Variance Estimation for Converting MIMIC Model Parameters to IRT Parameters in DIF Analysis

Randall MacIntosh, California State University, Sacramento

Sabrina Hashim, E*Trade

The purpose of this study is to document the delta method to compute the standard error of the estimates of the converted item response theory (IRT) discrimination and difficulty parameters derived from multiple-indicator, multiple-causes (MIMIC) model parameters. Discussed is the formulation of MIMIC models to explore differential item functioning in Mplus and how to obtain factor-analytic estimates that are converted easily into IRT parameters. Also described are the partial derivatives necessary to apply the delta method to estimate variances for the converted parameters. Both item difficulty and discrimination parameters estimated from MIMIC parameters were very close to the Multilog estimates. The variance estimates for most parameters were similar as well. *Index terms: MIMIC model, delta method, variance estimation, Mplus, item response theory, differential item functioning.*

Introduction

The correspondence between factor-analytic models and item response theory (IRT) models has become increasingly well documented in the literature. Takane and de Leeuw (1987) provide proofs that a factor analysis of dichotomous variables is equivalent to a two-parameter normal ogive IRT model. This equivalence is exploited by Muthen and colleagues in a series of studies (Muthen, 1985, 1988; Muthen, Kao, & Burstein, 1991; Muthen & Lehman, 1985) to assess differential item functioning (DIF) within a multiple-indicator, multiple-causes (MIMIC) model (Hauser & Goldberger, 1971; Jöreskog & Goldberger, 1975; Jöreskog & Sörbom, 1995a, 1995b).

The MIMIC model integrates causal variables with a confirmatory factor analysis. Muthen et al. (1991) provide equations to convert the MIMIC model parameters into item discrimination (a) and item difficulty (b) parameters for a two-parameter IRT model. The primary purpose of this study is to document a method to estimate the statistical properties of the converted IRT parameters derived from MIMIC parameters. The converted parameters are compared to those produced by IRT software.

MIMIC Models

The latent variable, η, is influenced by observed exogenous background variables x and a dummy variable z:

$$\eta = \gamma_x' X + \gamma_z' Z + \zeta \quad (1)$$

Applied Psychological Measurement, Vol. 27 No. 5, September 2003, 372–379
DOI: 10.1177/0146621603256021

The error term, ζ, is normally distributed and independent of x and z.

The indicators, y_j, are realizations of a latent response variable y_j^*. When a threshold, τ_j, is exceeded, $y_j = 1$ but is otherwise 0:

$$y_j = 1, \text{ if } y_j^* > \tau_j. \tag{2}$$

The continuous latent response variable is explained by the latent trait variable, η, and a direct effect from the dummy variable, z_k, which measures the sociodemographic characteristic that is suspected of causing differential item functioning:

$$y_j^* = \lambda_j \eta + \beta_j z_k + \varepsilon_j. \tag{3}$$

MIMIC to IRT

Following Muthen et al. (1991), the conversions from the MIMIC model parameters to IRT parameters are as follows:

$$a_j = \lambda_j (1 - \lambda_j^2 \psi)^{-1/2} \sigma_{\eta\eta}^{1/2}, \tag{4}$$

$$b_{jk} = [(\tau_j - \beta_j z_k)\lambda_j^{-1} - \mu_\eta]\, \sigma_{\eta\eta}^{-1/2}. \tag{5}$$

The latent variable, η, has a mean of μ_η and a variance of $\sigma_{\eta\eta}$. In Equation (4), a_j and λ_j are, respectively, the discrimination parameter and factor loading for item j, and ψ is the error variance for the latent variable. In Equation (5), the item difficulty parameter (b_{jk}) is a function of the item threshold (τ_j), the factor loading, factor mean, and factor variance. Equation (5) also includes a group indicator dummy variable term ($\beta_j z_k$) that allows for the DIF test. If the group dummy is significant, it is evidence that the item difficulty shifts along the latent continuum for group k, depending on the sociodemographic characteristic measured by the indicator, such as race or gender. Equations (4) and (5) can be simplified greatly if the latent variable is standardized to $\mu_\eta = 0$ and $\sigma_{\eta\eta} = 1$. This study uses Mplus (Muthen & Muthen, 1998-2001). The application of the delta method is not described in the software documentation. The steps to carry the latent variable standardization out in Mplus are described below in a later section.

Advantages of This Psychometric Approach

Muthen (1988) lists several advantages of this approach. First, it permits the estimation of IRT-type parameters, including DIF estimates. But it also permits the testing of hypotheses and estimation of how background variables influence the latent trait. It allows for the *simultaneous* consideration of how the background variables may reflect subject heterogeneity and lead to DIF. Furthermore, these background variables may be categorical or continuous. Finally, it permits the testing and relaxation of the IRT requirement of unidimensionality and conditional independence by generalizing the model and extending it by allowing for the introduction of multiple latent variables. More details and examples may also be found in Muthen (1985); Muthen, Grant, and Hasin (1993); Muthen et al. (1991); Larson (2000); and Mast and Lichtenberg (2000).

Variance Estimation

As shown above, these converted IRT parameters are nonlinear functions of MIMIC model parameter estimates, which are random variables. The new quantities are also random variables and have some sampling variance as well. But variance estimates for the converted parameters require computations that are not yet programmed into standard statistical or modeling programs.

Delta Method

The delta method provides an estimate of the variance of the new random variable, which is a complex function of other random variables. This is accomplished by pre- and postmultiplication of the original parameters' variance-covariance matrix by a row vector of partial derivatives. The partial derivatives are taken for the new variable with respect to the parameters in Equations (4) and (5). For example, the delta method applied to the item discrimination parameter (a) is

$$\text{Var}(\hat{a}) = \mathbf{DSD}', \tag{6}$$

where **S** is the variance-covariance matrix of the MIMIC parameters in Equation (4), and **D** is a vector of partial derivatives of a with respect to λ, ψ, and σ. When the latent variable is standardized, $\sigma_{\eta\eta} = 1$ and is dropped from Equation (4). Setting the latent variable's variance equal to 1 is accomplished by fixing ψ to a constant value as explained below. Then the discrimination parameter can be expressed in terms of *only* the factor loading (λ_j) and a constant (ψ). In this case, the **D** matrix of derivatives becomes simply a scalar (noted as d) as there is only one partial derivative, $\partial a/\partial\lambda_j$ (a with respect to λ_j). The variance-covariance matrix is similarly simplified and becomes just the variance of λ_j. The resulting computation is

$$\text{Var}(a) = d^2\ \text{Var}(\lambda_j). \tag{7}$$

The partial derivative of a with respect to λ is

$$d = \frac{\partial a}{\partial \lambda} = (1 - \lambda^2\psi)^{-3/2}. \tag{8}$$

Similarly, the delta method may also be applied to the difficulty parameter in Equation (5). Again, assuming the latent variable has been standardized, the difficulty parameter is expressed in terms of the threshold, the DIF coefficient (where appropriate), and the factor loading. The partial derivatives of the item difficulty parameter (b) with respect to τ, βz, and λ are as follows (subscripts have been dropped to ease presentation):

$$\frac{\partial b}{\partial \tau} = \lambda^{-1}, \tag{9}$$

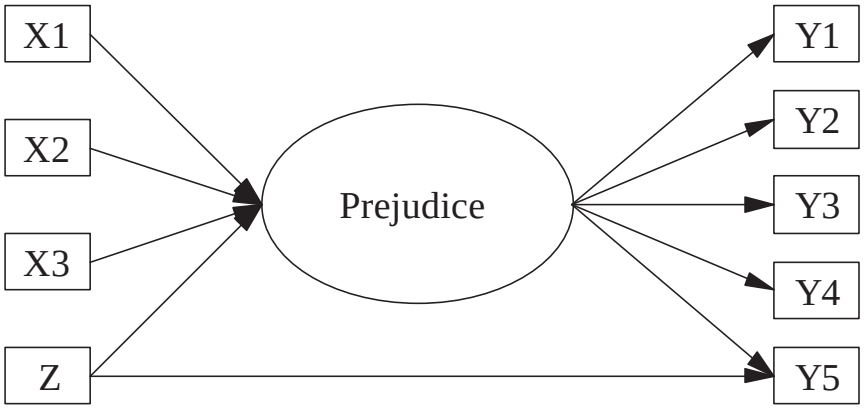
$$\frac{\partial b}{\partial \beta z} = -\lambda^{-1}, \tag{10}$$

$$\frac{\partial b}{\partial \lambda} = (\beta z - \tau)\lambda^{-2}, \tag{11}$$

Equation (11) is further simplified for the dummy variable reference category by dropping the βz term. In this case, the Equation (10) derivative is also dropped from the **D** matrix when $z_k = 0$ and for items when the βz term is not significant. In the latter case, the model can be reestimated and any nonsignificant DIF parameters dropped from the final model. For items without DIF and for the reference category for biased items, variances can be estimated using a vector of derivatives shown in Equations (9) and (11) (which, as noted above, is modified by dropping the βz term because $z_k = 0$). For items with significant DIF, the delta vector includes all three derivatives in Equations (9) through (11), as shown above for the group scored $z_k = 1$.

The delta method described above is a first-order approximation using a Taylor expansion series (Rao, 1973; Wolter, 1985). Second-order and higher order approximations are available by extending the Taylor series expansion. The first-order approximation, however, typically yields satisfactory results. Wolter (1985) warns that it may not be satisfactory for application to highly skewed populations. An application of the delta method can be found in Bollen and Stine (1990), and the method is also discussed in Stuart and Ord (1987).

Figure 1
A Multiple-Indicator, Multiple-Causes (MIMIC) Model for Differential Item Functioning (DIF) Analysis



Standardizing the Latent Variable

As noted above, the computations are greatly simplified if the latent variable is standardized ($\mu_\eta = 0$ and $\sigma_{\eta\eta} = 1$). In Mplus, the mean of the latent variable is zero when the exogenous predictor variables in Figure 1 (the X variables) are recoded by obtaining deviation scores from their respective means. Standardizing the latent variable's variance requires two steps. The model is run once, and the estimated R^2 for the latent variable is obtained. In a second run, the error variance of the latent variable (ψ) is fixed to $1 - R^2$, estimated in the first run. Success in properly fixing this value can be verified by checking the variance of the latent variable in the optional Tech 4 output obtained in the second run. As this sets the scale for the latent variable, the program default that fixes the first item loading to 1.0 is to be overridden in the second run as well.

An Example

Data for this example are from the 1994 General Social Survey (GSS). Several questions probe racial prejudice among Euro-Americans toward African Americans. The items selected for this example are as follows (GSS variable names in brackets):

[RACSEG] "White people have the right to keep Blacks out of their neighborhoods if they want to, and Blacks should respect those rights."
[RACFEW] "Would you yourself have any objection to sending your children to a school where a few of the children are Black?"
[RACEPRES] "If your party nominated a Black for President, would you vote for him if he were qualified for the job?"
[RACMAR] "Do you think there should be laws against marriage between Blacks and Whites?"
[RACPUSH] "Blacks shouldn't push themselves where they are not wanted."

The items were recoded so that higher scores on the latent variable reflect higher levels of prejudice.[1]

[1]The prejudice scale items are initially recoded, so 1 = *prejudiced response* and 0 = *nonprejudiced response*. Note that several of the original items are Likert-type scales, and these are collapsed into dichotomies so the indicators conform with the theoretical measurement model. The predictors are EDUC, POLVIEWS, FUND, and SEX. The

An inverse relationship is hypothesized between education (X1) and prejudice, whereas political conservatism (X2) and religious fundamentalism (X3) are expected to be positively related to racial prejudice. Gender differences are not anticipated in the overall level of prejudice, but this variable is included to permit a test of whether DIF may be present for some items because of differences in sex role norms and socialization. There are a total of 517 people with complete data on all the variables. The MIMIC model in Figure 1 shows the three X variables and the Z (gender) variable as influencing the prejudice latent variable, which is measured by the five observed indicators. There is a direct effect between gender (Z) and the RACPUSH scale item (Y5). If this coefficient is significant, it represents DIF, or item bias.

Results

This model represents a good overall fit to the data, χ^2 (16) = 20.8, $p > .19$. Table 1 shows the MIMIC parameters and standard errors. The parameters labeled lambda (λ_j) are the factor loadings, taus (τ_j) are the item thresholds, βz is the direct effect that represents the DIF coefficient, and the gammas (γ_x) are the effects of the independent variables on the latent variable. Panel A of Table 2 shows the IRT item discrimination and difficulty parameter estimates based on the MIMIC model parameters. The standard error columns for the converted parameters are also shown.

The delta variance estimate for the item Y1 (RACSEG) discrimination parameter is obtained by multiplying the squared derivative by the factor loading variance:

$$3.1458^{2*}.004096 = .040534.$$

The corresponding standard error is .201.

Similarly, the delta computations for the item Y1 (RACSEG) difficulty estimate require the lambda and tau MIMIC parameters' derivative matrix:

$$\mathbf{D} = [1.13895 - 1.39709],$$

and their covariance matrix taken from the Mplus output:

$$\mathbf{S} = \begin{bmatrix} .005625 & .000000 \\ .000000 & .004096 \end{bmatrix}.$$

Applying Equation (6) to these matrices produces var(b) = .01529 and a corresponding $SE(b)$ = .1236. To compute the variance estimates for the items that include the DIF parameter, expand the **D** vector to include the derivative of b with respect to βz (Equation (10)) and expand the parameter covariance matrix by a row and column to include the βz variance-covariance. Care must be taken to make sure the **D** vector and the rows and columns of the parameter covariance matrix correspond correctly when multiplying.

An examination of the converted IRT parameters shows that four of the items have similar discrimination parameter estimates in the range from about 1 to 1.5. The exception is item Y2 (RACFEW), which is not a very discriminating item. This is because it is an extremely difficult item to endorse. Its difficulty estimate of 4.62 is several standard errors beyond the next most difficult item, as only 12 of the 517 survey participants endorsed RACFEW.

---

trichotomous religious fundamentalism variable is collapsed into a dichotomy (1 = *fundamentalist*, 0 = *moderate or liberal*). Gender is recoded (1 = male, 0 = female). EDUC and POLVIEWS were not collapsed and are treated as interval-level measures. After collapsing as explained above, all four independent variables are subjected to a final recode by deviating each from its respective sample mean to establish a zero mean for the latent variable. The sample means are EDUC = 13.59, SEX(Male) = .46, POLVIEWS = 4.07, and FUND = .25.

Table 1

Multiple-Indicator, Multiple-Causes (MIMIC) Model Estimates

	Estimate	SE
λ		
Y1	.878	.064
Y2	.437	.143
Y3	.785	.089
Y4	.951	.073
Y5	.816	.074
βz		
Y5	.304	.132
τ		
Y1	1.077	.075
Y2	2.018	.128
Y3	1.304	.086
Y4	1.231	.089
Y5	.033	.059
γ		
Education	–.161	.023
Political Conservatism	.123	.039
Fundamentalist	.423	.121
Male	–.005	.124
Ψ		
Prejudice	.693	

The DIF coefficient for item Y5 (RACPUSH) was found to be significant. About 55% of the men endorsed this item, compared to only 44% of the women. Table 2 shows how this translates into different item difficulty estimates. The *b* parameter for men is –.332 and .040 for women.

Panel A of Table 2 also shows the standard errors for the IRT parameters estimated using the delta method. Panel B shows parameter estimates produced by Multilog (Thissen, 1991). A two-parameter logistic (2PL) model was specified in which item difficulties vary by gender for item Y5. The Multilog discrimination estimates and their standard errors are divided by 1.7 to place them on the same scale as the converted IRT parameters in Panel A. A similar comparison of other data by Muthen et al. (1991) found consistency between the discrimination estimates produced by converting MIMIC parameters and those produced by Bilog[2]. They also found larger discrepancies

[2]The two-parameter logistic model was estimated via Multilog, a widely used item response theory estimation program that was available to the author, whereas Bilog was not. The problem statement used was >PRO RA IND NI=10 NG=2 NE=517;>TEST ALL L2;>EQU AJ ITEMS=(6,7,8,9,10) WITH=(1,2,3,4,5);>EQU BJ ITEMS=(6,7,8,9) WITH=(1,2,3,4);>END.

Table 2

Converted IRT Parameters, Variances, and 2PL Estimates

A. MIMIC to IRT Conversion Item	*a*	SE Delta	*b*	SE Delta
Y1	1.29	.201	1.23	.124
Y2	.469	.177	4.62	1.55
Y3	1.04	.182	1.66	.200
Y4	1.56	.320	1.30	.126
Y5 Female	1.11	.187	.040	.072
Y5 Male	1.11	.187	–.332	.183
B. Multilog 2PL Estimates				
Item	a^a	SE^a	*b*	*SE*
Y1	1.41	.21	1.19	.11
Y2	.512	.19	4.65	1.76
Y3	1.08	.18	1.61	.18
Y4	1.34	.20	1.23	.11
Y5 Female	1.13	.13	.18	.11
Y5 Male	1.13	.13	–.29	.12

Note. MIMIC = multiple-indicator, multiple-causes; IRT = item response theory; 2PL = two-parameter logistic model.
a. Multilog discrimination parameter estimates and their standard errors have been divided by 1.7 to place them on the same scale as the converted MIMIC/IRT parameters in Panel A.

between the two techniques for difficulty parameters. Here, a correspondence between both sets of parameters is found. Note that the variance estimates are also similar. The greatest exception is found for the Y4 discrimination parameter as the Multilog standard error estimate is about one third smaller than the value produced by the delta method. Differing types of estimation may explain some of the discrepancies. Multilog uses marginal maximum likelihood, and Mplus uses a weighted least squares method.

Discussion and Recommendations

Muthen et al. (1991) show that IRT parameters converted from MIMIC models are comparable in magnitude to those produced by widely used IRT software. Our results found that both item difficulties and discrimination parameter values estimated from MIMIC parameters were very close to the Multilog estimates. For most parameters, the variance estimates were similar as well. The computations shown for the delta method variance estimates may be carried out easily in a spreadsheet.

Care must be used by those who wish to apply this method, however, when obtaining the parameter covariances from the Mplus Tech 3 output. Parameters are identified by number in the Mplus output. The program places the DIF coefficient among the gamma coefficients, and the factor loading for any item tested for DIF is in the beta coefficient matrix.

References

Bollen, K. A., & Stine, R. A. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115-140.

Hauser, R. M., & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. *Sociological Methodology*, *2*, 81-117.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *10*, 631-639.

Jöreskog, K. G., & Sörbom, D. (1995a). *LISREL 8 user's reference guide*. Chicago: Scientific Software International, Inc.

Jöreskog, K. G., & Sörbom, D. (1995b). *Prelis 2 user's reference guide*. Chicago: Scientific Software International, Inc.

Larson, S. L. (2000). Rural-urban comparisons of item responses in a measure of depression. *Dissertation Abstracts International Section A: Humanities & Social Sciences*, *60*(8-A), 3157.

Mast, B. T., & Lichtenberg, P. A. (2000). Assessment of functional abilities among geriatric patients: A MIMIC model of the functional independence measure. *Rehabilitation Psychology*, *45*(1), 49-64.

Muthen, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, *10*, 121-132.

Muthen, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer and H. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Lawrence Erlbaum.

Muthen, B. O., Grant, B., & Hasin, D. (1993). The dimensionality of alcohol abuse and dependence: Factor analysis of DSM-III-R and proposed DSM-IV criteria in the 1988 National Health Interview Survey. *Addiction*, *88*(8), 1079-1090.

Muthen, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: An application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*, 1-22.

Muthen, B. O., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics*, *10*, 133-142.

Muthen, L. K., & Muthen, B. O. (1998-2001). *Mplus user's guide*. Los Angeles: Muthen & Muthen.

Rao, C. R. (1973). *Linear statistical inference and its application*. New York: John Wiley.

Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics* (Vol. 1). New York: Oxford University Press.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Thissen, D. (1991). *Multilog user's guide*. Chicago: Scientific Software International, Inc.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

Acknowledgments

The authors wish to acknowledge helpful comments by Bengt Muthen and Alex Piquero on an earlier draft.

Author's Address

Address correspondence to Randall MacIntosh, Sociology Department, 6000 J Street, Sacramento, CA 95819-6005, e-mail: rmacintosh@csus.edu. The Mplus command file used in this study may be viewed at http://webpages.csus.edu/~rmac/.