Effect of Noncompensatory Multidimensionality on Separate and Concurrent Estimation in IRT Observed Score Equating.

A. A. Béguin
Citogroup
anton.beguin@citogroep.nl

B. A. Hanson
ACT, Inc.
Hanson@act.org

March 22, 2001

Abstract

In this article, the results of a simulation study comparing the performance of separate and concurrent estimation of a unidimensional item response theory (IRT) model applied to multidimensional noncompensatory data are reported. Data were simulated according to a two-dimensional noncompensatory IRT model for both equivalent and nonequivalent groups designs. The criteria used were the accuracy of estimating a distribution of observed scores, and the accuracy of IRT observed score equating. In general, unidimensional concurrent estimation resulted in lower or equivalent total error than separate estimation, although there were a few cases where separate estimation resulted in slightly less error than concurrent estimation. Estimates from the correctly specified multidimensional model generally resulted in less error than estimates from the unidimensional model. The results of this study, along with results from a previous study where data were simulated using a compensatory multidimensional model, make clear that multidimensionality of the data affects the relative performance of separate and concurrent estimation, although the degree to which the unidimensional model produces biased results with multidimensioanl data depends on the type of multidimensionality present.

Index terms: item response theory, noncompensatory multidimensional IRT, multidimensional equating, nonequivalent groups design, EPDIRM, BILOG-MG

The latent variable in unidimensional IRT (item response theory) models is unidentified up to a linear transformation. In each calibration, restrictions on the parameters are imposed to define the scale on which the parameters are measured. In a common item nonequivalent group design two forms of a test with some items in common are administered to samples from two populations. If item parameters for the two forms are estimated independently, the parameter estimates for the different forms will not be on the same scale. These estimates are brought on a common scale via minimization of some loss function. Techniques for this purpose have been developed by Haebara (1980), Marco (1977), Loyd and Hoover (1980) and Stocking and Lord (1983). An alternative procedure to obtain estimates on a common scale is concurrent estimation of multiple forms. Using a so-called marginal maximum likelihood (MML) procedure, the parameters of the IRT model are directly estimated on a common scale (Bock & Zimowski, 1996; Glas & Verhelst, 1989). Kiefer and Wolfowitz (1956) have shown that the MML estimator is strongly consistent under fairly reasonable regularity conditions. Therefore, in concurrent estimation standard asymptotic theory for confidence intervals and the distribution of statistics computed using MML estimates directly applies.

A number of studies have been carried out to compare the performance of concurrent and separate estimation (Hanson & Béguin, 1999; Kim & Cohen, 1998; Petersen, Cook & Stocking, 1983; Wingersky, Cook & Eignor, 1987). These studies used data that were simulated from the same unidimensional model also used for parameter estimation. With real data, the simple unidimensional model may not be appropriate and this could affect the performance of unidimensional separate and concurrent estimation. One source of misspecification is multidimensionality of the data. In this paper, the effect on performance of unidimensional separate and concurrent estimation will be studied for data that in fact follow a multidimensional noncompensatory IRT model (Ackerman, 1987; Embretson, 1980, 1984; Maris, 1993, 1995; Sympson, 1978; Spray, Davey, Reckase, Ackerman & Carlson, 1990).

Two classes of multidimensional IRT models for dichotomously scored items can be distinguished, compensatory and noncompensatory models. In compensatory multidimensional models (Lord & Novick, 1968; McDonald, 1967; Reckase, 1985 and Ackerman, 1996a and 1996b) the probability of a correct response is based on the sum of the proficiencies on the different dimensions. Consequently, a higher proficiency on one of the dimensions compensates for a lower proficiency on one of the other dimensions. In noncompensatory models (Ackerman, 1987; Embretson, 1980, 1984; Maris, 1993, 1995; Sympson, 1978; Spray, Davey, Reckase, Ackerman & Carlson, 1990) the probability of a correct response is based on a product of the proficiencies on the different dimensions. Consequently, a low proficiency on one of the dimensions can not be compensated with a high proficiency on one of the other dimensions.

Most of the research in multidimensional IRT has focused on the compensatory models. These models were first presented by Lord and Novick (1968) and McDonald (1967). These authors use a normal ogive to describe the probability of a correct response. McDonald (1967,1997) developed an estimation procedure based on an expression for the association between pairs of items derived from a polynomial expansion of the normal ogive. This procedure is implemented in NOHARM (Normal-Ogive Harmonic Analysis Robust Method, Fraser, 1988). An alternative using all information in the data, and therefore labeled "Full Information Factor Analysis", was developed by Bock, Gibbons, and Muraki, (1988). This approach is a generalization of the marginal maximum likelihood (MML) and Bayes modal estimation procedures for unidimensional IRT models (see, Bock & Aitkin, 1981, Mislevy, 1986), and has been implemented in TESTFACT (Wilson, Wood, and Gibbons, 1991). A comparable model using a logistic rather than a normal-ogive representation has been studied by Andersen (1985), Glas (1992), Reckase (1985, 1997) and Ackerman (1996a and 1996b).

Noncompensatory IRT models for dichotomous items were introduced by Sympson (1978). He proposed a multidimensional multiplicative generalization of the three-parameter logistic (3-PL) model (Birnbaum, 1968; Lord,1980). A multicomponent Rasch model was introduced by Embretson (1980, 1984). An estimation procedure for this model based on the EM algorithm (Dempster, Laird & Rubin, 1977) was developed by Maris (1993, 1995).

Considerable attention has been given to the effect of noncompensatory multidimensionality on parameter estimates of unidimensional IRT models. Ansley and Forsyth (1985) examined unidimensional estimates obtained from two-dimensional data generated using a noncompensatory model. They found that the unidimensional estimates of discrimination and proficiency parameters were highly related to the average over dimensions of their multidimensional counterparts. Ackerman (1987) compared the performance of unidimensional IRT estimates under two-dimensional compensatory- and non-compensatory models. He found similar patterns in the unidimensional estimates for both multidimensional models. Finally, Spray, Davey, Reckase, Ackerman and Carlson (1990) compared data generated under compensatory and noncompensatory models. They concluded that the models were indistinguishable from a practical standpoint.

A number of multidimensional equating procedures have been proposed. Hirsch (1989) proposed a procedure that calibrates the separate estimates of separate multidimensional two-parameter logistic models for the two forms in a common-examinee design on a common scale. Davey, Oshima and Lee (1996) proposed a procedure to calibrate the estimates of two multidimensional three-parameter models for the two forms in a common-item- or common-examinee-design on the same scale. Li and Lissitz (1998) used simulation studies to compare a number of different procedures to calibrate the parameters of multidimensional IRT models on the same scale. Bolt (1999) used simulation studies to investigate whether unidimensional IRT true-score equating is more adversely affected by the presence of multidimensionality than conventional linear- and equipercentile equating. He found that for correlations between dimensions equal to 0.7 or larger, IRT true-score equating performed slightly better than the conventional procedures. At lower correlations, IRT-equating performed almost as good as equipercentile equating. Finally, Béguin, Hanson and Glas (2000) compared the effect of multidimensionality on unidimensional IRT equating based on separate and concurrent estimation. They found that in some nonequivalent group conditions the error for both unidimensional equating methods was very large compared to the effect of multidimensional equating.

In this paper, the performance of separate and concurrent estimation of a unidimensional three-parameter logistic (3-PL) model (Birnbaum, 1968; Lord, 1980) applied to multidimensional data is compared. To obtain a benchmark to evaluate these unidimensional estimates, a two-dimensional noncompensatory normal-ogive model with guessing (labeled NCMP-PNO) is estimated. In this model, the probability of a correct response of a person i on an item j, denoted $Y_{ij} = 1$, is written as

$$P(Y_{ij} = 1; \theta_i, \alpha_j, \beta_j, \gamma_j) \;=\; \gamma_j + (1 - \gamma_j) \prod_{q=1}^{2} \Phi(\alpha_{jq}\theta_{iq} - \beta_{jq}) \qquad (1)$$

where Φ denotes the standard normal cumulative distribution function, γ_j is the guessing parameter, β_{jq} is the difficulty parameter on the q^{th} dimension, θ_{iq} is the proficiency of person i on dimension q, and α_{jq} is the discrimination parameter of item j.

The NCMP-PNO model will be estimated by an adapted version of a Markov Chain Monte Carlo (MCMC) estimation procedures (Béguin, 2000, Béguin & Glas, in press) for a multidimensional compensatory IRT model. This procedure is a generalization to incomplete designs of procedures that use Gibbs sampling (Gelfand & Smith, 1990) with data-augmentation to estimate models in the normal ogive context.

Using these procedures the posterior number correct score distribution is easily obtained by sampling response patterns during each iteration of the Gibbs sampler. These response patterns are simulated based on the probability of a correct response given the values of the parameters in the current iteration of the Gibbs sampler. A nice property of this procedure is that the uncertainty of the parameter estimates is taken into account in the estimation of the number-correct observed score distribution.

Data

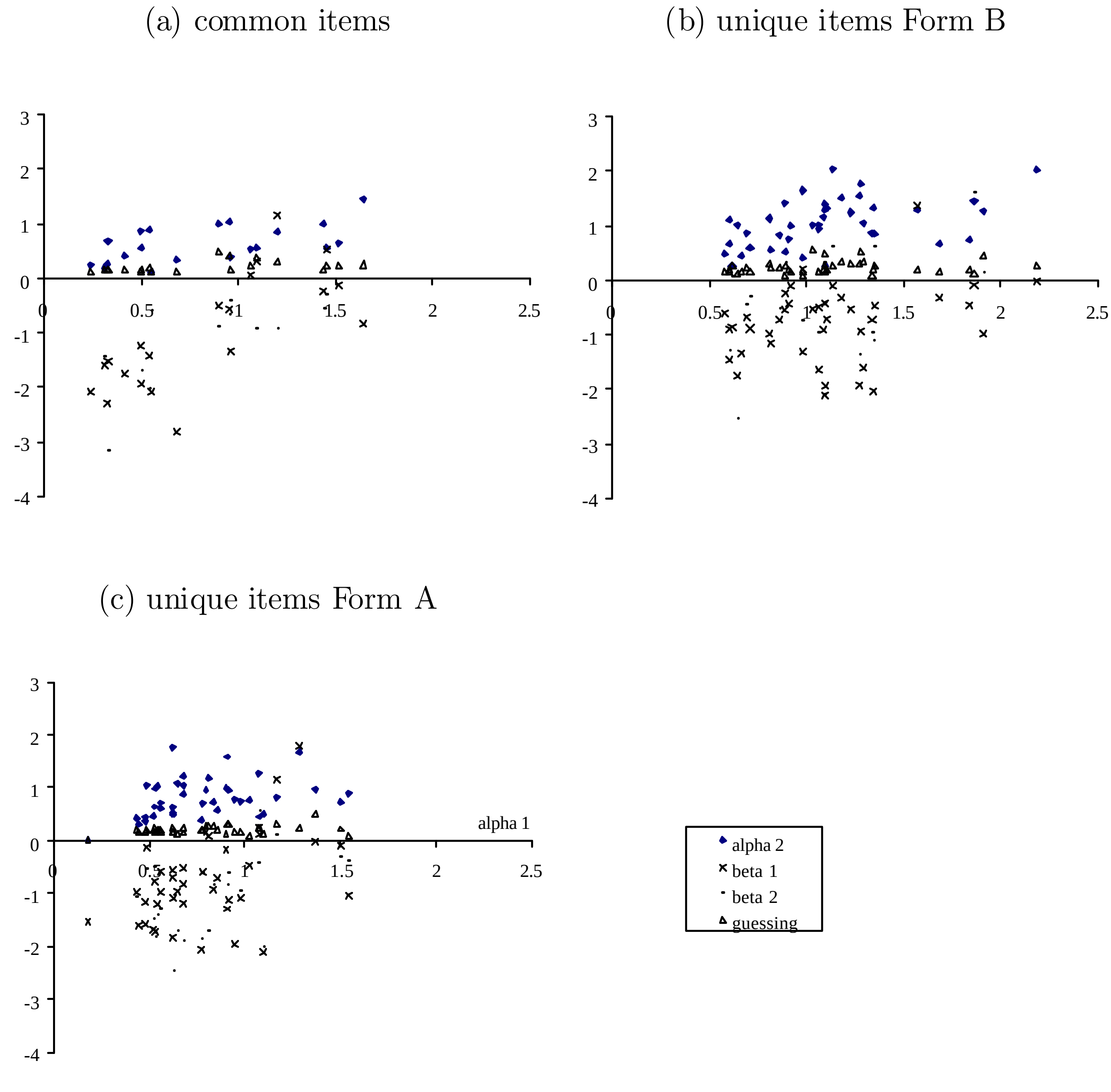
To simulate data with realistic properties item parameter estimates of the NCMP-PNO model obtained on data from examinations at the end of secondary education in the Netherlands will be used to simulate data. The original data used in this study consist of examinations in language comprehension.

Two forms of three different examinations were used: 1) two forms of the examination '*language comprehension in English at MAVO level*' for the years 1993 and 1999, 2) two forms of the examination '*language comprehension in German at MAVO level*' for the years 1995 and 1999, and 3) two forms of the examination '*language comprehension in French at MAVO level*' for the years 1995 and 1999. These forms and examinations were selected from a larger pool of forms and examinations in such a way that these examinations represent realistic conditions with a different amount of correlation between the latent proficiencies of the NCMP-PNO model. Each form of each examination contained 50 dichotomously scored items. The original two forms of each examination had no items in common, but additional data from persons who responded to items from both forms were available. The data collection design is beyond the scope of this article, for a detailed description of refer to Glas and Béguin (1996) or Béguin (2000). As mentioned above, the NCMP-PNO model item parameter estimates for the items on each examination are obtained using a two-dimensional MCMC estimation procedure. In this estimation procedure, the item parameters are estimated under the assumption of different proficiency distributions for the two groups in the design. So this procedure can be labeled a multiple-group concurrent estimation procedure. The correlation between the two latent proficiencies for the English, German, and French examinations were 0.0, 0.3 and 0.5, respectively.

To simulate data according to a common-item nonequivalent group design for each examination 10 items were randomly selected from each of the two forms. These 20 items were used as common items in two test forms, say A and B, constructed from items on the original two forms. Form A was created using the 20 selected common items and the 40 remaining items from one of the original forms. Form B contained the 20 common items and the 40 remaining items from the other original form. So, Form A contained all 50 items from the oldest form, the original 1993 or 1995 form, and 10 items from the original 1999 form. Form B contained 10 items from the original 1993 or 1995 form and all 50 items from the original 1999 form. To give an illustration of the item parameters used for generating the data, the parameter values for the examinations in French language comprehension are given in Figure 1. The values of the discrimination parameters on the second dimension, α_2, the difficulty parameters β_1 and β_2, and the guessing parameter γ are plotted against the value of the discrimination parameter on the first dimension, α_1.

Method

Samples of item responses for forms A and B for each of the examinations are generated under two different conditions. These conditions differ in the mean vectors of the bivariate normal proficiency distributions for the two populations taking Forms A and B. The mean proficiency on the first dimension for the population taking Form A is 0.0 in all conditions while the mean proficiency on the first dimension for the population taking Form B is either 0.0 or 0.5. The mean proficiency for the second dimension is 0.0 in all conditions. Combining the two levels of mean proficiency difference with the three examinations produced six study conditions. Table 1 contains a summary of the conditions. Conditions 1 and 4 use the English forms with correlation between the dimensions of 0.0, conditions 2 and 5 use the German


(a) common items

(b) unique items Form B

(c) unique items Form A

Figure 1. Parameter values

Table 1. Overview of the conditions

Condition	Mean Proficiency		Covariance
	Form A	Form B	Both Forms
1	(0,0)	(0,0)	$\begin{pmatrix} 1 & \\ 0.0 & 1 \end{pmatrix}$
2	(0,0)	(0,0)	$\begin{pmatrix} 1 & \\ 0.3 & 1 \end{pmatrix}$
3	(0,0)	(0,0)	$\begin{pmatrix} 1 & \\ 0.5 & 1 \end{pmatrix}$
4	(0,0)	(0.5,0)	$\begin{pmatrix} 1 & \\ 0.0 & 1 \end{pmatrix}$
5	(0,0)	(0.5,0)	$\begin{pmatrix} 1 & \\ 0.3 & 1 \end{pmatrix}$
6	(0,0)	(0.5,0)	$\begin{pmatrix} 1 & \\ 0.5 & 1 \end{pmatrix}$

forms with correlation between the dimensions of 0.3, and conditions 3 and 6 use the French forms with correlation between the dimension of 0.5. The first three conditions can be considered equivalent groups conditions, since the proficiency distributions of the populations administered Form A and B are the same. The last three conditions are nonequivalent group conditions. The conditions will be identified using the examination and an indication of whether the groups are equivalent or nonequivalent. For example, condition 5 in Table 1 will be referred to as the nonequivalent condition for the German examination.

Estimation of the parameters

For each condition, 20 samples of both forms were generated with 2000 persons per form. Two unidimensional estimation programs were used, BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1996) and EPDIRM (Hanson, 2000). In each sample and each condition, three sets of parameter estimates were obtained using BILOG-

MG and EPDIRM, two sets for each separate form (separate estimation) and one for both forms simultaneously (concurrent estimation). Also for each sample the NCMP-PNO model was estimated.

For both BILOG-MG and EPDIRM, normal population distributions of the latent variable were assumed. In the nonequivalent groups conditions the mean and standard deviation of the normal distribution for the group taking Form B was estimated. Default priors were used for the a and b parameter in BILOG-MG. The prior used for the a parameters is a lognormal distribution with a mean of 0 and a standard deviation of 0.5. The prior used for the c parameters is a beta distribution with parameters 6 and 16. To support convergence an additional $N(0, 2)$ prior distribution was used on the b parameter. In the nonequivalent groups conditions the priors are updated at each iteration, so priors used in the final iteration will be somewhat different from the initial priors. Appendix A gives the BILOG-MG control files used to obtain parameter estimates for each simulated sample.

Default four–parameter beta priors were used for the a, b, and c parameters in EPDIRM. The priors used for the a, b, and c parameters were Beta(1.75, 3, 0, 3), Beta(1.01, 1.01, -6, 6), and Beta(3.5, 4.0, 0, 0.5), respectively, where Beta(p, q, l, u) represents a four–parameter beta prior with shape parameters p and q, lower limit l, and upper limit u. The item parameter priors used in EPDIRM are less informative than the item parameter priors used in BILOG-MG. Appendix B gives the EPDIRM control files used to obtain parameter estimates for the simulated samples.

The MCMC procedure consisted of 3000 iterations with a burn-in period of 1000 iterations. Results of Albert (1992) show that this is sufficient. As starting values for the NCMP-PNO model, the true parameters were used for the item parameters, and $\boldsymbol{\theta} = \mathbf{0}$ was used for the proficiency of each simulee. The priors on the item parameters were $\alpha \sim N(1., 0.5)$, $\beta \sim N(-1, 1)$ and $\gamma \sim \text{Beta}(20 * \gamma_{true}, 20 * (1 - \gamma_{true}))$.

In the separate estimation conditions, the parameters of Form A and Form B had to be calibrated on a common scale. This was done with the Stocking and Lord (1983) method (see also Kolen & Brennan, 1995) which was among the best performing methods in the comparison by Hanson & Béguin (1999). The three conditions where the groups of simulees that were administered Form A and B had equal proficiency distributions (conditions 1 through 3 in Table 1) can be considered equivalent groups conditions. In an equivalent groups design, it is not necessary to assume different population distributions for the groups taking Form A and Form B. Consequently, in the condition were the parameters are estimated separately for the two forms, no linking is necessary to bring the two sets of estimates on a common scale. Analogously, one can assume a single population distribution for both samples when concurrent estimation is applied. In this study, a single population distribution was assumed for the estimation of the unidimensional models in the conditions where the populations administered Form A and B had equal proficiency distributions. Consequently, in separate estimation, no scaling was performed and in concurrent estimation using BILOG-MG and EPDIRM, a single group was specified. In the NCMP-PNO model, different population distributions were estimated due to the current limitations of the available software. Because Hanson and Béguin (1999) found indications that separate estimation with scaling improved performance in equivalent group conditions, separate estimation with scaling was also performed in conditions where the populations administered Form A and B had equal proficiency distributions.

In the separate estimation conditions, two sets of item parameter estimates for the common items are available. In this study, the Form A item parameter estimates were used as the parameter estimates of the common items for the purpose of computing the criteria used to evaluate the quality of item parameter scaling. An alternative would be using the average of the item parameter estimates obtained on the two forms (Kim & Cohen, 1998).

Evaluation of scaling

To evaluate the quality of item parameter scaling, differences in results of equating scores on Form B to scores on Form A were assessed. Two criteria based on IRT observed-score (OS) equating of number-correct (NC) scores (Zeng & Kolen, 1995) were used. This technique uses the estimated number correct score distributions of both forms in one population. Here, the score distributions of Forms A and B were estimated for the population taking Form A.

Estimating score distributions

Using the estimated item and population parameters, the compound binomial distribution was used to generate the score distribution of a simulee with multidimensional proficiency $\boldsymbol{\theta}$. The score distribution for the simulees administered Form A, say a sample from a population A with a multivariate normal ability distribution having mean $\boldsymbol{\mu}_A$ and covariance matrix $\boldsymbol{\Sigma}_A$, can be computed by integrating over the population distribution of $\boldsymbol{\theta}$, that is,

$$f(r) = \int \cdots \int \sum_{\{x|r\}} f(x\,|\boldsymbol{\theta})g(\boldsymbol{\theta} \mid \boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A)d\boldsymbol{\theta}, \tag{2}$$

where $\{x\,|r\}$ stands for the set of all possible response patterns resulting in a score r, and $f(x\,|\boldsymbol{\theta})$ is the probability of item response pattern x given latent proficiency vector $\boldsymbol{\theta}$. In the case of normal distributed populations, the integral can be computed using Gauss-Hermite quadrature (Abramowitz & Stegun, 1972). At each of the quadrature points, a recursion formula by Lord and Wingersky (1984) can be used to obtain $\sum_{\{x|r\}} f(x\,|\boldsymbol{\theta})$, the probability of obtaining number correct score r given proficiency $\boldsymbol{\theta}$. To obtain accurate results, 180 quadrature points were used in the unidimensional case and 100 quadrature points were used for each dimension in the multidimensional case.

In the conditions where an MCMC estimation procedure was used, the score distribution was estimated as follows. After the burn-in period for the Gibbs-sampler, after every 20 iterations, the procedure by Lord and Wingersky was applied with the currently drawn values of the person and item parameters. The estimated score distribution was the mean over 100 thus obtained score distributions. A nice property of this procedure is that the uncertainty of the parameter estimates is taken into account in the estimation of the score distribution.

In the conditions where a unidimensional model was used, the observed score distributions needed for the criteria described in the next section were calculated with Guass–Hermite quadrature using a univariate standard normal distribution.

Criteria

To evaluate the equating precision in the 6 conditions the following two criteria were used. The first criterion was based on the differences between the estimated and true observed score distributions on Form B for the population administered Form A, where the true distribution is the distribution under the model used to generate the data. The second criterion was based on comparing equivalent score points from the observed score equating function with the true equivalent score points based on the model used to generate the data for the population that took Form A. The evaluation of the score distributions served two purposes. On one hand, comparison of score distributions provided an evaluation of model fit. On the other hand, it provided insight into the quality of the equating process, since the score distributions play a crucial role in IRT number-correct equating.

Let $f_{true,r}$ be the expected frequency of score point r on Form B for a sample of examinees from the population administered Form A as computed using the parameters of NCMP-PNO model from which the data were generated. Let f_{hr} be the frequency of score point r on Form B for the population that took Form A as estimated using item parameter estimates from replication h. To compare the score distributions, the mean over score points of the mean squared error (MSE) was calculated by summing over the 20 samples and the $k+1$ score points, that is,

$$MSE = \frac{1}{20(k+1)} \sum_{h=1}^{20} \sum_{r=0}^{k} (f_{hr} - f_{true,r})^2. \tag{3}$$

The MSE can be decomposed into a term representing the mean over score points of the squared bias (mean bias) and a term representing the mean over score points of the variance (mean variance):

$$MSE = \frac{1}{k+1}\sum_{r=0}^{k}(\overline{f_r} - f_{true,r})^2 + \frac{1}{20(k+1)}\sum_{h=1}^{20}\sum_{r=0}^{k}(f_{hr} - \overline{f_r})^2, \tag{4}$$

where $\overline{f_r}$ is the mean over replications, that is,

$$\overline{f_r} = \frac{1}{20}\sum_{h=1}^{20} f_{hr}. \tag{5}$$

A measure of model fit can be obtained if the terms of (3) are divided by the true frequency. This results in the test-statistic

$$X^2 = \frac{1}{20(k+1)}\sum_{h=1}^{20}\sum_{r=0}^{k}\frac{(f_{hr} - f_{true,r})^2}{f_{true,r}}. \tag{6}$$

Although, the distribution of this statistic in the present application is unknown (Glas & Verhelst, 1989), the values provide an –admittedly fallible– basis for comparison.

For the second criterion, equivalent score points of Form B equated to Form A estimated using various models were compared with the equivalent score points obtained with the true model. Let $s_{true,r}$ be the integer score point on Form A that is equivalent with the score point r on Form B, based on the rounded IRT observed score equating function computed using the true item parameters and the true latent proficiency distribution for the group taking Form A. Let s_{hr} be an analogous score point estimated in replication h. Furthermore, let $p_{r,true}$ be the probability in the population taking Form A of obtaining a score r on Form B based on the true parameters values. To compare the equivalent score points, a weighted mean squared error (WMSE) was calculated by summing over samples and score points. The score points were weighted by $p_{h,true}$, which resulted in

$$WMSE = \frac{1}{20}\sum_{r=0}^{k} p_{r,true}\sum_{h=1}^{20}(s_{hr} - s_{true,r})^2. \tag{7}$$

The WMSE can be decomposed into terms representing the weighted sum of the squared bias (weighted bias) of equated score points and weighted sum of the variance (weighted variance) of the equated score points, so,

$$WMSE = \sum_{r=0}^{k} p_{r,true}(\overline{s_r} - s_{true,r})^2 + \frac{1}{20}\sum_{r=0}^{k} p_{r,true}\sum_{h=1}^{20}(s_{hr} - \overline{s_r})^2, \tag{8}$$

where $\overline{s_r}$ is the mean equivalent score of score point r over replications, that is,

$$\overline{s_r} = \frac{1}{20}\sum_{h=1}^{20} s_{hr}. \tag{9}$$

The weighted mean absolute error (WMAE) is obtained if the squared error in (7) is replaced by the absolute value of the error, so

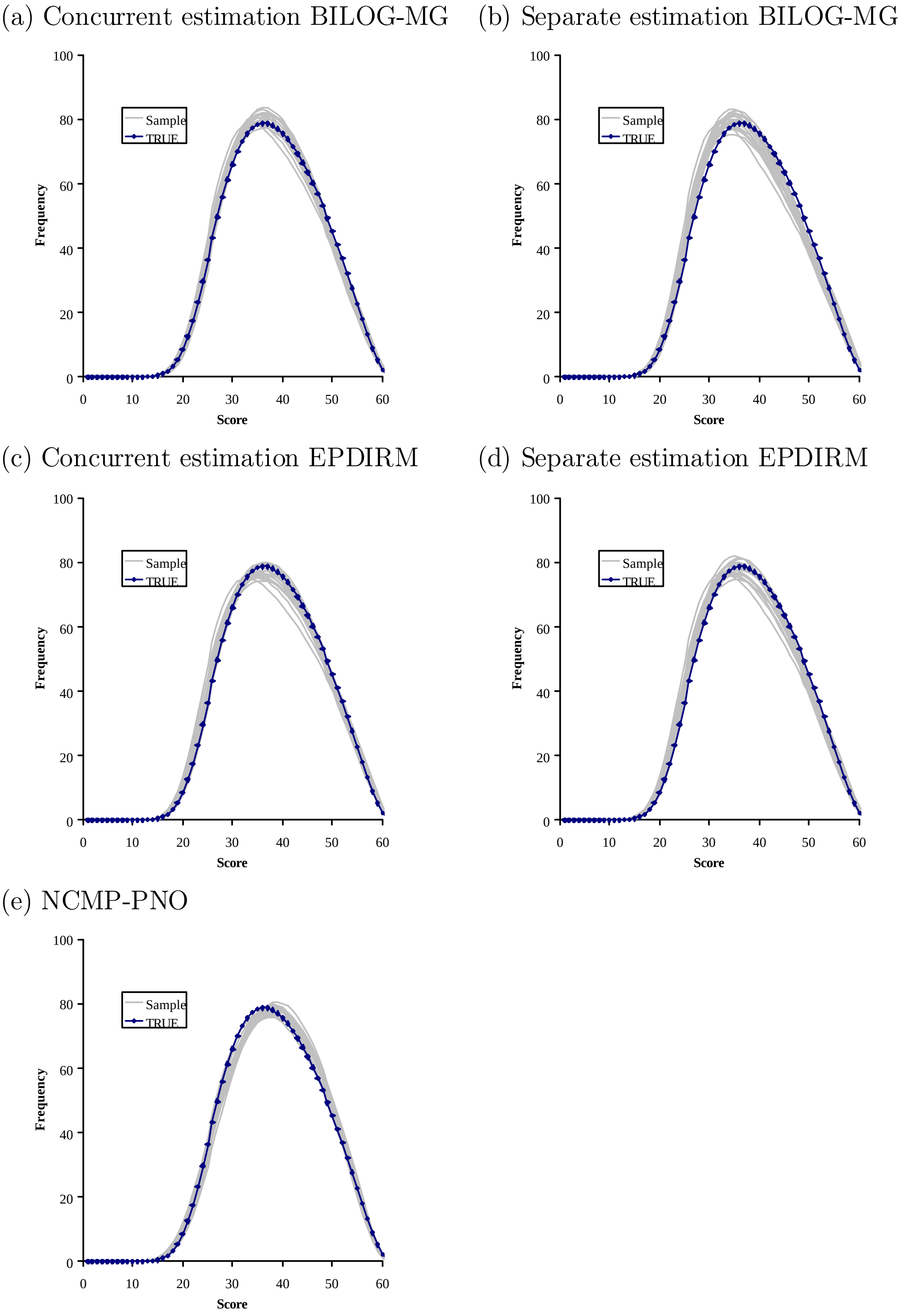
$$WMAE = \frac{1}{20}\sum_{r=0}^{k} p_{r,true} \sum_{h=1}^{20} |s_{hr} - s_{true,r}| \,. \tag{10}$$

Results

Three factors are investigated in this study 1) concurrent versus separate estimation 2) EPDIRM versus BILOG-MG concurrent estimation 3) unidimensional versus multidimensional noncompensatory estimation. All BILOG-MG, EPDIRM, and MCMC runs converged except for 4 Form B data sets in the German examination nonequivalent groups condition (separate estimation) for which BILOG-MG did not converge. Convergence was achieved for these four data sets by re–running BILOG-MG with the number of Newton steps set to zero.

First, the true and estimated frequency distributions of Form B were compared. To illustrate the results, the estimated frequency distributions for the French examination for nonequivalent groups using the NCMP-PNO model, BILOG-MG and EPDIRM are plotted in Figure 2. The frequency distribution obtained using the true model used to generate the data is plotted together with the estimated frequency distributions of the 20 samples. In Figure 2 it can be seen that the unidimensional estimation procedures show a larger variation between the score distributions of the different samples than the score distributions obtained using the NCMP-PNO model. The scores in the samples obtained using the unidimensional estimation procedures are in general somewhat lower than the scores obtained using the true values. This effect is stronger in the separate estimation conditions (Figure 2b and 2d), especially when based on the BILOG-MG estimates. In some of the samples the score distributions based on the EPDIRM estimates (Figure 2c and 2d) show a larger deviation from the true score distribution. Finally, in Figure 2e it can be seen that the scores in the samples estimated using the NCMP-PNO model are somewhat higher than the true score distribution.

Table 2 gives the mean squared error, squared bias and variance for the estimated Form B distributions, along with the value of the X^2–statistic, for the various conditions and estimation methods. The first three columns of Table 2 identify the combination of model and program, condition, and equating method for which results are presented. The first column gives the model and program used for estimation. The second column identifies the study condition by giving the examination followed by a 0 or 5, where 0 means equivalent groups (first dimension mean of 0.0 for the



(a) Concurrent estimation BILOG-MG

(b) Separate estimation BILOG-MG

(c) Concurrent estimation EPDIRM

(d) Separate estimation EPDIRM

(e) NCMP-PNO

Figure 2. Score distributions for Form B for the French Examination with nonequivalent groups condition, determined using the true proficiency distribution of the population administered Form A.

population taking Form B), and 5 for nonequivalent groups (first dimension mean of 0.5 or the group population Form B). The third column gives the equating method — "sepNS" for separate estimation with no item parameter scaling (only for the equivalent groups conditions), "sep" for separate estimation with item parameter scaling using the Stocking-Lord method, and "con" for concurrent estimation. The mean squared error, bias, and variance presented in Table 2 are plotted in Figure 3. The top two plots in Figure 3 present the MSE results from Table 2 for the equivalent and nonequivalent groups cases, respectively. The middle two plots present squared bias, and the bottom two plots present variance.

The performance of the unidimensional and multidimensional models differ in the equivalent and nonequivalent groups conditions. In the equivalent groups conditions the NCMP-PNO model resulted in a MSE and squared bias that were in general smaller than the MSE and squared bias obtained using unidimensional models for the English and German examinations. The unidimensional model using concurrent estimation and separate estimation with no scaling resulted in a somewhat lower MSE and squared bias for the French examination. The variance obtained with unidimensional concurrent estimation or with separate estimation with no scaling was smaller than the variance obtained with the NCMP-PNO model. Comparing separate and concurrent estimation in the equivalent groups conditions both BILOG-MG and EPDIRM resulted in a MSE, squared bias and variance that were in general smaller for the concurrent estimation method than for the separate estimation method. The only exception occurred for the French examination where the separate estimation method with no scaling using EPDIRM resulted in the same MSE as the concurrent estimation method. The separate estimation method with no scaling had a slightly lower variance but this effect was offset by a slightly higher squared bias than in the concurrent estimation method. Separate estimation with no scaling resulted in a lower MSE, squared bias and variance than separate estimation with scaling.

Comparing separate and concurrent estimation in the nonequivalent groups conditions both BILOG-MG and EPDIRM resulted in a MSE, squared bias and variance that were in general smaller for the concurrent estimation method than for the separate estimation method. The exceptions occurred for the English and German examinations when the EPDIRM program was used. For the English examination the separate estimation resulted in a lower MSE and squared bias. For the German examination only the squared bias was lower. The NCMP-PNO model performed better than the other estimation procedures for the English and German examinations. For the French examination both EPDIRM and BILOG-MG concurrent estimation procedures performed better than NCMP-PNO.

With respect to the relative performance of BILOG-MG and EPDIRM, in con-

Table 2. Mean squared error of estimated Form B distribution

Model/Program	Condition	Equating	MSE	Bias	Variance	X^2
EPDIRM	English–0	sepNS	5.8	5.2	0.6	349.
EPDIRM	English–0	sep	7.1	5.4	1.8	392.
EPDIRM	English–0	con	5.6	5.1	0.5	339.
BILOG-MG	English–0	sepNS	10.6	10.0	0.6	589.
BILOG-MG	English–0	sep	11.7	10.0	1.7	627.
BILOG-MG	English–0	con	9.8	9.2	0.5	547.
NCMP-PNO	English–0	con	1.5	0.6	1.0	68.
EPDIRM	German–0	sepNS	4.0	3.2	0.8	323.
EPDIRM	German–0	sep	6.0	3.3	2.7	387.
EPDIRM	German–0	con	3.5	2.8	0.7	279.
BILOG-MG	German–0	sepNS	8.0	7.2	0.8	601.
BILOG-MG	German–0	sep	10.0	7.2	2.8	671.
BILOG-MG	German–0	con	6.8	6.1	0.7	497.
NCMP-PNO	German–0	con	1.9	0.5	1.4	78.
EPDIRM	French–0	sepNS	1.4	0.4	1.0	88.
EPDIRM	French–0	sep	6.5	2.8	3.7	367.
EPDIRM	French–0	con	1.4	0.3	1.1	79.
BILOG-MG	French–0	sepNS	3.0	1.9	1.0	151.
BILOG-MG	French–0	sep	7.6	4.3	3.4	424.
BILOG-MG	French–0	con	2.6	1.6	1.0	128.
NCMP-PNO	French–0	con	5.0	3.3	1.6	207.
EPDIRM	English–5	sep	6.0	4.4	1.5	499.
EPDIRM	English–5	con	7.0	5.6	1.4	616.
BILOG-MG	English–5	sep	9.5	7.9	1.5	754.
BILOG-MG	English–5	con	4.6	3.1	1.5	380.
NCMP-PNO	English–5	con	1.6	0.3	1.3	76.
EPDIRM	German–5	sep	6.5	3.9	2.6	312.
EPDIRM	German–5	con	6.4	4.1	2.3	379.
BILOG-MG	German–5	sep	10.9	8.4	2.5	588.
BILOG-MG	German–5	con	4.6	2.3	2.3	217.
NCMP-PNO	German–5	con	3.6	1.7	1.9	135.
EPDIRM	French–5	sep	9.2	6.2	3.0	345.
EPDIRM	French–5	con	5.9	3.1	2.8	245.
BILOG-MG	French–5	sep	13.1	10.0	3.2	409.
BILOG-MG	French–5	con	4.5	2.0	2.5	143.
NCMP-PNO	French–5	con	6.9	5.2	1.8	247.

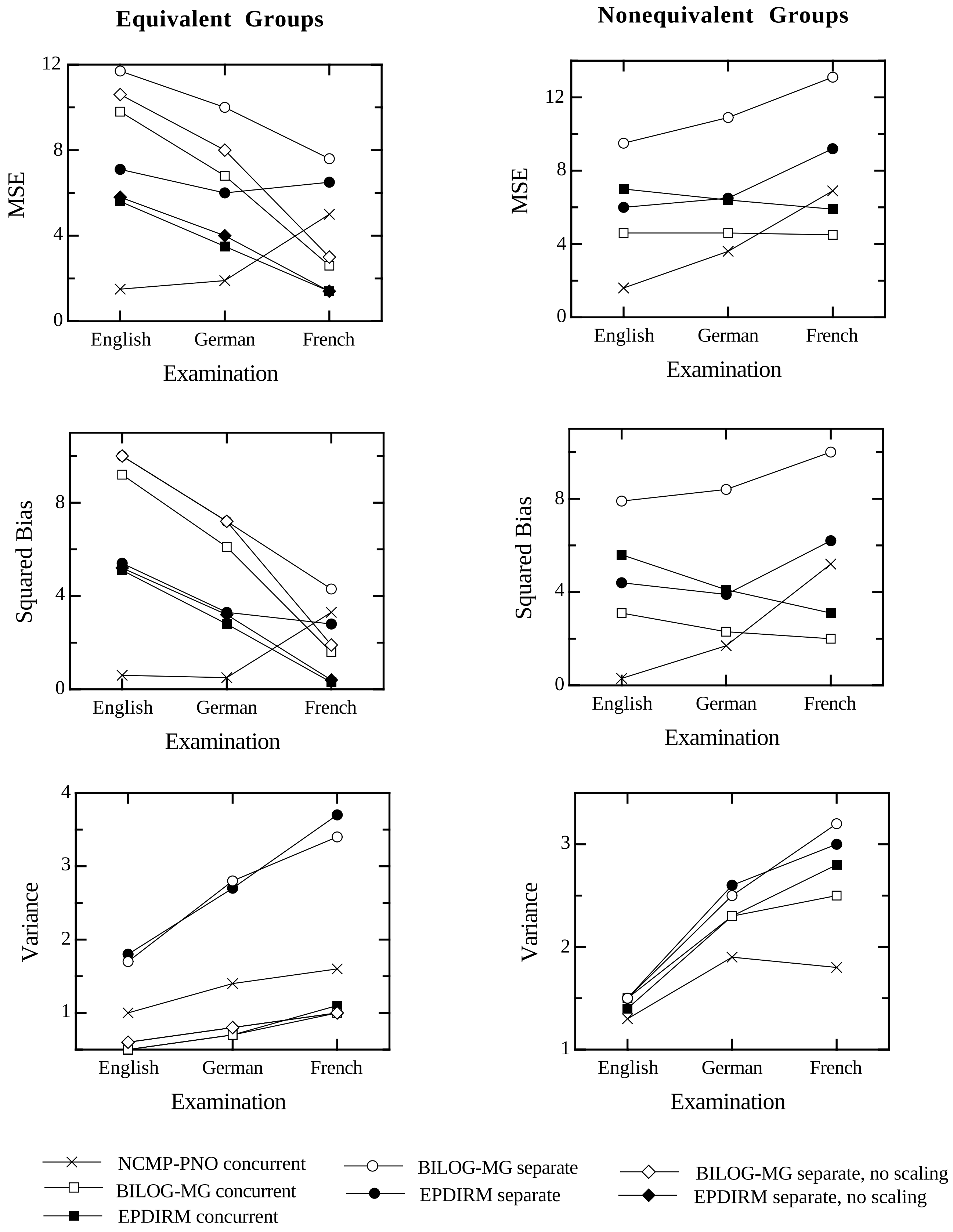


Figure 3. Average Error of Estimated Form B Distributions.

current estimation BILOG-MG performed better than EPDIRM, while in separate estimation EPDIRM performed better.

Comparing the X^2 for the different conditions and estimation methods, within conditions the pattern of results was similar to those for MSE. In general the unidimensional concurrent estimation procedures results in lower X^2 than the separate procedures. Exceptions are found with the EPDIRM program in the nonequivalent groups condition for the English and German examinations. The NCMP-PNO model performs best in all conditions for the English and German examinations, but worse than the unidimensional model using concurrent estimation for the French examination. Note that it is unclear how these results should be interpreted, since the distribution of the X^2 statistic is unknown and can differ over conditions. The X^2 is calculated by summation over 1,220 cells. However the degrees of freedom of each X^2 value is unknown.

Table 3 gives the weighted mean squared error, weighted squared bias, and weighted variance for the estimated equating functions of Form B scores to equivalent Form A scores, along with the WMAE, for the various conditions and estimation methods. The weighted mean squared error, weighted squared bias, and weighted variance presented in Table 3 are plotted in Figure 4.

The differences in the weighted variance that resulted from different estimation procedures and examinations were relatively small in the equivalent groups conditions. In the nonequivalent conditions the weighted variance differed some among examinations with French having the highest weighted variance, and English the lowest weighted variance. Larger differences in the results among the conditions and the estimation procedures occurred in the weighted squared bias, and consequently also in the WMSE. For the equivalent groups conditions, the NCMP-PNO model had the lowest WMSE across conditions for each examination. All the univariate procedures resulted in similar levels of weighted squared bias and WMSE for the English and German examinations. In these conditions the differences between EPDIRM and BILOG-MG procedures were small. For the French examination concurrent estimation resulted in the lowest squared bias and WMSE, and separate estimation with scaling resulted in the highest squared bias and WMSE for the unidimensional model.

In the nonequivalent groups results reported in Table 3, the NCMP-PNO model had the lowest WMSE except for the English examination were the unidimensional concurrent estimation procedures performed better. The unidimensional concurrent estimation procedures performed better than the separate estimation procedures. The EPDIRM concurrent estimation procedure performed slightly better than its BILOG-MG counterpart for the English examination, while EPDIRM performed somewhat worse than BILOG-MG for the French examination. Finally, the WMSE's

Table 3. Weighted error of equated scores

Model/Program	Condition	Equating	WMSE	Bias	Variance	WMAE
EPDIRM	English–0	sepNS	0.52	0.36	0.16	0.52
EPDIRM	English–0	sep	0.55	0.40	0.15	0.55
EPDIRM	English–0	con	0.56	0.43	0.13	0.56
BILOG-MG	English–0	sepNS	0.53	0.37	0.16	0.52
BILOG-MG	English–0	sep	0.54	0.39	0.15	0.54
BILOG-MG	English–0	con	0.54	0.41	0.13	0.54
NCMP-PNO	English–0	con	0.22	0.07	0.14	0.22
EPDIRM	German–0	sepNS	0.36	0.25	0.11	0.36
EPDIRM	German–0	sep	0.40	0.27	0.13	0.40
EPDIRM	German–0	con	0.37	0.26	0.11	0.37
BILOG-MG	German–0	sepNS	0.37	0.26	0.11	0.37
BILOG-MG	German–0	sep	0.40	0.27	0.13	0.40
BILOG-MG	German–0	con	0.37	0.25	0.12	0.37
NCMP-PNO	German–0	con	0.23	0.09	0.14	0.23
EPDIRM	French–0	sepNS	0.43	0.32	0.11	0.43
EPDIRM	French–0	sep	0.54	0.37	0.17	0.52
EPDIRM	French–0	con	0.40	0.30	0.10	0.40
BILOG-MG	French–0	sepNS	0.42	0.32	0.10	0.41
BILOG-MG	French–0	sep	0.50	0.34	0.16	0.49
BILOG-MG	French–0	con	0.35	0.30	0.05	0.35
NCMP-PNO	French–0	con	0.23	0.09	0.14	0.23
EPDIRM	English–5	sep	0.21	0.10	0.11	0.21
EPDIRM	English–5	con	0.16	0.06	0.10	0.16
BILOG-MG	English–5	sep	0.20	0.10	0.11	0.20
BILOG-MG	English–5	con	0.19	0.09	0.10	0.19
NCMP-PNO	English–5	con	0.21	0.07	0.14	0.21
EPDIRM	German–5	sep	0.30	0.16	0.14	0.30
EPDIRM	German–5	con	0.27	0.14	0.13	0.27
BILOG-MG	German–5	sep	0.29	0.15	0.14	0.29
BILOG-MG	German–5	con	0.27	0.14	0.13	0.27
NCMP-PNO	German–5	con	0.25	0.12	0.12	0.25
EPDIRM	French–5	sep	1.24	1.10	0.14	1.01
EPDIRM	French–5	con	0.97	0.81	0.16	0.85
BILOG-MG	French–5	sep	1.11	0.95	0.16	0.93
BILOG-MG	French–5	con	0.73	0.57	0.16	0.67
NCMP-PNO	French–5	con	0.33	0.16	0.17	0.33

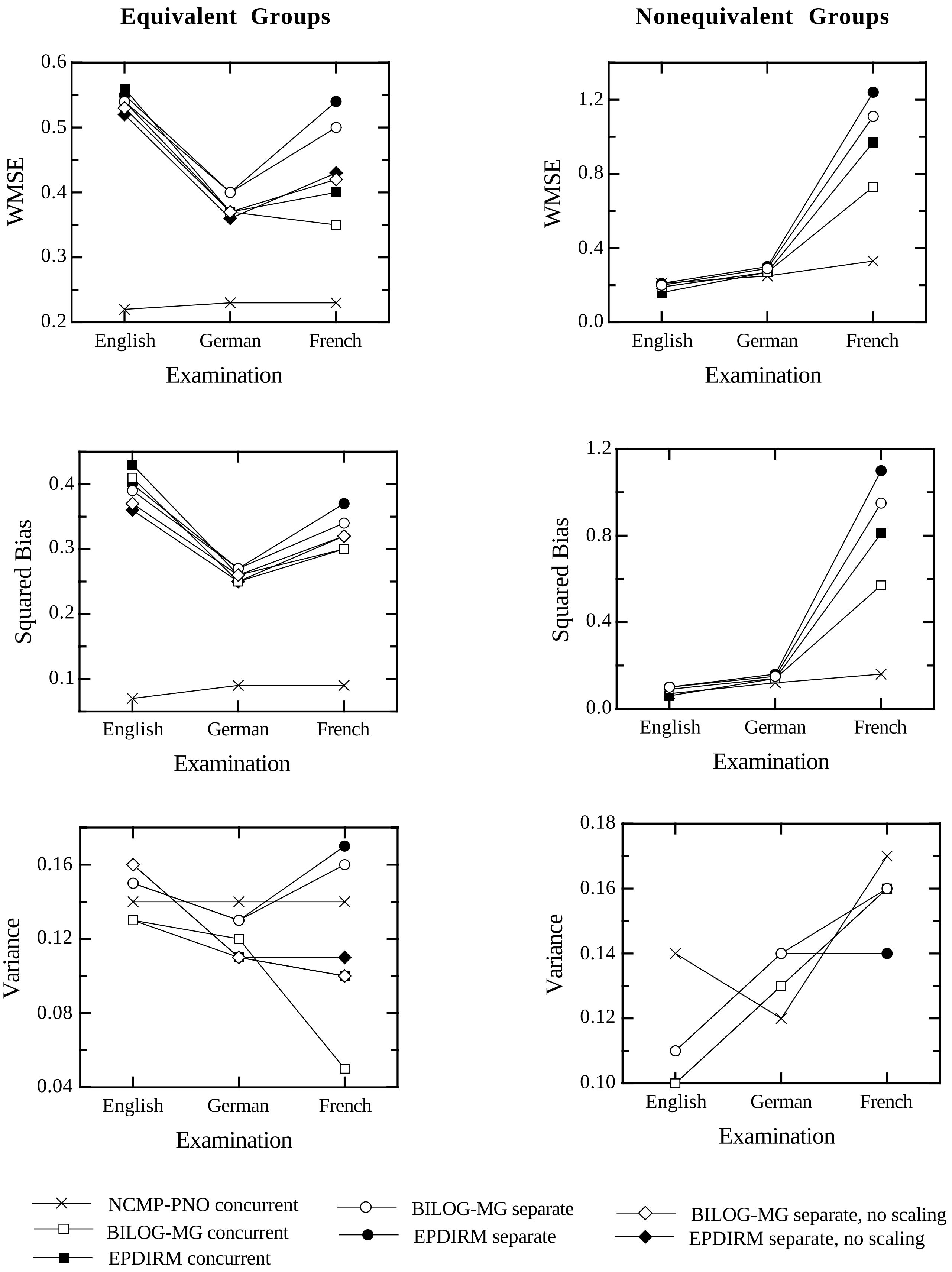


Figure 4. Average Weighted Error of Estimated Equating Functions.

were much higher for the French examination than the other two examinations. The WMSE's for the German examination were slightly higher than those for the English examination.

Conclusions

In this study, the effect of the estimation method on equating results was compared for the case where unidimensional models were applied to multidimensional noncompensatory data. As with any simulation study, considerable caution in drawing conclusions should be taken, due to the small number of conditions investigated. In the present study, the results pertain only to the six different conditions used. The only aspects varied in the conditions were the use of three different sets of forms, with different correlations between the two dimensions, and the difference between the mean of the proficiency distributions of the populations. There was no variation in data collection designs or the number of respondents in the design.

For the unidimensional model concurrent estimation generally resulted in lower or equivalent total error than separate estimation, although there were a few cases where separate estimation resulted in slightly less error than concurrent estimation. These results are consistent with the results in Béguin, Hanson and Glas (2000) which simulated data from a compensatory multidimensional model, and Hanson and Béguin (1999) where data were simulated from a unidimensional model. Comparing the two estimation programs, there tended to be larger differences between the total error in concurrent and separate estimation for BILOG-MG than for EPDIRM.

Separate estimation without scaling resulted in similar or better performance than separate estimation with scaling. This result is in line with the results reported in Béguin, Hanson and Glas (2000), but differs from the results of Hanson and Béguin (1999). This suggests that computing a scaling transformation in the case of equivalent groups is beneficial when a unidimensional model is correctly specified, but can be detrimental when a unidimensional model is used with multidimensional data.

In general, the EPDIRM and BILOG-MG programs produced similar results, although there some cases where there were systematic differences between the two programs. For the Form B distribution criterion the MSE was smaller for EPDIRM than BILOG–MG in all equivalent groups conditions, although for the nonequivalent groups conditions BILOG–MG had smaller MSE than EPDIRM when concurrent estimation was used, but not when separate estimation was used. For the equating criterion EPDIRM and BILOG-MG had similar WMSE over all conditions for the English and German examinations. For the French examination BILOG-MG resulted in lower WMSE across all conditions. The differences between the results using BILOG-MG and EPDIRM may be at least partly due to the different priors that were used by the two programs. For both programs default priors were used, except for separate estimation with BILOG-MG in the nonequivalent groups cases, where a standard normal prior was put on the b parameters rather than the default of no prior. The priors used in EPDIRM were generally less informative than the priors used in BILOG-MG. For concurrent estimation in the nonequivalent groups cases the priors were updated at each EM–step in BILOG-MG, but constant priors were used in EPDIRM.

The multidimensional model resulted in lower total error than the unidimensional model in most conditions. The principal exception is for the Form B distribution criterion on the French examination where the total error and bias for the multidimensional model was greater than for the unidimensional model using concurrent estimation. The difference in the total error tended to be greater in the equivalent groups conditions as opposed to the nonequivalent groups conditions. These results differ from those found in Béguin, Hanson and Glas (2000) where data were simulated using a compensatory multidimensional model. Béguin, Hanson and Glas (2000) found that the difference in total error between the unidimensional and multidimensional models was small for the equivalent groups conditions, but was uniformly very large for the nonequivalent groups conditions.

The total error and bias for the multidimensional model increased from the English to the German to the French examinations, especially for the Form B distribution criterion. This effect was also observed for some of the nonequivalent groups conditions when using the univariate model. Since the correlation between the dimensions increased from the English to the German to the French examinations this implies that the total error for the multidimensional model increased with increasing correlation between the dimensions, although this is confounded by the fact that the examinations differed as well as the correlation between dimensions. This effect of increasing total error with increasing correlation was also observed in Béguin, Hanson and Glas (2000) for the unidimensional model in the nonequivalent groups conditions, whereas here this effect was observed for the multidimensional model in both the equivalent and nonequivalent groups conditions, and only for some of the results using the unidimensional model in the nonequivalent groups conditions.

The results in this study with regard to differences in performance of the multidimensional versus the unidimensional model differs from the results in Béguin, Hanson and Glas (2000). The major difference in these two studies is the type of multidimensional model used (compensatory versus noncompensatory). It appears that the bias of the unidimensional results were less and the bias of the multidimensional results were greater for the noncompensatory model as compared to the compensatory model, at least based on the Form B distribution criterion. This effect can be seen by comparing the results in Figure 2 with the results in Figures 2 and 3 in Béguin, Hanson and Glas (2000). It is possible that some bias in the multivariate results is caused by the priors used for the item parameters in the noncompensatory model estimation. As was shown in Figure 2e, the estimated French frequency distributions for the nonequivalent groups condition were somewhat positively biased. This could be due to the choice of the prior distribution of β used in the estimation procedure. In this case, when the estimated parameter values of β tend to be larger than -1, the prior decreases the β values, which will lead to positively biased score distributions.

The differences in results between this study and the studies by Hanson and Béguin (1999) and Béguin, Hanson and Glas (2000) illustrate the sensitivity of the results of simulation studies to the true simulation model. The results of this study and Béguin, Hanson and Glas (2000) make clear that multidimensionality of the data affects the relative performance of separate and concurrent unidimensional estimation methods, although the degree to which the unidimensional model produces biased results with multidimensional data depends on the type of multidimensionality present in the data.

Appendix A – Control Files for BILOG-MG

Separate Estimation

```
>GLOBAL DFNAME=’NCME05A.1’,NPARM=3,NTEST=1, SAVE;
>SAVE PAR=’SEP05A01.PAR’;
>LENGTH NITEMS=60;
>INPUT NTOT=60,SAMPLE=2000,NALT=4,NID=4;
>ITEMS INUM=(1(1)60);
>TEST TNAME=EN;
(4A1,T6,60A1)
>CALIB NQPT=40,CYCLE=40,TPRIOR,NEWTON=15;
```

Concurrent Estimation - Equivalent Groups

```
>GLOBAL DFNAME=’NCME05C.1’,NPARM=3,NTEST=1, SAVE;
>SAVE PAR=’CON05A01.PAR’;
>LENGTH NITEMS=100;
>INPUT NTOT=100,SAMPLE=4000,NALT=4,NID=2,NFORM=2;
>ITEMS INUM=(1(1)100);
>TEST TNAME=EN;
>FORM1 LEN=60, INUMBERS=(1(1)60);
>FORM2 LEN=60, INUMBERS=(41(1)100);
(2A1,1X,I1,1X,60A1)
>CALIB NQPT=40,CYCLE=40,TPRIOR,NEWTON=5;
```

Concurrent Estimation - Nonequivalent Groups

```
>GLOBAL DFNAME=’NCME15C.1’,NPARM=3,NTEST=1, SAVE;
>SAVE PAR=’CON15N01.PAR’;
>LENGTH NITEMS=100;
>INPUT NTOT=100,SAMPLE=4000,NALT=4,NID=2,NGROUP=2,NFORM=2;
>ITEMS INUM=(1(1)100);
>TEST TNAME=EN;
>FORM1 LEN=60, INUMBERS=(1(1)60);
>FORM2 LEN=60, INUMBERS=(41(1)100);
>GROUP1 GNAME=’A’,LEN=60,INUMBERS=(1(1)60);
>GROUP2 GNAME=’B’,LEN=60,INUMBERS=(41(1)100);
(2A1,1X,I1,T4,I1,1X,60A1)
>CALIB NQPT=40,CYCLE=40,TPRIOR,NORMAL,REFERENCE=1,NEWTON=20;
```

Appendix B – Control files for EPDIRM

Separate Estimation

```
# 60 items
epdirm_start 60

# read item responses
read_examinees NCME05A.1 { @6 60i1}

# compute starting values
starting_values

# compute EM iterations
EM-steps

epdirm_end
```

Concurrent Estimation - Equivalent Groups

```
# 100 items
epdirm_start 100

# Items on form A
set items(a) [seq 1 60]

# Items on form B
set items(b) [seq 41 100]

# Responses are read from columns 6-65 for both forms
set respFmt(a) { @6 60i1}
set respFmt(b) { @6 60i1}

# Form is read from column 1 of record
set formFmt a1

# Read item responses for examinees who took form A
read-examinees_missing NCME00A.1 $formFmt items respFmt

# Read item responses for examinees who took form B
read-examinees_missing NCME00B.1 $formFmt items respFmt
```

```
# compute starting values
starting_values

# compute EM iterations
EM_steps -max_iter 300

epdirm_end
```

Concurrent Estimation - Nonequivalent Groups

```
# 100 items, 2 groups, allow unique latent variable
# points for each group
epdirm_start 100 -num_groups 2 -unique_points

# Items on form A
set items(a) [seq 1 60]

# Items on form B
set items(b) [seq 41 100]

# Responses are read from columns 6-65 for both forms
set respFmt(a) { @6 60i1}
set respFmt(b) { @6 60i1}

# Form is read from column 1 of record
set formFmt a1

# Read item responses for examinees who took form A (group 1)
read_examinees_missing NCME05A.1 $formFmt items respFmt 1

# Read item responses for examinees who took form B (group 2)
read_examinees_missing NCME05B.1 $formFmt items respFmt 2

# compute starting values
starting_values

# compute EM iterations
EM_steps -estim_dist_mean_sd -max_iter 300

epdirm_end
```

References

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of Mathematical Functions.* New York: Dover Publications.

Ackerman, T.A. (1987) *A comparison study of the unidimensional IRT estimation of compensatory and noncompensatory multidimensional item response data* (ACT research report series 87-12). Iowa-City, IA: ACT inc.

Ackerman, T. A. (1996a). Developments in multidimensional item response theory. *Applied Psychological Measurement 20,* 309-310.

Ackerman, T. A. (1996b). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement 20,* 311-329.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50,* 3-16.

Ansley, R. A., & Forsyth, T. N. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Béguin, A. A. (2000). *Robustness of Equating High-Stakes Tests*, Doctoral thesis, Enschede: University of Twente.

Béguin, A. A. & Glas, C. A. W. (in press). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika.*

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000).textit Effect of Multidimensionality on Separate and Concurrent Estimation in IRT Equating. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, IL.
(Available at http://www.b-a-h.com/papers/paper0002.html)

Birnbaum, A. (1968). Some latent trait models. In: F.M. Lord, & M.R. Novick (Eds.), *Statistical theories of mental test scores.* Reading (Mass.): Addison-Wesley

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika, 46,* 443-459.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement 12,* 261-280.

Bock, R. D., & Zimowski, M. F. (1996). Multiple group IRT. in W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer-Verlag.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in education, 12*, 383-407.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking Multidimensional Item Calibrations. *Applied Psychological Measurement, 20*, 405-416.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Embretson, S. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479-494.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186.

Fraser, C. (1988). NOHARM: *A Computer Program for Fitting Both Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory.* NSW: University. of New England.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85,* 398-409.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson, (Ed.), *Objective measurement: theory into practice, Vol. 1,* (pp.236-258), New Jersey: Ablex Publishing Corporation.

Glas, C. A. W., & Béguin, A.A. (1996). *Appropriateness of IRT observed score equating* (Research Report 96-04). Enschede: University of Twente.

Glas, C. A. W. , and Verhelst, N. D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635-659.

Haebera, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144-149.

Hanson, B. A. (2000). *Estimation Program for Dichotomous Item Response Models (EPDIRM).* (Available at http://www.b-a-h.com/software/epdirm/).

Hanson, B. A., & Béguin, A. A. (1999). *Separate versus concurrent estimation of IRT item parameters in the common item equating design.* ACT Research Report 99-8. Iowa City, IA: ACT inc.

Hirsch, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement, 26* , 337-349.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887-903.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22,* 131-143.

Kolen, M. J., & Brennan, R. L. (1995). *Test Equating.* New York: Springer.

Li, Y. H., & Lissitz, R. W. (1998). *An evaluation of multidimensional IRT equating methods by assessing the accuracy of transforming parameters onto a target test metric.* Paper presented at the annual meeting of the National Council of Measurement in education, San Diego, CA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*, 453-461.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 4,* 11-22.

Marco, G. L., (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of educational Measurement, 14,* 139-160.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika, 58*, 445-470.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*, 523-548.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric monographs, No.15.*

McDonald, R. P. (1997). Normal-ogive multidimensional model. In: W. J.van der Linden and R. K. Hambleton (eds.). *Handbook of Modern Item Response Theory.* (pp.257-269). New York: Springer.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Petersen, N. S.,.Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In: W. J.van der Linden and R. K.Hambleton (eds.). *Handbook of Modern Item Response Theory.* (pp.271-286). New York: Springer.

Spray, J. A., Davey, D. C., Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1990) *Comparison of two logistic multidimensional* item response theory models (ACT research report series ONR90-8). Iowa-City, IA: ACT inc.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7* , 201-210.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82-98). Minneapolis: University of Minnesota..

Wilson, D. T., Wood, R., & Gibbons, R. (1991) *TESTFACT: Test scoring, Item statistics, and Item Factor Analysis.* (Computer Software). Chicago, IL: Scientific Software International, Inc.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987) *Specifying the characteristics of linking items used for item response theory item calibration* (ETS Research Report 87-24). Princeton NJ: Educational Testing Service.

Zeng, L., & Kolen, M. J. (1995). An alternative approach for IRT observed- score equating of number-correct scores. *Applied Psychological Measurement, 19,* 231-240.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *Bilog MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items.* Chicago: Scientific Software International, Inc.