Effect of Multidimensionality on Separate and Concurrent Estimation in IRT Equating

A. A. Béguin
University of Twente

B. A. Hanson
ACT, Inc.

C. A. W. Glas
University of Twente

April 3, 2000

Abstract

The relative performance of separate and concurrent unidimensional IRT estimation could be affected by multidimensionality of the data. This paper reports the results of a simulation study comparing the relative performance of unidimensional estimation methods on multidimensional data. Data based on a two-dimensional IRT model are simulated according to equivalent and nonequivalent groups designs. The results of separate and concurrent unidimensional estimation are compared with the results of concurrent estimation under the two-dimensional model. In this study it becomes clear that multidimensionality of the data can effect the relative performance of separate and concurrent unidimensional estimation methods. The relative performance of separate and concurrent estimation was different for the equivalent and nonequivalent groups conditions. In the nonequivalent group conditions, the error for the unidimensional estimation methods was very large compared to the error obtained using two-dimensional IRT estimation.

1 Introduction

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. Using these estimates, the forms 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 on confidence intervals and the distribution of statistics computed using MML estimates directly applies.

A number of studies has 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 item parameter estimation. With real data the simple unidimensional model would probably be misspecified by some extent. This misspecification could affect the relative performance of unidimensional separate and concurrent estimation. A 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 follows a multidimensional IRT model (Lord & Novick, 1968; McDonald, 1967; Reckase, 1985 and Ackerman, 1996a and 1996b). In this paper the effect of multidimensionality of the data on performance of separate and concurrent estimation will be studied.

Multidimensional IRT models for dichotomously scored items 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).

Considerable attention has been given to the effect of multidimensionality on parameter estimates of unidimensional IRT models. Ansley and Forsyth (1985) examined the unidimensional estimates from two-dimensional data generated using a noncompensatory model. Ackerman (1987a) compared performances of unidimensional IRT estimates under a two-dimensional compensatory- and non-compensatory model. Ackerman (1987b) also investigated the effect of using multidimensional items in a computerized adaptive testing procedure based on a unidimensional IRT model. He found that respondents with different proficiency composites tend to receive tests with a different content. Stout (1987, 1990) introduced the concept of essential dimensionality. Analogous to the common practise in factor analysis, only

the major dimensions present in the data are used while the minor dimensions are ignored. From his research it was concluded that the existence of exactly one major dimension -so called essential unidimensionality- provides a justification of the use of IRT models that require unidimensionality. Finally, Spray, Abdel-fattah, Huang and Lau (1997) investigated the effects of a multidimensional item pool and latent proficiency space on the accuracy of the decisions made in computerized classification testing using the 3-PL model. They found that the procedure was fairly robust against violations of unidimensionality.

Also a number of multidimensional equating procedures have been proposed. Hirsch (1989) proposed a procedure that places the separate estimates of a multidimensional two-parameter logistic model for both forms of a common-examinee design on a common scale. Davey, Oshima and Lee (1996) proposed a procedure to place the estimates of a multidimensional three-parameter model for both forms of 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 place the parameters of multidimensional IRT models on the same scale. Finally, 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.

In this paper, performance of separate and concurrent estimation of a unidimensional three-parameter logistic (3-PL) model (Birnbaum, 1968; Lord, 1980) applied on multidimensional data are compared. To obtain benchmarks to evaluate these unidimensional estimates, a 3 parameter normal ogive model (3PNO) and its two dimensional counterpart ((2+2)PNO) are estimated. In these models the probability of a correct response of a person i on an item j, denoted $Y_{ij} = 1$, is written as

$$P(Y_{ij} = 1; \theta_i, \alpha_j, \beta_j, \gamma_j) \quad = \quad \gamma_j + (1 - \gamma_i)\Phi(\eta_{ij} - \beta_j) \tag{1}$$

where γ_j is the guessing parameter, β_j the difficulty parameter of the item j, Φ denotes the standard normal cumulative distribution function, and η_{ij} is a weighted proficiency. For the 3PNO model $\eta_{ij} = \alpha_j \theta_i$ with θ_i the proficiency of person i and α_j, the discrimination parameter of item i. For the (2+2)PNO model $\eta_{ij} = \sum_{q=1}^{2} \alpha_{jq}\theta_{iq}$ with α_{jq} the discrimination parameter or factor loading of item j on the q^{th} dimension, and θ_{iq} the proficiency parameter of person i on the q^{th} dimension. The 3PNO and (2+2)PNO model are estimated using a Markov chain Monte Carlo (MCMC) estimation procedures (Béguin, 2000). These procedures are generalizations to incomplete designs of procedures that use Gibbs sampling (Gelfand

& Smith, 1990) with data-augmentation to estimate models in the normal ogive context (Albert, 1992; Béguin & Glas, 1998).

2 Data

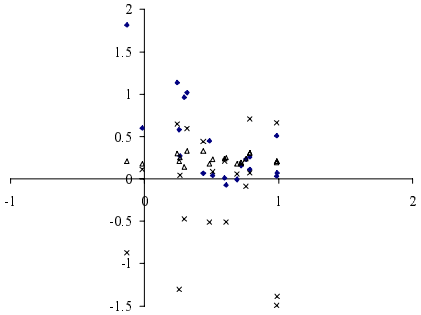
To simulate data with realistic properties, this study uses item parameter estimates of the (2+2)PNO model obtained on data from examinations at the end of secondary education in the Netherlands. These data consist of 50 dichotomously scored items from the examination '*language comprehension in English*' for the years 1992 and 1995. The original examinations have no items in common, but additional data are available where persons responded to items from both examinations. The data collection design is beyond the scope of this paper. For a description of the data-collection design applied to examinations in the Netherlands the reader is referred to Glas and Béguin (1996) or Béguin (2000). As mentioned above, the item parameters are obtained using a two dimensional (2+2)PNO MCMC estimation procedure. In this estimation procedure the item parameters are estimated assuming different proficiency distributions for the groups in the design. So this procedure is a multiple-group concurrent estimation procedure.

To simulate data according to a common-item nonequivalent group design, 10 items were randomly selected from each of the two examinations. These 20 items were used as items common to two test forms, A and B. Form A was created using the 20 selected items and the 40 remaining items from one of the examinations. Form B contained the 20 selected items and the 40 remaining items from the other examination. So, Form A contained all 50 items from the 1992 examination and 10 items from the 1995 examination, and Form B contained 10 items from the 1992 examination and all 50 items from the 1995 examination. To give an impression of the item parameters used in estimating the data, in Figure 5.1, the values of the factor loading on the second dimension, α_2, the difficulty parameter β, and the guessing parameter γ are plotted against the value of the factor loading on the first dimension α_1.

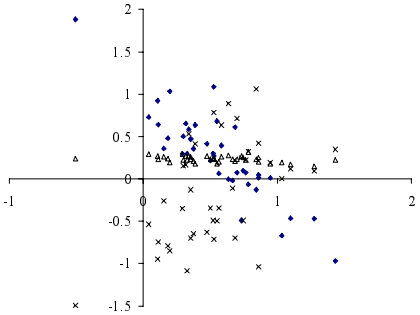
3 Method

Samples of item responses for forms A and B are obtained for 6 different conditions. These conditions differ in the covariance of the proficiency distribution and in difference in mean proficiency between the two populations. Three levels of covariance between dimensions and two levels of mean proficiency difference between the two populations, are assumed. The three levels of covariance have a unit variance on

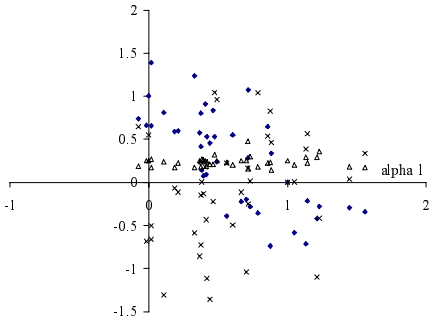
(a) common items



(b) unique items Form B



(c) unique items Form A



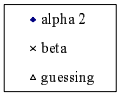


Figure 1. Parameter values

Table 1. Overview of the conditions

Condition	Mean Proficiency		Covariance
	Form A	Form B	Both Forms
1	(0,0)	(0,0)	$\begin{pmatrix} 1 & \\ 0.5 & 1.25 \end{pmatrix}$
2	(0,0)	(0,0)	$\begin{pmatrix} 1 & \\ 0.7 & 1.49 \end{pmatrix}$
3	(0,0)	(0,0)	$\begin{pmatrix} 1 & \\ 0.9 & 1.81 \end{pmatrix}$
4	(0,0)	(1,0)	$\begin{pmatrix} 1 & \\ 0.5 & 1.25 \end{pmatrix}$
5	(0,0)	(1,0)	$\begin{pmatrix} 1 & \\ 0.7 & 1.49 \end{pmatrix}$
6	(0,0)	(1,0)	$\begin{pmatrix} 1 & \\ 0.9 & 1.81 \end{pmatrix}$

the first dimension and differ in the covariance and the variance on the second dimension. The three conditions have a covariance of .5, .7 and .9 with a variance on the second dimension of 1.25, 1.49 and 1.81, respectively. The mean proficiency on the first dimension for the Form A respondents is 0 in all conditions while the mean proficiency for the Form B respondents is either 0 or 1. The mean proficiency for the second dimension is 0 in all conditions. For a summary of the characteristics of the conditions see Table 1. The first three conditions can be considered equivalent groups conditions, since the proficiency distributions of the populations administered Form A and B are equal. The equivalent groups design differs in the covariance and variance on the second dimension. The last three conditions are nonequivalent group conditions with the same covariances as in the first three conditions. The conditions will be referred to using the covariance between dimensions and an indication of whether the groups are equivalent or nonequivalent. For example, condition 5 in Table 1 will be referred to as: nonequivalent .7 covariance condition.

3.1 Estimation of the parameters

For each condition 20 samples of both forms are generated with 2000 persons per form. In each sample and each condition three sets of parameter estimates using BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1996) are obtained, two sets for each separate form (separate estimation) and one for both forms simultaneously (concurrent estimation). Also for each sample 2 MCMC runs are performed, a 3PNO run and a (2+2)PNO run.

In BILOG-MG, it was specified that normal population distributions were assumed. Further, the default priors were used on the a,and c parameters. To support convergence an additional $N(0, 2)$ prior distribution was used on the b parameter. Appendix A gives the BILOG-MG control files used to obtain parameter estimates for each simulated sample. The MCMC procedure consisted of 3000 iterations with a burn-in period of 1000 iterations. As starting values for the 3-PNO model $\alpha = 1$, $\beta = 0$ and $\gamma = \gamma_{true}$, for all items, and proficiency, $\theta = 0$, for each respondent were used . The priors on the item parameters were $\pi(\alpha) \sim N(1., 0.25)$, $\pi(\beta) \sim N(0, 0.5)$ and $\pi(\gamma) \sim \text{Beta}(6, 16)$, which is equivalent to 20 prior observations with probability 0.25. The prior on γ is the same as the default prior used in BILOG-MG. As starting values for the (2+2)PNO model the true parameters were used for the item parameters, and $\boldsymbol{\theta} = \mathbf{0}$ was used for the proficiency of each respondent. The priors on the item parameters were $\pi(\alpha) \sim N(0., 0.25)$, $\pi(\beta) \sim N(0, 0.5)$ and $\pi(\gamma) \sim \text{Beta}(50 * \gamma_{true}, 50 * (1 - \gamma_{true}))$. In the (2+2)PNO model, more parameters had to be estimated, and to support the stability of the estimates a more informative prior was used on γ. Furthermore, the mean of the prior on α was set equal to 0, because in the (2+2)PNO model the α parameters are not restricted to be positive.

In the separate estimation conditions, the parameters of Form A and Form B must be brought on a common scale. This scaling is performed 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 populations administered Form A and B have equal proficiency distributions (see Table 1), can be considered equivalent groups conditions. In an equivalent groups design it is not necessary to assume different populations for the groups taking Form A and Form B. Consequently, in separate estimation of the forms, no linking is necessary to bring the forms on a common scale. By the same token, one can assume a single population for both samples if concurrent estimation is applied. In this study, a single population is assumed in the estimation of the unidimensional models in the conditions where the populations administered Form A and B have equal proficiency distributions. Consequently, in separate estimation no scaling is performed and in BILOG-MG concurrent estimation a single group is specified. In the (2+2)PNO model different population distributions are esti-

mated 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 have equal proficiency distributions.

In the separate estimation conditions there will be two sets of item parameter estimates for the common items. 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 the item parameter scaling. An alternative would be using the average of the item parameter estimates for the two forms (Kim & Cohen, 1998).

3.2 Evaluation of scaling

To evaluate the quality of the item parameter scaling, differences in results of equating scores on Form B to scores on Form A are assessed. Two criteria are used that are based on the IRT observed-score (OS) equating of number-correct (NC) scores (Zeng & Kolen, 1995) on Form B to scores on Form A. In this technique, equipercentile equating (Kolen & Brennan, 1995) is performed on the score distributions of both forms computed for one population. Here, the score distributions of Form A and B were estimated for the population administered Form A.

Estimating score distributions.

Using the estimated item and population parameters, the compound binomial distribution was used to generate the score distribution of respondents of a given proficiency, $\boldsymbol{\theta}$. The score distribution for population a can be computed by integrating over the population distribution of $\boldsymbol{\theta}$, that is,

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

where $\{x\,|r\}$ stands for the set of all possible response patterns resulting in a score r. 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) was used to obtain $f_r(x\,|\boldsymbol{\theta})$, the score distribution of respondents of a given proficiency, $\boldsymbol{\theta}$. To obtain accurate results it is necessary to apply a large number of quadrature points. Therefore, 180 quadrature points are used in the unidimensional case and 100 quadrature points are 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,

every 20 iterations the procedure by Lord and Wingersky was applied with current values of the person and item parameters of the Gibbs-sampler. The estimated score distribution was the mean over 100 thus obtained score distributions. A nice property of this procedure was that the uncertainty of the parameter estimates was taken into account in the estimation of the score distribution.

Criteria

Evaluation of performance of equating in the 6 conditions was based on two different criteria. First, the differences between the Form B score distributions estimated under various models and the population Form B score distribution computed with the model used to generate the data were compared. Second, equivalent score points from the observed score equating function under different models were compared with the equivalent score points obtained with the model used to generate the data. 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 in 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 frequency of score point r based on the parameters of (2+2)PNO that were used in generating the data. Let f_{jr} be the frequency of score point r estimated from sample j. To compare these 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,

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

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):

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

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

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

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

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

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

In the second criterium, equivalent score points from IRT observed score equating obtained using various models were compared with the equivalent score points obtained with the model used to generate the data. Let $s_{true,r}$ be the score point on the new examination that is equivalent with the score point r on the reference examination, based on the true parameter values. Let s_{jr} be the score point on the new examination that is equivalent with a score point r, and let equivalence be based on the parameters estimates of sample j. Furthermore, let $p_{r,true}$ be the probability of obtaining a score r 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 the $k+1$ score points of the reference examination. The score points were multiplied by $p_{j,true}$, which resulted in

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

The weighted mean squared error 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,

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

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

$$\overline{s_j} = \frac{1}{20}\sum_{i=1}^{20} s_{ij}. \tag{9}$$

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

$$WMAE = \frac{1}{20}\sum_{r=0}^{k} p_{r,true}\sum_{j=1}^{20}\left|s_{jr} - s_{true,r}\right|. \tag{10}$$

4 Results

Two factors are investigated in this study: 1) the performance of equating using concurrent versus separate estimation. 2) the performance of equating using MCMC (2+2)PNO versus BILOG-MG or 3PNO concurrent estimation. Except for two of the data sets in the separate estimation condition all BILOG-MG and MCMC runs

converged. The two runs that did not converge were separate estimation runs on group A samples from nonequivalent groups conditions. One was a 0.5 covariance samples and the other was a 0.9 covariance sample. Although the convergence criterium was not met after 40 EM and 20 Newton iterations, the item parameter estimates of these samples where within reasonable bounds.

In general, the results obtained using BILOG-MG concurrent estimation were similar to the results obtained using MCMC estimation of the 3PNO model. Furthermore, the results obtained using BILOG-MG for both concurrent and separate estimation were also similar to results obtained using a preliminary version of an open-source IRT estimation toolkit (Hanson, 2000).

First, the true and estimated frequency distribution of Form B were compared. To illustrate the results, the frequency distributions for the nonequivalent groups with covariance 0.9 condition are plotted in Figure 2. The frequency distribution obtained using the true parameters is plotted together with the estimated frequency distributions of the 20 different samples.

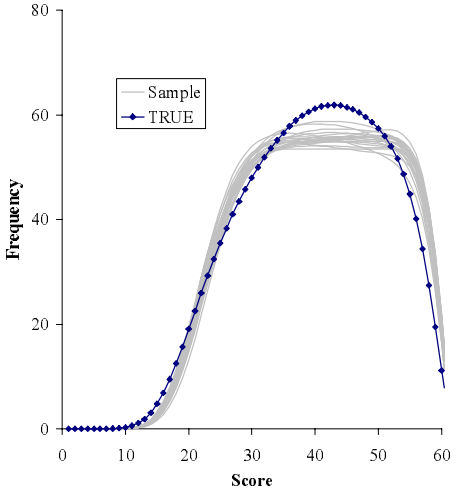
From Figure 2 it becomes clear that the estimated frequency distributions based on the unidimensional BILOG-MG estimates deviated from the frequency distribution based on the true parameters. For both the separate and concurrent estimates, the estimated frequencies are too small in the top of the distribution and to large in the upper tail. As might be expected, the frequency distributions based on the estimates from the (2+2)PNO model showed only minor deviations from the frequency distribution based on the true parameters. Figure 3 presents the estimated frequency distributions for the .5 and .7 covariance nonequivalent groups condition and based on the BILOG-MG concurrent estimation. It becomes clear that the deviation observed in the .9 covariance condition (Figure 2b) is smaller for the .7 and .5 conditions.

In Table 2, the mean squared error, bias and variance are given for the different conditions and estimation methods, together with the value of the X^2-statistic.

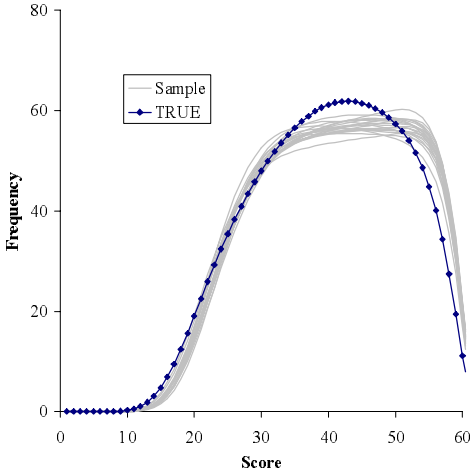
The performance of the unidimensional model relative to the multidimensional model differs in the equivalent versus nonequivalent groups conditions. In the equivalent groups conditions the MSE based on the unidimensional model is of the same order of magnitude as the MSE obtained using the (2+2)PNO model. In the nonequivalent groups conditions the MSE based on the unidimensional model was far larger than the MSE based on the (2+2)PNO model. The bias in the unidimensional estimation method increased with the increase in covariance and variance in the second proficiency dimension.

In the equivalent groups conditions the MSE, bias and variance were smaller

(a) Separate estimation
BILOG-MG



(b) Concurrent estimation
BILOG-MG



(c) (2+2)PNO MCMC estimation

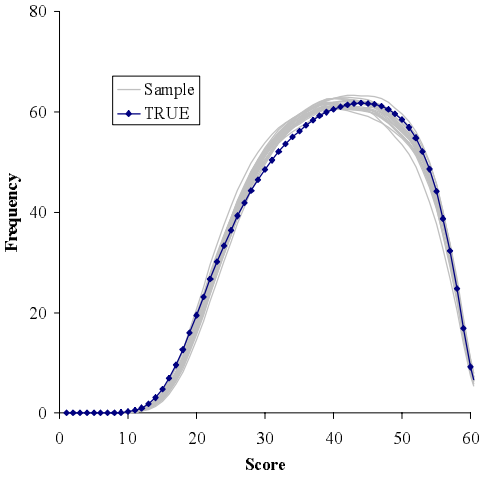


Figure 2. Score distributions for Form B in the nonequivalent groups covariance .9 condition, determined using the true proficiency distribution of the population administered Form A.

(a) Covariance .7 with concurrent estimation BILOG-MG

(b) Covariance .5 with concurrent estimation BILOG-MG

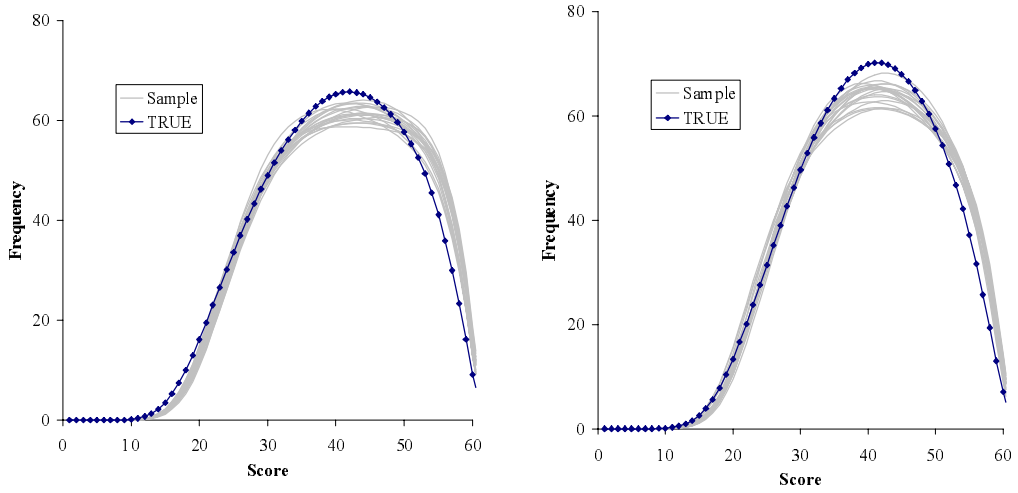


Figure 3. Score distributions for Form B determined using the true proficiency distribution of the population administered Form A.

for the concurrent estimation condition than for the separate estimation condition. From comparison of the separate estimation method with and without scaling, it is observed that the 'no scaling' condition resulted in a lower MSE, variance, and bias.

In the nonequivalent groups conditions, the MSE and bias were larger for the concurrent estimation method in the covariance .7 and .9 condition. In the covariance .5 condition the MSE and bias were larger for the separate estimation method. As mentioned before, for both concurrent and separate estimation methods the MSE and bias are very large compared to the MSE and bias based on the (2+2)PNO estimates. Similar to the results found in the equivalent groups conditions, the MSE and bias increased with the increase in covariance and variance in the second proficiency dimension.

Comparing the X^2 for the different conditions and estimation methods, the following results were found. The values of X^2 for the (2+2)PNO estimation method were slightly higher in the nonequivalent groups condition than in the equivalent groups condition. For the unidimensional estimation methods the values of X^2 were far higher in the nonequivalent groups condition than in the equivalent groups condition. Further, the values of X^2 were smaller for the concurrent estimation method than for the separate estimation method except for the .7 and .9 covariance nonequivalent groups conditions. Note that it is unclear how these results should be interpreted, since the distribution of the X^2 statistic is unknown and can differ

Table 2. Mean squared error of estimated frequency

Design	Cov.	Estimation	MSE	Bias	Variance	X^2
Equivalent groups	0.5	con	2.6	1.6	0.9	99.
		sep	3.7	2.0	1.7	143.
		sepNS	2.8	1.9	0.9	105.
		2+2PNO	3.7	2.3	1.4	115.
	0.7	con	2.5	1.7	0.8	106.
		sep	4.3	2.4	1.9	158.
		sepNS	2.7	1.9	0.8	111.
		2+2PNO	2.1	1.2	0.9	85.
	0.9	con	3.8	2.9	0.8	197.
		sep	4.4	3.2	1.3	201.
		sepNS	4.0	3.1	0.9	200.
		2+2PNO	2.5	1.7	0.8	102.
Nonequivalent groups	0.5	con	12.4	10.5	1.9	575.
		sep	15.0	13.2	1.8	696.
		2+2PNO	3.1	2.2	0.9	146.
	0.7	con	16.8	15.0	1.7	844.
		sep	15.7	13.6	2.1	765.
		2+2PNO	6.1	4.3	1.9	229.
	0.9	con	21.8	20.4	1.5	1093.
		sep	20.7	19.2	1.5	967.
		2+2PNO	3.4	2.3	1.1	158.

con: BILOG-MG concurrent estimation

sep: BILOG-MG separate estimation with scaling

sepNS: BILOG-MG separate estimation without scaling

2+2PNO: MCMC concurrent estimation of the (2+2)PNO model

over conditions. The X^2 is calculated by summation over 1,200 cells. However the degrees of freedom of each X^2 value is unknown.

Table 3. Weighted error of equated scores

Design	Cov.	Estimation	WMSE	Bias	Variance	WMAE
Equivalent groups	0.5	con	0.35	0.16	0.19	0.35
		sep	0.35	0.15	0.19	0.35
		sepNS	0.36	0.18	0.19	0.36
		2+2PNO	0.36	0.15	0.20	0.36
	0.7	con	0.27	0.10	0.17	0.27
		sep	0.33	0.13	0.20	0.33
		sepNS	0.27	0.08	0.18	0.27
		2+2PNO	0.27	0.11	0.16	0.27
	09	con	0.26	0.12	0.15	0.26
		sep	0.28	0.13	0.15	0.28
		sepNS	0.33	0.15	0.18	0.33
		2+2PNO	0.26	0.09	0.17	0.26
Nonequivalent groups	0.5	con	0.53	0.43	0.10	0.47
		sep	0.68	0.56	0.12	0.54
		2+2PNO	0.36	0.17	0.19	0.36
	0.7	con	1.46	1.28	0.18	0.94
		sep	0.95	0.86	0.09	0.66
		2+2PNO	0.39	0.27	0.12	0.39
	0.9	con	1.79	1.68	0.11	1.12
		sep	1.13	1.04	0.09	0.75
		2+2PNO	0.33	0.17	0.16	0.33

con: BILOG-MG concurrent estimation

sep: BILOG-MG separate estimation with scaling

sepNS: BILOG-MG separate estimation without scaling

2+2PNO: MCMC concurrent estimation of the (2+2)PNO model

In Table 3, the WMSE, weighted bias, weighted variance and WMAE are given for the equated score points determined for different conditions and estimation methods. In the three equivalent group conditions, there were not large differences among the estimation methods in terms of WMSE, bias, variance, and WMAE. The separate estimation methods resulted in somewhat higher values of the WMSE and WMAE. The no scaling condition resulted in a lower WMSE in the covariance .7 condition and in a higher WMSE in the covariance .9 condition. In the three nonequivalent

groups conditions, the differences between the unidimensional estimation methods and the (2+2)PNO estimation method were relatively large. This difference was larger in the higher covariance conditions. In the nonequivalent groups conditions, for the unidimensional estimation methods, the WMSE, bias and WMAE increased with increasing covariance. This effect was larger for the concurrent estimation condition. Comparing the separate and concurrent BILOG-MG estimation methods in the nonequivalent group conditions, the separate estimation method resulted in lower WMSE, bias, variance and WMAE for the covariance .7 and .9 condition, while the concurrent estimation method resulted in lower WMSE, bias, variance and WMAE in the .5 covariance condition.

5 Conclusions

In this study, the effect of the estimation method on equating results were compared when unidimensional models were applied on multidimensional data. As with any simulation study considerable caution needs to be exercised in drawing conclusions due to the small number of conditions investigated. In this case, the results pertain to only the two specific forms and six different conditions used in this study. The only aspect varied in the conditions was the difference between the proficiency distributions of the populations administered the forms. There was no variation in data collection designs or the number of respondents in the design.

In the equivalent groups conditions, the different unidimensional estimation methods resulted in criteria that were quite similar to the criteria obtained using the multidimensional (2+2)PNO model. In these conditions the separate estimation method where scaling is applied using the Stocking-Lord method resulted in generally higher criterium values than the unidimensional concurrent estimation method. Further, estimation without scaling resulted in similar or better performance than the separate estimation method with scaling. These results are opposite to those found in the simulation study by Hanson and Béguin (1999). They reported that concurrent estimation generally resulted in better performance than separate estimation, and better performance if scaling was applied in a equivalent groups design.

In the nonequivalent group conditions, the error for the unidimensional methods was very large compared to the error obtained using the (2+2)PNO model. From this result it must be concluded that performance of both separate and concurrent estimation methods are unsatisfactory in these conditions. The error increased with an increase in the covariance and variance of the second proficiency dimension. This effect was stronger for the concurrent estimation method than for the separate estimation method. Consequently the separate estimation method performed better in the .7 and .9 covariance nonequivalent groups conditions. From the current analyses

no well defined explanation for this effect could be found. In Figure 2 it is shown that the estimated frequency for the nonequivalent group and .9 covariance condition deviates from the true distribution both in the top and the upper tail of the distribution. In Figure 3, the score distributions based on BILOG-MG concurrent estimates are plotted for the nonequivalent .5 and .7 covariance condition. From Figure 3 it becomes clear that the misfit increases with increasing covariance. It is not clear if this misfit is the source of the increase of the error. Also no explanation is apparent for the occurrence of this misfit. Further research and additional simulation studies are needed to clarify this result.

In the current study it remains unclear what effect of multidimensionality one can expect in practise. Although the item parameter values used in this study were estimates from an empirical dataset, it is unknown to what extent the different multidimensional conditions are realistic. The item parameter values are obtained together with proficiency distribution estimates from a concurrent multidimensional MML estimation assuming multivariate normal proficiency distributions. Together, these item and population parameter estimates are realistic parameter values, but it is unclear if the item parameter values are also realistic if different population distributions are assumed. To study the relative effects of different degrees of multidimensionality under realistic conditions the parameter values of the various conditions must be obtained from datasets that possess the degree of multidimensionality one is interested in.

In general, the unidimensional IRT models resulted in reasonable estimated score distributions when applied on multidimensional data from an equivalent groups design. In this study the nonequivalent groups conditions, led to large deviations between the true and estimated score distributions. This effect increased with larger covariance and second dimension variance and this increase is larger for the concurrent estimation method. The differences in results between this study and the study by Hanson and Béguin (1999) illustrates the sensitivity of the results of simulation studies to the model used to simulate the data. The results of this study make it clear that multidimensionality of the data affects the relative performance of separate and concurrent unidimensional estimation methods.

Appendix A

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;
```

References

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

Ackerman, T.A. (1987a) *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. (1987b) *The use of unidimensional item parameter estimates of multidimensional items in adaptive testing* (ACT research report series 87-13). 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. & Glas, C. A. W. (1998). *MCMC estimation of multidimensional IRT models* (Research report 98-14). Enschede: University of Twente.

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

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.

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 Toolkit for Item Response Models (ETIRM).* (Available at http://www.b-a-h.com/software/cpp/etirm.html).

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.

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.

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., Abdel-fattah A.A., Huang, C., & Lau, C.A. (1997) *Unidimensional approximations for a computerized test when the item pool and latent space are multidimensional* (ACT research report series 97-5). 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.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55,* 293-326.

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.