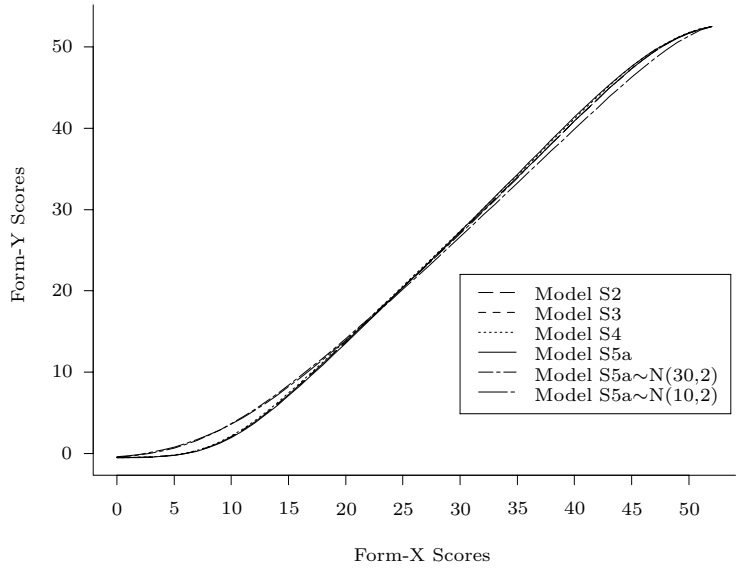

Figure 3. The equipercentile functions between scores on Tests- X and Y for different models in Table 3.

Obviously, the missing-at-random assumption does not raise an issue for the model-data fitting of the Spanish language data because the ignorable and nonignorable models provide similar comparable scores across the x-score range. However, the small sample size causes trouble. Figure A4 plots the X and Y distributions based on Model S5a. Notably, the observed data provide no score information at the lower tail in the X distribution. The log-linear model slightly improves the data sparsity occurring at the lower tail, but not too much. For $x = 0, \ldots, 6$, therefore, comparable scores on Y are almost zero. Figure 4 plots the standard errors of comparable score for different models. Clearly, the standard errors corresponding to smaller x scores are larger than is desirable. The Bayesian

procedure works much better than its ML counterpart in terms of estimating comparable scores at the lower tail. Figure A5 plots X and Y distributions based on the Bayesian estimates with $\sigma^2 = 10$, and $\lambda = 2$. As indicated in Figures 3 and 4, the Bayesian procedure provides reasonable estimates of comparable scores at the lower tail, and significantly reduces the standard errors for those comparable scores as well.

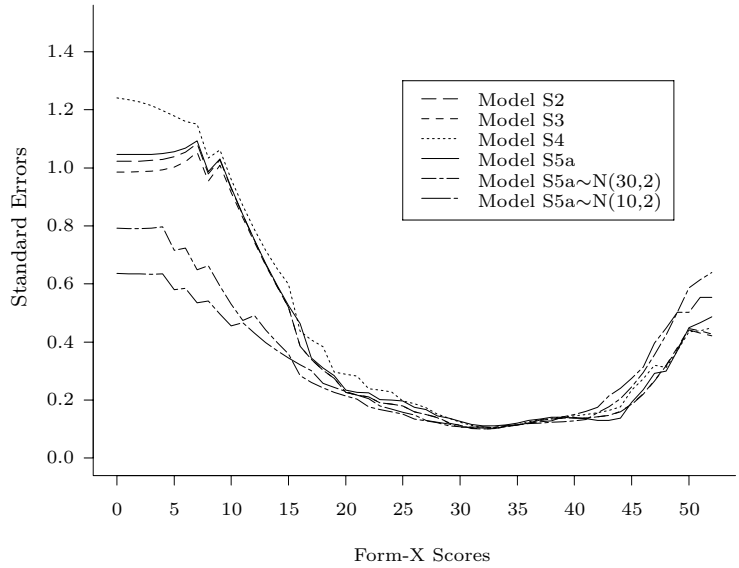

Figure 4. The standard errors of comparable scores for different models in Table 3.

6. Conclusion

The fittings of the NAEP reading data suggest that the ignorable model has more difficulty in shifting the mode of the Y distribution to a lower score than the nonignorable model does. The fittings of the Spanish language data suggest that the L variable works even better than the C variable for predicting sample-selection bias. In applications, when common items are not sensitive to substantial bias in test taker groups, comparability studies will benefit highly from adding background variables to the imputation models. The nonignorable models can also be considered for correcting bias in test taker groups when useful background data are not available. The proposed parametric approach of imputing score distributions requires more computer memory and the EM convergence can sometimes be too slow. So far, the application of the log-linear model is limited to short tests and a small number of background variables. In this study,

plotting complete-data distributions is recommended as a useful device for model selection because a valid model must yield score distributions that at least make intuitive sense. The Bayesian procedure is a compromise between ignorable and nonignorable models; it has been proposed in this study as a tool for small-sample applications. Empirical studies also suggest that the procedure is particularly useful for smoothing the tails of score distributions. Although the values of σ^2 and λ can be chosen arbitrarily, the plottings of equipercentile functions and the standard error of comparable scores can be used as devices for deciding the σ^2 and λ values in applications. In conclusion, this study recommends the use of the generalized log-linear model for comparability studies.

Acknowledgement

The author thanks Alison Snieckus and Steve Wang for their assistance in preparing the data, and Philip E. Cheng, Eugene G. Johnson, Donald B. Rubin, Juliet P. Shaffer, referees and Associate Editor for their useful comments on the draft manuscript. Support for this research was provided by the National Assessment of Educational Progress, US Department of Education.

Appendix

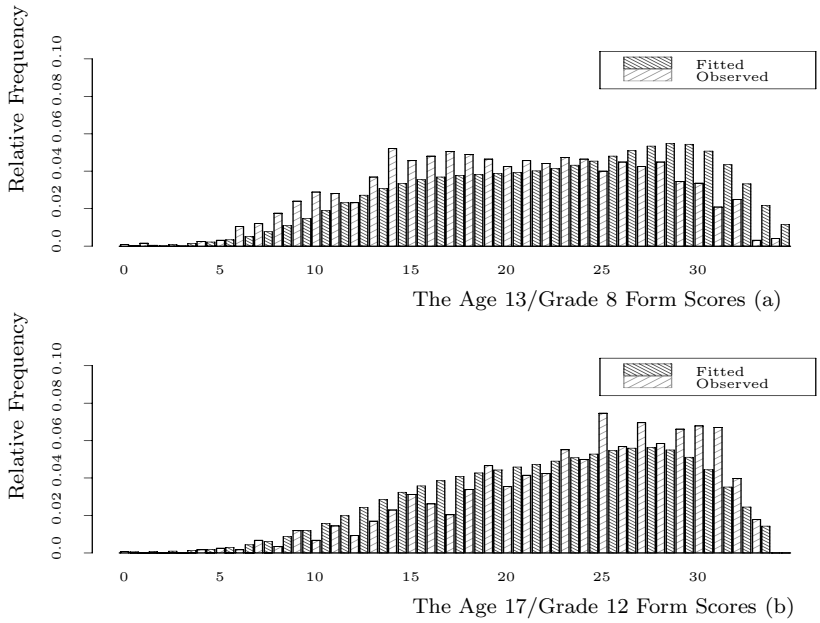

Figure A1. The observed and fitted score distributions corresponding to model N2c in Table 2 for the (a) Age 13/Grade 8 and (b) Age 17/Grade 12 Forms.

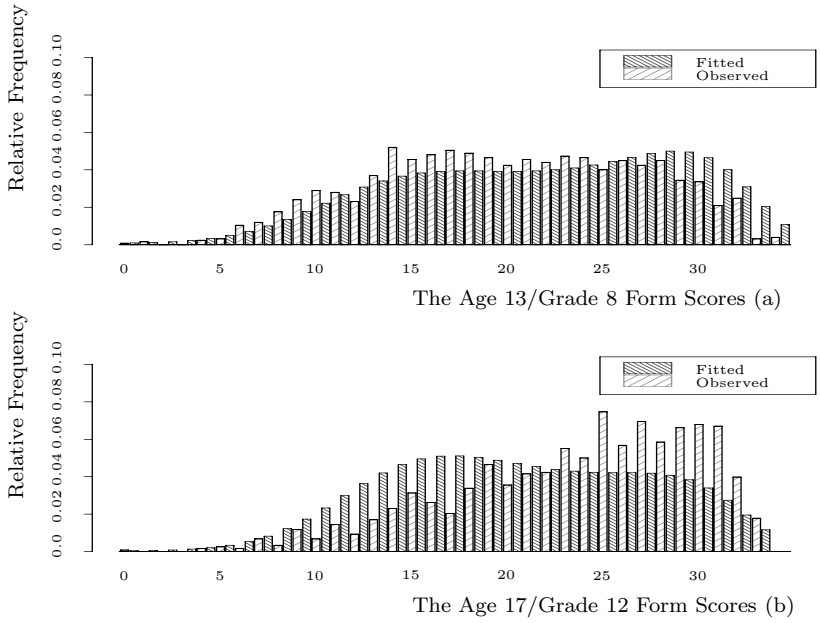

Figure A2. The observed and fitted score distributions corresponding to model N4a in Table 2 for the (a) Age 13/Grade 8 and (b) Age 17/Grade 12 Forms.

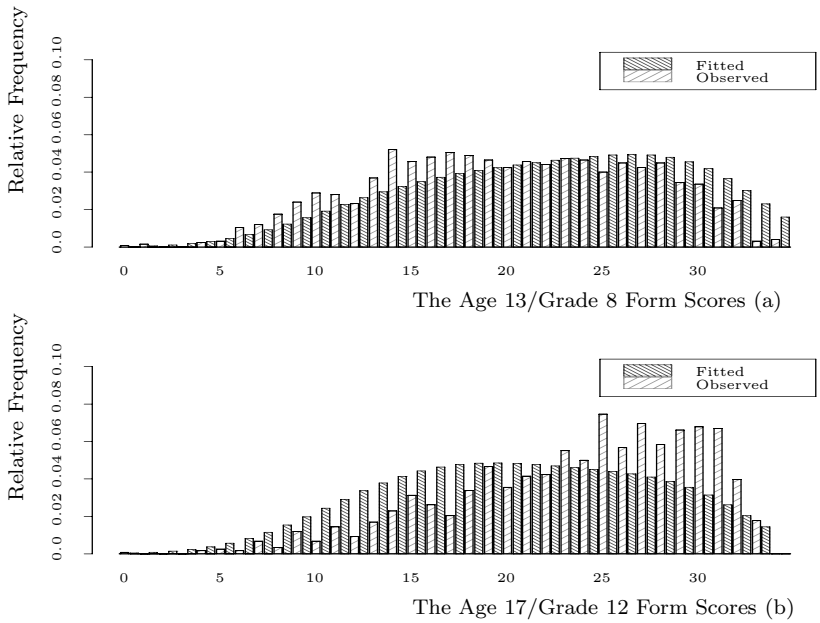

Figure A3. The observed and fitted score distributions corresponding to the Bayesian estimate of model N4a in Table 2 with $\sigma^2 = 30, \lambda = 2$ for the (a) Age 13/Grade 8 and (b) Age 17/Grade 12 Forms.

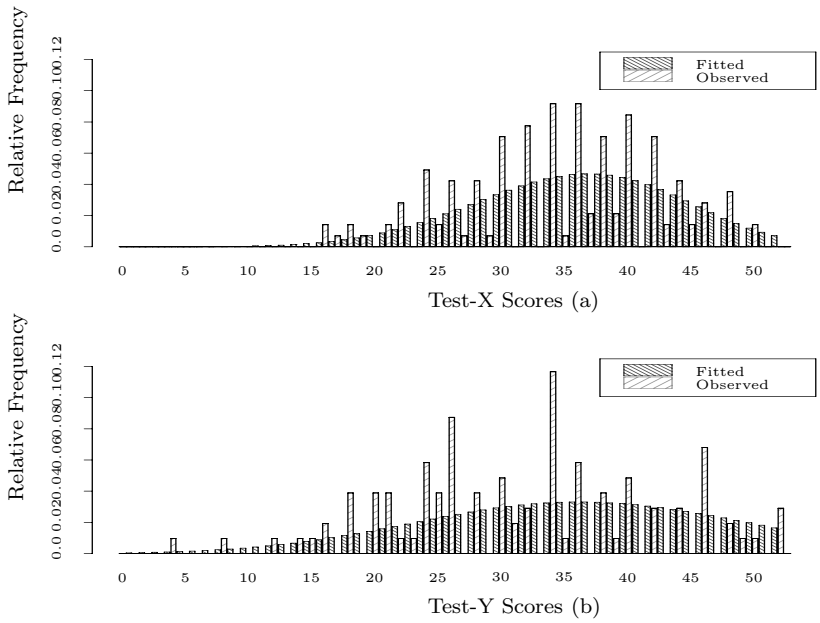

Figure A4. The observed and fitted score distributions corresponding to model S5a in Table 3 for (a) Test-X and (b) Test-Y.

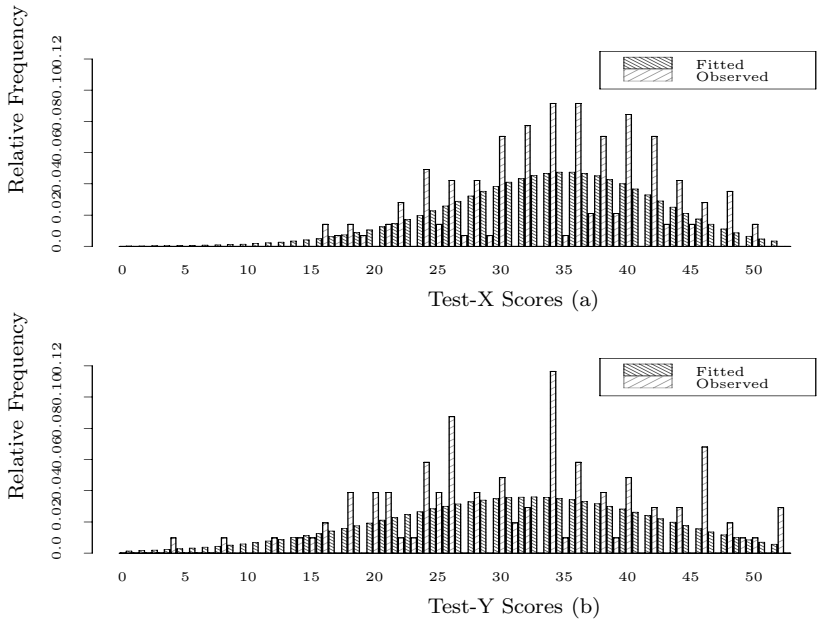

Figure A5. The observed and fitted score distributions corresponding to the Bayesian estimate of model S5a in Table 3 with $\sigma^2 = 10, \lambda = 2$ for (a) Test-X and (b) Test-Y.

References

Allen, N. L., Holland, P. W. and Thayer, D. T. (1994). Estimating scores for an optional test section using information from a common section. Research Report 94-18, Educational Testing Service, Princeton.

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service. (Reprint of chapter in *Educational Measurement*, Second Edition, 1971 (Edited by R. L. Thorndike), 508-600, American Council on Education Washington.)

Bahadur, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37**, 577-580.

Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Amer. Statist. Assoc.* **83**, 62-69.

Braun, H. I. and Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In *Test Equating* (Edited by P. W. Holland and D. B. Rubin), 9-49. Academic Press, New York.

Fay, R. E. (1986). Causal models for patterns of nonresponse. *J. Amer. Statist. Assoc.* **81**, 354-365.

Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Statist.* **42**, 1957-1961.

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

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

Holland, P. W. and Thayer, D. T. (1987). Notes on the use of log-linear models for fitting discrete probability distributions, Technical Report No. 87-79, Educational Testing Service, Princeton.

Holland, P. W. and Thayer, D. T. (1989). The kernel method of equating score distributions. Research Report 89-7, Educational Testing Service, Princeion

Jarjoura, D. and Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent population design. *J. Eductional Statistics* **10**, 143-277.

Liou, M. and Cheng, P. E. (1995a). Equipercentile equating via data-imputation techniques. *Psychometrika* **60**, 119-136.

Liou, M. and Cheng, P. E. (1995b). Asymptotic standard error of equipercentile equating. *J. Ed. Behavioral Statist.* **20**, 259-286.

Little, R. J. A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bull. Internat. Statist. Inst.* **15**, 1-15.

Little, R. H. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley, New York.

Little, R. J. A. and Rubin, D. B. (1994). Test equating from biased samples, with applications to the Armed Services Vocational Aptitude Battery. *J. Ed. Behavioral Statist.* **19**, 309-335.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *J. Ed. Measurement* **30**, 23-29.

Marco, G. L., Abdel-fattah, A. A. and Baron, P. A. (1992). Methods used to establish score comparability on the enhanced ACT assessment and the SAT. College Board Report No. 92-3. Educational Testing Service, Princeton.

Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Amer. Statist. Assoc.* **86**, 899-909.

Rosenbaum, P. R. and Thayer, D. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British J. Math. Statist. Psych.* **40**, 43-49.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Rubin, D. B. and Thayer, D. (1978). Relating tests given to different samples. *Psychometrika* **43**, 3-10.

Rubin, D. B. and Schafer, J. L. (1988). Imputation strategies for estimating the undercount. Proceedings of the Fourth Annual Research Conference, Bureau of the Census, March, Virginia.

Rubin, D. B. (1995). Personal communication.

Thomasson, G. L., Bloxom, B. and Wise, L. (1994). Initial operational test and evaluation of forms 20, 21, and 22 of the Armed Service Vocational Aptitude Battery (ASVAB). DMDC Technical Report 94-001, Defense Manpower Data Center, Monterey.

Wright, N. K. and Dorans, N. J. (1993). Using the selection variable for matching or equation. Research Report 93-4, Educational Testing Service, Princeton.

Institute of Statistical Science, Academia Sinica, Taipei 115.

E-mail: mliou@stat.sinica.edu.tw

(Received May 1996; accepted December 1997)