that are available in the framework of the methodology of Muthén (1984), extending IRT modeling. Some more complex types of modeling would also seem to be useful. We may consider multiple latent ability variables with separate sets of items measurements. The interrelations of these abilities can be explored in relation to other relevant variables. Furthermore, with categorical x variables, a powerful simultaneous analysis of multiple groups of examinees can be carried out. This makes it possible to investigate invariance hypotheses regarding both measurement and structural parameters, which would be valuable, for example, in studies of test item bias.

References

Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries* (International Association for the Evaluation of Educational Achievement: International Studies in Evaluation I). Stockholm: Almqvist & Wiksell.

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

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551–560.

Muthén, B. (1979). A structural probit model with latent variables. *Journal of the American Statistical Association*, *74*, 807–811.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, *22*, 43–65.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*, 407–419.

Author

BENGT MUTHÉN, Associate Professor, Graduate School of Education, UCLA, Los Angeles, CA 90024. *Specializations*: Latent variable structural equation modeling with categorical and other nonnormal data.

Journal of Educational Statistics
Summer 1985, Volume 10, Number 2, pp. 133–142

MULTIPLE GROUP IRT MODELING: APPLICATIONS TO ITEM BIAS ANALYSIS

BENGT MUTHÉN
and
JAMES LEHMAN
University of California, Los Angeles

KEY WORDS. *Measurement invariance, factor analysis, random disturbances.*

ABSTRACT. This article shows the applicability of new methodology for multiple-group factor analysis of dichotomous variables. Situations are considered where the same set of test items has been administered to more than one group of examinees. The new methodology is contrasted with the IRT approach to item bias analysis. An example is given in which females and males have taken a certain biology test.

The purpose of this article is to show the applicability of IRT-related methodology developed by Muthén and Christoffersson (1981). We will be concerned with the situation in which the same set of dichotomously scored test items has been administered to more than one group of examinees. In such situations, issues of item invariance are of primary concern. IRT modeling has been proposed to study these issues in an effective way, particularly under the rubric of item bias (see, e.g., Linn, Levine, Hastings, & Wardrop, 1981; Lord, 1977, 1980). Here, biased items (not showing invariance across groups) are singled out and modified or discarded, after which the values of the latent ability variable are estimated for each individual. The alternative approach to be discussed here differs from the above in three important respects: (a) for each item, the model is in a certain sense more general than conventional IRT models; (b) a simultaneous analysis is performed of the various groups under certain testable invariance restrictions of parameters; and (c) individual values of the latent ability variable are not estimated (as is also the case in marginal maximum likelihood estimation), but rather group means and variances.

The new approach gives a very powerful multiple-group analysis. An important product of the new approach is that items identified as biased by conventional IRT analysis need not be discarded but can still be used to estimate group parameters. In the next sections the new methodology is outlined and illustrated by an achievement test example.

The research of the first author was supported by Grant No. SES-8312583 from the National Science Foundation.

#15

Multiple-Group Modeling

Let us consider the multiple groups to be analyzed as samples from different populations with certain forms of cross-population parameter invariance. Let the superscript g denote population membership. In line with Muthén and Christoffersson (1981), this article limits attention to modeling with normal ogives and zero guessing parameter values. Consider the two-parameter normal ogive model for item i in group g,

$$Pr(y_i^{(g)} = 1|\eta^{(g)}) = \Phi[a_i(\eta^{(g)} - b_i)], \tag{1}$$

where $i = 1, 2, \ldots, p$ and $g = 1, 2, \ldots, G$. Let

$$E(\eta^{(g)}) = \alpha^{(g)}, \; V(\eta^{(g)}) = \psi^{(g)}, \tag{2}$$

denoting the structural parameters, assumed to vary across groups. Note that in Equation 1 the measurement parameters a_i and b_i have no group superscript, reflecting a hypothesis of group-invariant measurement parameters. Hence, for a second group g', we would expect the conditional expectation function (the item characteristic curve, ICC)

$$\Phi[(a_i(\eta^{(g')} - b_i)] \tag{3}$$

to hold. Given an individual with ability η_0, say, Equations 1 and 3 therefore give the same probability. Since this is true for all values η_0, item i will be said to be "unbiased" (c.f., Linn, Levine, Hastings, & Wardrop, 1981, p. 161). The aim of this paper is to show that the invariance hypothesis reflected in Equations 1 and 3 is unnecessarily restrictive.

We may consider a less restrictive invariance hypothesis as follows. Consider again two groups, g and g'. Whereas for group g we assume that Equation 1 holds, the conditional expectation function for group g' is assumed to be

$$\Phi[d_i^{(g')}a_i(\eta^{(g')} - b_i)]. \tag{4}$$

The admissible range for the new measurement parameter $d_i^{(g')}$ is restricted to positive values. We note that with a single group to be analyzed, the d_i and a_i parameters are confounded. However, this not the case with several groups. Whereas group g can be said to be described by a "standardized ICC" with $d_i^{(g')} = 1$, group g' is allowed to have a "flatter" ($d < 1$) or "steeper" ($d > 1$) ICC, passing through the same point of inflection at $\eta = b_i$. Whereas in group g the a_i's and b_i's are identified and estimable as usual, using the additional group g' information clearly makes the d_i's identified and estimable.

We may interpret the new d-parameters as reflecting a weakening ($d < 1$) or strengthening ($d > 1$) of the relationship between y_i and η due to factors other than η that are independent of η. Such random disturbance factors may include other, incidental abilities called upon to solve item i, or distractions

in the test-taking situation. Although such random disturbance factors are thought to be present in both group g and g', the new modeling explicitly recognizes that the strength of influence of such factors may differ across the groups. Furthermore, it would seem that such differences in the interaction between item and test-taking group are not of primary interest when trying to determine whether a set of items contains biased items. The more important aspect of item bias is whether the direct influence of η itself on y_i's is invariant or not across groups. It follows that we are less concerned if d_i's are noninvariant across groups than if a_i's and b_i's are noninvariant across groups. In our approach, structural parameters describing group differences are in fact identified and estimable even when d_i's vary across groups, as long as a_i's and b_i's do not.

It may be noted that with an item for which the a and b parameters are invariant across groups but d is not, item bias appears in the form of ICC's that cross each other. We should make clear that we do not believe that this is the only or most common form of bias. Bias in the a or b parameter can of course still occur. However, allowing d parameters to differ across groups allows for less restrictive modeling.

The concept of d-parameters will be viewed here to arise naturally in the following way, drawing from Muthén and Christoffersson (1981). Assume that underlying each observed response variable y_i for a certain group (population) there is a continuous latent response variable y_i^*, such that

$$y_i = \begin{cases} 1, \text{ if } y_i^* > \tau_i \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

$$y_i^* = \lambda_i \eta + \varepsilon_i, \tag{6}$$

where the τ_i's are threshold parameters for the items, the λ_i's are loading parameters, and the ε_i's are residuals with zero means that are uncorrelated among themselves and with the latent variable (the factor) η. The residuals are taken to be multivariate normal, and the factor is taken to be normal, yielding multivariate normal y^*'s. While the normality of the residuals corresponds to the IRT normal ogive specification, this model has the added specification of a normal factor. Let $V(\varepsilon_i) = \theta_{ii}$ and let the mean and variance of the factor be denoted as above.

This model gives the conditional probability curve of Equation 1. Muthén and Christoffersson, among others, have shown that using the group superscript g,

$$a_i^{(g)} = \lambda_i^{(g)} \theta_{ii}^{(g)-1/2}, \tag{7}$$

$$b_i^{(g)} = \tau_i^{(g)}/\lambda_i^{(g)}, \tag{8}$$

when relating the new model to the conventional two-parameter normal ogive

(with d_i's = 1). It follows that the extended model of Equation 4 has $d_i = \theta_{ii}^{-1/2}$ and $a_i = \lambda_i$, $b_i = \tau_i \lambda_i^{-1}$. Hence, the d-parameters are given the interpretation of inverted standard deviations of residuals in linear regressions of latent response variables on the factor; the larger the standard deviation, the stronger the deflation. The thresholds and the loadings are measurement parameters describing the properties of the items. From Equations 7 and 8 we see that with group-invariance of thresholds and loadings, the conventional IRT parameter b is still invariant, whereas the conventional IRT parameter a is not, unless residual variances are also invariant. Hence, the new model is more flexible.

To shed further light on the modeling of Equations 5 and 6 in multiple-group situations, it is interesting to draw parallels with factors analysis of continuous observed variables. Here, the y^*'s in Equation 6 are directly observed. Meredith (1964) applied classic selection formulas due to Pearson and Lawley in the context of factor analysis (see also Muthén & Jöreskog, 1983). Consider a total population for the observed y^*'s, divided into subpopulations by a set of selection variables. If the regressions of the y^*'s on the selection variables are linear and homoscedastic in the total population, it was shown that in each subpopulation the factor analysis model still holds if the selection variables are only indirectly related to the observed variables via the factors. The intercepts (set to zero in Equation 6), factor loadings, and residual variances of Equation 6 then remain the same in each subpopulation, whereas the factor mean and variance will differ across subpopulations. This would be an argument for invariant residual variances in addition to the invariance of thresholds and loadings. Then invariance of d's would hold, and consequently invariance of the two-parameter normal ogive conditional probability curves. However, it is our experience that in practice, working with observed continuous variables, the additional invariance specification for residual variances frequently does not hold. The assumptions underlying the Pearson-Lawley-Meredith selection results are presumably too restrictive for many practical settings. As is the case for our IRT-related approach, the invariance of residual variances is an unnecessary requirement in multiple-group factor analysis with observed continuous variables.

Consonant with Lord (1977, 1980), we suggest that plots of estimated item measurement parameters can be instructive. Under certain forms of invariance, item parameters will be linearly related across groups. Here, we will deduce corresponding results from the Muthén and Christoffersson (1981) model. For simplicity, consider two groups of individuals, where estimates of thresholds, loadings, and residual variances have been obtained from each group separately. In single-group analyses, the Muthén and Christoffersson methodology may use the parameter standardization of, say, factor mean of zero and variance one, and latent response variable variance of one, so that

$$\theta_{ii} = 1 - \lambda_i^2. \tag{9}$$

Using asterisks for parameters under this standardization, Muthén and Christoffersson deduce:

$$\tau_i^* = (\tau_i - \lambda_i \alpha)\sigma_{ii}^{-1/2}, \tag{10}$$

$$\lambda_i^* = \lambda \psi^{1/2} \sigma_{ii}^{-1/2}, \tag{11}$$

$$\theta_{ii}^* = \theta_{ii} \sigma_{ii}^{-1}. \tag{12}$$

Here σ_{ii} refers to the true (unstandardized) variance of y_i^* in Equation 6. It is useful to consider the functions (c.f., Equations 7 and 8),

$$a_i^* = \lambda_i^* \theta_{ii}^{*-1/2}, \tag{13}$$

$$b_i^* = \tau_i^* \lambda_i^{*-1}. \tag{14}$$

Considering two groups f and h, it follows by straightforward algebra that under group invariance of τ_i and λ_i,

$$a_i^{*(f)} = \psi^{(f)1/2} \psi^{(h)-1/2} \theta_{ii}^{(h)1/2} \theta_{ii}^{(f)-1/2} a_i^{*(h)}, \tag{15}$$

$$b_i^{*(f)} = \psi^{(f)-1/2}(\alpha^{(h)} - \alpha^{(f)}) + \psi^{(f)-1/2} \psi^{(h)1/2} b_i^{*(h)}. \tag{16}$$

From Equation 16 we find that the b_i^*'s are linearly related with intercept and slope, depending on structural parameters. Hence, a plot of estimated b^*'s should indicate linearity and the relative magnitude across groups of factor means and variances. From Equation 15, however, we find that the a^*'s from different groups are linearly related only if the residual variances are invariant for each item. If this additional invariance holds, the slope in the a^* plot will be the inverse of that in the b^* plot.

Muthén and Christoffersson (1981) present a limited information, generalized least-squares procedure to estimate the parameters in multiple-group situations. Large-sample standard errors of estimates are obtained as well as a large-sample chi-square measure of fit to the restrictions imposed by the model.

An Application

As an application we will now carry out an item bias type analysis of a set of biology achievement items from the IEA study (Comber & Keeves, 1973). The items were administered in grade 12 to males and females in England. Sample sizes were 1,075 for females and 743 for males. A total of 16 5-choice biology items were used. These were dichotomously scored for our analyses.

A disadvantage of the Muthén and Christoffersson (1981) methodology is that no provision is made for the possibility of guessing. If there is guessing, the analysis may be distorted. We attempted to circumvent this potential problem by first carrying out a traditional three-parameter logistic IRT anal-

ysis by LOGIST after which items with a considerable amount of guessing were deleted from further analyses. A guessing estimate larger than .2 was deemed large enough to warrant deletion. For females, items 3, 5, 11, 14, and 16 were deleted, while for males, items 5, 6, 11, and 16 were deleted. This yields 10 common items to be studied across sex.

In the analyses that follow we use the LISCOMP computer program (Muthén, 1984). We first carry out an analysis for each sex separately. Unidimensionality may then be tested by the large-sample chi-square statistic. With 35 degrees of freedom, females obtained a value of 32.77 and males 39.84. Hence, unidimensionality cannot be rejected for either group. The estimates for each group are given in Table I. The estimates are given in the parameterization indicated by asterisks in Equations 10–14.

Let us now create and plot estimates of the a^*'s and b^*'s of Equations 13 and 14 to investigate whether or not approximately linear relationships hold across sex. This indicates to what extent measurement parameter invariance holds. In Figures 1 and 2, plots are presented for males against females for estimated a^*'s and b^*'s, respectively. Although only ten data points are available, it seems that the a^* plot does not indicate linearity, whereas the b^* plot does. The correlations are .47 and .94, respectively. In line with our previous discussion, this could indicate that invariance of thresholds and loadings holds but that invariance of residual variances does not. Hence, if the results of these few data points were to be taken seriously, item parameter invariance for the classic two-parameter normal ogive does not hold; invariance of a's seems to be violated. Item bias analysis in the traditional IRT framework would discard items for which the non-invariance of a's was significant (see, e.g., Lord, 1980, p. 223). Here, a substantial amount of items may be lost.

TABLE I

Measurement Parameter Estimates Analyzing Females and Males Separately

	Females			Males		
Item	Threshold	Loading	Res. var.	Threshold	Loading	Res. var.
1	.572	.287	.918	.192	.278	.922
2	.052	.374	.860	−.039	.494	.756
4	.177	.316	.900	−.076	.097	.991
7	−.946	.420	.824	−.595	.390	.848
8	−1.117	.643	.587	−1.299	.509	.741
9	−1.485	.357	.873	−1.456	.537	.712
10	−1.049	.509	.741	−1.130	.614	.623
12	−.282	.467	.782	−.375	.534	.715
13	.295	.302	.909	.113	.482	.768
15	−1.045	.293	.914	−1.348	.468	.781

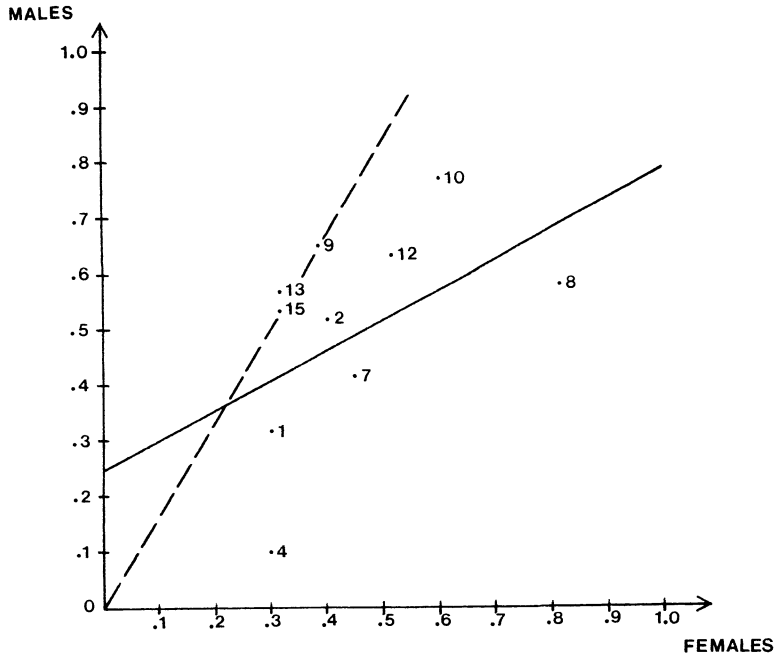

FIGURE 1. Plot of a^*'s for ten items. Correlation = .47, Estimated regression line (solid line): Intercept = .27, Slope = .54

In Figure 2, the estimated regression line for the 10 items is given. By Equation 6, the intercept and slope estimates suggest that the female factor variance is about 36% of the male variance and that the male factor mean is larger than the female mean. Inverting this slope estimate should give us the slope for a^*'s in Figure 1 and the intercept should be zero. This is the broken line in Figure 1. Relying on this b^* based slope, we get an indication of which items seem to deviate from the residual variance invariance specification.

Following the Muthén and Christoffersson (1981) methodology, we may now perform a simultaneous analysis of females and males. Here we may first test the hypothesis of all thresholds and loadings being sex-invariant, jointly. With 78 degrees of freedom we obtained a chi-square value of 79.38. Hence, there is no indication of deviation from this invariance hypothesis. The methodology allows us to go further and test more restrictive hypotheses. In the computer software used here, the residual variances do not appear as parameters but instead $\sigma^{-1/2}$, using the variances of the latent response variables y^*. However, given invariance of loadings, but allowing for differing factor variances, invariance of residual variances would imply that these σ's are equal over items, but possibly different over sex. Testing the hypothesis of residual variance invariance in this way results in a chi-square of 161.22 with 87 degrees

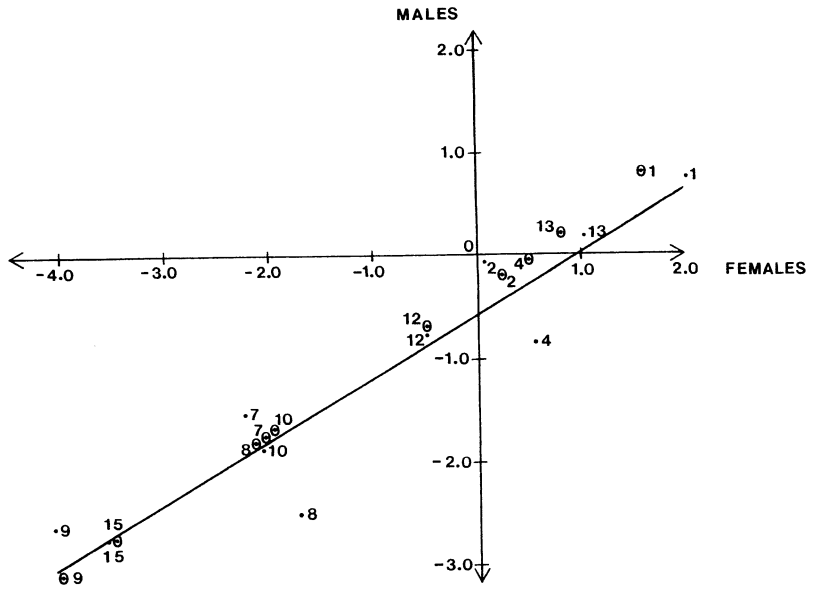

FIGURE 2. Plot of b^*'s for ten items. Correlation = 0.94, Estimated regression line (solid line): Intercept = −.58, Slope = .60

of freedom. In terms of the traditional two-parameter normal ogive, these results imply that the hypothesis of invariant a's does not hold, but only invariance of b's. The estimated b^*'s from the model with 78 degrees of freedom fall along a straight line and are denoted by circles in Figure 2. It is seen that the corresponding dots from the two single-group analyses are well approximated by this line.

We have now rejected the hypothesis of overall invariance of residual variances. Hence, a certain form of item bias has been detected in this analysis; the ICC's are not all the same across sex. One could go further and try to single out those items that contribute heavily to the rejection of this hypothesis. The plot in Figure 1 and the estimates in Table II can guide in such an effort. We do not feel knowledgeable enough about the items or the subjects to try to intepret the cause of the type of bias found. Also, we do not feel that this small, illustrative example warrants a serious substantive conclusion. A further important caveat is that even among the items used there may be a small amount of guessing remaining, which will affect the a parameter and thereby the d parameter estimation.

As discussed previously, deletion of items is not necessary in this case since the structural parameters for the groups can still be estimated under this type of partial non-invariance. In Table II the estimates are given for the model

with 78 degrees of freedom. Standard errors of estimates are given in parentheses. The ratios estimate/standard error have approximately standard normal distributions in large samples. At the bottom of Table II the estimates of the structural parameters are given. The factor mean is standardized to zero for females. The male mean is significantly larger than zero, implying a higher mean ability for this group. In the metric of the male ability, the estimated male mean corresponds to roughly one-third of the estimated standard deviation. The ability variance obtained a higher estimated value for males. A test of ability variance equality over sex was performed by rerunning the analysis, setting the variances equal. A chi-square difference test of this equality, however, resulted in a strong rejection, obtaining the value 13.41 with one degree of freedom.

Discussion

In these analyses, the more relaxed form of item parameter invariance was found to be valuable. Despite a certain form of item bias detected, a well fitting multiple-group model was found, yielding precise estimates of structural parameters. The approach does not necessitate the estimation of individual ability estimates. However, it is quite possible to go further and perform

TABLE II

Estimated Parameters for the Simultaneous Analysis of Females and Males[a]

	Common Measurement Parameters		$\sigma^{-1/2b}$		Residual Variance	
Item	Threshold	Loading	Females	Males	Females	Males
1	.554 (.039)	1.000[c] (—)	1.000[c] (—)	.553 (.091)	.886	3.035
2	.079 (.034)	1.075 (.218)	1.000[c] (—)	.883 (.199)	.868	1.011
4	.162 (.037)	.965 (.212)	1.000[c] (—)	.237 (.134)	.894	17.585
7	−.953 (.044)	1.355 (.252)	1.000[c] (—)	.520 (.049)	.791	3.267
8	−1.123 (.046)	1.562 (.267)	1.000[c] (—)	.882 (.066)	.722	.712
9	−1.491 (.056)	1.103 (.228)	1.000[c] (—)	.865 (.056)	.861	1.051
10	−1.062 (.045)	1.458 (.252)	1.000[c] (—)	.872 (.064)	.758	.816
12	−.278 (.036)	1.400 (.258)	1.000[c] (—)	.763 (.115)	.777	1.257
13	.278 (.035)	.957 (.186)	1.000[c] (—)	1.037 (.190)	.896	.715
15	−1.047 (0.45)	.877 (.179)	1.000[c] (—)	1.124 (.078)	.912	.611

Structural Parameters

	Females	Males
Factor mean	0.000[c] (—)	.174 (.045)
Factor variance	.114 (.033)	.235 (.071)

[a] Standard errors are given in parentheses.
[b] Inverted standard deviation for the latent response variable.
[c] Parameter fixed for identification purpose.

such estimation given the item parameter estimates. It should be noted that other types of structural parameters may be of interest. In our application, the ability variable was in effect related to a categorical variable, the dichotomous variable sex. In the more general framework of Muthén (1984), we may also consider structural regression coefficients, relating the ability variable to other continuous variables. Furthermore, multiple ability variables may be considered. Finally, some item analysis applications may require multiple-group estimation where only a subset of items is in common for the groups, for example, in linking test results via anchor items. Such situations can also be handled. We have noted the limitation of the above modeling to situations with negligible amounts of guessing. The concept of d parameters can in principle be generalized to situations with guessing, and such developments would seem to be desirable.

References

Comber, L. C., & Keeves, J. P. (1973). *Science education in nineteen countries* (International Association for the Evaluation of Educational Achievement: International Studies in Evaluation I). Stockholm: Almquist & Wiksell.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29,* 177–185.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46,* 407–419.

Muthén, B., & Jöreskog, K. G. (1983). Selectivity problems in quasi-experimental studies. *Evaluation Review, 7,* 139–174.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5,* 159–173.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology.* Amsterdam: Swets and Zeitlinger.

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

Authors

BENGT MUTHÉN, Associate Professor, Graduate School of Education, UCLA, Los Angeles, CA 90024. *Specializations*: Latent variable structural equation modeling with categorical and other nonnormal data.

JAMES D. LEHMAN, PhD Candidate, Graduate School of Education, UCLA, Los Angeles, CA 90024. *Specialization*: Measurement.

Journal of Educational Statistics
Summer 1985, Volume 10, Number 2, pp. 143–160

STANDARD ERRORS OF EQUIPERCENTILE EQUATING FOR THE COMMON ITEM NONEQUIVALENT POPULATIONS DESIGN

DAVID JARJOURA and MICHAEL J. KOLEN
The American College Testing Program

KEY WORDS. *Equipercentile equating, common item equating, standard errors, standard errors of equating.*

ABSTRACT. An equating design in which two groups of examinees from slightly different populations are administered different test forms that have a subset of items in common is widely used. A procedure for equipercentile equating under this design has been previously outlined, but standard errors for this rather complex procedure have not been provided. This paper provides these standard errors and a simulation that verifies the equations for large samples. A real data example is provided for considering issues involved in using these procedures.

Common item equating with random samples from nonequivalent populations as described by Angoff (1971, pp. 579–583) refers to an equating design in which two groups of examinees from slightly different populations are each administered different test forms that have a subset of items in common. The subset of items that are in common for the two forms (and thus the two groups) are used to equate one test form to the other. Typically some linear equating procedure is used (e.g., Tucker or Levine Equally Reliable, Angoff, 1971). More recently, item response theory procedures for observed score equating have been described (Lord, 1982a).

For this design, an equipercentile (curvilinear) equating procedure, referred to as the frequency estimation method by Angoff (1971), was outlined by Braun and Holland (1982, pp. 21–23). Very little research on this procedure has been reported, and it is often thought of as impractical because it is presumed that very large sample sizes are necessary, even for reasonably accurate results. However, asymptotic standard errors have not been derived previously, so the practicality of this procedure with typical test administration sample sizes is questionable.

In this paper, standard errors of equipercentile equating under the common item nonequivalent populations design are derived and the equations are evaluated using simulated and actual test data. The magnitude of the standard errors is used to judge the practicality of the equipercentile procedure.

The common item nonequivalent populations design is regularly used in testing programs that have a large, continuously growing bank of items. In such programs, a single new form of a test is typically administered on every