

A Note on the Conversion of Item Parameters Standard Errors

Chun Wang

University of Washington

Xue Zhang

Key Laboratory for Applied Statistics (KLAS), School of Mathematics and Statistics, Northeast Normal University

Both authors made **equal** contributions to the paper, and the order of authorship is alphabetical. Correspondence concerning this manuscript could be addressed to either Chun Wang at: 312E Miller Hall, College of Education, University of Washington, 2012 Skagit Ln, Seattle, WA 98105, wang4066@uw.edu, or Xue Zhang at: School of Mathematics and Statistics, Northeast Normal University, 5268 Renmin Road, Changchun, Jilin Province, China, zhangx815@nenu.edu.cn.

The project is supported by Institute of Education Sciences grant R305D170042 (originally R305D160010), National Science Foundation grant SES-165932, and National Natural Science and Social Science Foundations of China 11571069.

Citation: Wang, C., & Zhang, X. (2019). A note on the conversion of item parameters standard errors. *Multivariate Behavioral Research*, 54, 307-321.

Related code can be downloaded at: <https://sites.uw.edu/pmetrics/publications-and-source-code/>

A Note on the Conversion of Item Parameters Standard Errors

Abstract

The relations among alternative parameterizations of the binary factor analysis (FA) model and two-parameter logistic (2PL) item response theory (IRT) model have been thoroughly discussed in literature (e.g., Lord & Novick, 1968; Takane & de Leeuw, 1987; McDonald, 1999; Wirth & Edwards, 2007; Kamata & Bauer, 2008). However, the conversion formulas widely available are mainly for transforming parameter estimates from one parameterization to another. There is a lack of discussion about the standard error (SE) conversion among different parameterizations, when SEs of IRT model parameters are often of immediate interest to practitioners. This paper provides general formulas for computing the SEs of transformed parameter values, when these parameters are transformed from FA to IRT models. These formulas are suitable for unidimensional 2PL, multidimensional 2PL, and bi-factor 2PL models. A simulation study is conducted to verify the formula by providing empirical evidence. A real data example is given in the end for an illustration.

Key words: item response theory, standard error, factor analysis

Standard errors of item parameter estimates in item response theory (IRT), or more generally error covariance matrices, play an important role because the uncertainty in item parameter estimates is often carried over in subsequent analysis, such as test form assembly, IRT scoring, equating and linking (e.g., Cheng & Yuan, 2010; Mislevy, Wingersky, & Sheehan, 1993; Thissen & Wainer, 1990). Moreover, obtaining the SEs is also a prerequisite for conducting hypothesis testing to evaluate differential item functioning (Cai, Yang, & Hansen, 2011; Woods, Cai & Wang, 2013) or item parameter drift (Bock, Muraki, & Pfeifferberger, 1988); as well as developing asymptotic adjustments for limited-information goodness-of-fit statistics (Cai, 2008; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006).

In IRT when full information maximum likelihood (FIML) estimation method is used, the parameter error covariance matrix can be computed as the inverse of the Fisher information matrix. Based on the statistical theory from the standard discrete multivariate analysis (Rao, 1973), the error covariance matrix computed this way is considered “gold standard” (Tian, Cai, Thissen, & Xin, 2012). However, the computation burden of this Fisher information based SE (FISE) increases drastically as test length increases because the number of possible response patterns increases exponentially. Two computationally feasible alternatives are the empirical cross-product approach (XPD) and the supplemented expectation maximization (SEM) approach (Cai, 2008). Past research has demonstrated that XPD, though computationally most efficient, works well only when sample size is much larger than test length, otherwise it produces upward bias (Paek & Cai, 2014). On the other hand, the SEM approach is based on numerically differentiating an implicit function defined by EM iterations (a.k.a., EM map) and it generally performs well under a variety of different conditions.

Item factor analysis (FA) which is rooted in categorical confirmatory factor analysis

offers an alternative to IRT parameterization. It assumes that ordered-categorical item responses are discrete representations of underlying continuous latent responses. Different from FIML that is often used in IRT framework, weighted least squares (WLS) estimation is usually adopted within the FA framework. A challenge with WLS is that the size of the optimal weight matrix becomes exceedingly large and increases rapidly when test length increases, hence a statistically less efficient yet computationally more feasible alternative is the diagonally weighted least squares method (Satorra & Bentler, 1990; Wirth & Edwards, 2007). This reduction in efficiency leads to biased standard errors and hence the robust standard error is recommended.

Previous research has discussed the advantages and limitations of both IRT and FA frameworks, along with their preferred estimation algorithms, i.e., FIML versus WLS¹ (Wirth & Edwards, 2007). Transformation formulas are available for practitioners to transform parameter estimates from one framework to the other (Kamata & Bauer, 2008; Wang, Kohli, & Henn, 2016). However, when the parameters from FA model are transformed to the IRT parameterization, there is a lack of documentation on how to compute the SEs of the transformed parameters (i.e., convert the standard errors) accordingly, and this note tends to fill the gap. Forero and Maydeu-Olivares (2009) mentioned of using the delta method for conversion, but the details were not provided. We believe the conversion formulas will be useful in at least three scenarios: (1) when researchers choose to estimate the FA model via WLSMV (either because of their familiarity with the FA model or because WLSMV is much faster with high-dimensional models) but later want to report the IRT item parameter estimates, the SEs of the transformed parameter values can be obtained directly from the plug-in equations provided in this note; (2)

¹ Other estimation methods are also available, such as Monte Carlo EM, Markov chain Monte Carlo, etc., but they are not the focus of this note.

when researchers want to compare the performance of a sandwich estimator versus XPD/SEM, the transformation is needed to put SEs from the two models on the same metric; (3) when researchers want to compare the performance of WLSMV and MML with respect to the SEs of item parameter estimates, the transformation formula facilitates a direct comparison, whereas prior studies had to conduct hundreds of replications to obtain the empirical standard deviation (e.g., Finch, 2010; DeMars, 2012).

To align with Kamata and Bauer's (2008) argument, in this note, the conversion formulas are provided for four different FA parameterizations (marginal vs. conditional, and reference indicator vs. standardized factor) that cover the majority of the applications. The conversion formulas are also general enough to be used with unidimensional, multidimensional, and bi-factor models. The rest of the paper is organized as follows. We first briefly introduce the IRT model and factor analytic model along with the four parameterizations. These four parameterizations were extensively discussed in Kamata and Bauer (2008). Then we discuss the standard error transformation and provide the conversion transformation table. The simulation study is then given, followed by a real data example. The final conclusion is given in the end.

IRT Models, FA Models, and Four Parameterizations

In this section, we will briefly introduce the IRT models, factor analytic models, and the four different parameterizations (Kamata & Bauer, 2008).

IRT and FA Models

Starting with the simplest unidimensional IRT (UIRT) two-parameter model, the item response function is defined as

$$p_{ij} = p(y_{ij} = 1 | a_i, d_i, \theta_j) = f(a_i \theta_j + d_i) \quad (1)$$

where y_{ij} denotes the item response of person j to item i . In Equation (1), a_i and d_i are known as

the item slope (discrimination) and intercept (threshold) parameters respectively, whereas θ_j denotes the latent trait of person j . $f(\cdot)$ is the cumulative distribution function (CDF), chosen to be either a normal ogive or logistic CDF. Here we use the intercept-slope notation to make the notation consistent with the other models. For instance, the multidimensional IRT (MIRT) is a natural extension of UIRT, with the item response function defined as

$$p_{ij} = p(y_{ij} = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}_j) = f(\mathbf{a}_i^T \boldsymbol{\theta}_j + d_i), \quad (2)$$

where $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jK})^T$ denotes a column vector of K latent traits, and \mathbf{a}_i is a vector of K slope parameters. This notation is used throughout Reckase (2009) (Equation 4.5, p. 86). In a general MIRT model, an item can load on either one of the K dimensions (i.e., simple structure) or on multiple dimensions (i.e., complex structure), and all dimensions of $\boldsymbol{\theta}$ are correlated.

The bi-factor model originally proposed by Holzinger and Swineford (1937) and popularized by Gibbons and Hedeker (1992) represents a unique type of factor structure. In this structure, each item loads on one general factor and one specific factor, and all factors are independent (Reise, 2012). Hence, only two elements in \mathbf{a}_i are non-zero.

For readers interested in the traditional discrimination-difficulty notation for the unidimensional two-parameter model,

$$p(y_{ij} = 1 | a_i, b_i, \theta_j) = f(a_i(\theta_j - b_i)), \quad (3)$$

or the simple structure MIRT model assuming item i measures the k th latent trait, i.e.,

$$p(y_{ij} = 1 | a_i, b_i, \theta_j) = f(a_{ik}(\theta_{jk} - b_{ik})), \quad (4)$$

the conversion formula for computing the SE of b_i (or b_{ik}) are provided as well.

All different two-parameter IRT models can be reparameterized in the FA framework.

Let y_{ij}^* denote the continuous latent response variable governing the observed binary response y_{ij} , then y_{ij}^* is written as the additive linear form

$$y_{ij}^* = \nu_i + \lambda_{i1}\xi_{j1} + \dots + \lambda_{iK}\xi_{jK} + \varepsilon_{ij}. \quad (5)$$

In Equation (5), ν_i is the intercept, usually fixed at zero for identification; λ_{ik} is the factor loading on the k th factor corresponding to item i ; ξ_{jk} is the latent factor score for individual j on factor k and ε_{ij} is the residual for person j on item i . y_{ij}^* is then dichotomized to form the binary observed y_{ij} based on the following rule

$$y_{ij} = \begin{cases} 1 & \text{if } y_{ij}^* \geq \tau_i \\ 0 & \text{if } y_{ij}^* < \tau_i \end{cases}, \quad (6)$$

where τ_i is the threshold for item i . If a unidimensional model is considered, then $K = 1$.

For bi-factor model, K equals the number of group factors plus one.

Four Parameterizations

The exposition in this section will closely mirror Kamata and Bauer (2008), readers familiar with this reference can skip this section. In item factor analysis, the scale of the latent response variable can be fixed by two different parameterizations. On one hand, the variance of y_{ij}^* is constrained to be 1 for all items such that the residual variance, $V(\varepsilon_{ij})$, is estimated as $V(\varepsilon_{ij}) = 1 - \lambda_i^T \text{cov}(\xi)\lambda_i$. This unit variance constraint for y_{ij}^* is rooted in the weighted least squares estimation method for binary FA, which involves the use of tetrachoric correlations (Kamata & Bauer, 2008). In fact, the tetrachoric correlation matrix is essentially a covariance matrix between underlying latent response variables with unit variance (Millsap & Yun-Tein, 2004). Following the naming convention in Kamata and Bauer (2008), this parameterization fixes the marginal distribution for the continuous latent variable y_{ij}^* and hence it is referred to as

the *marginal* parameterization. On the other hand, if fixing the residual variance $V(\varepsilon_{ij})$ to 1, then the marginal variance of y_{ij}^* is computed as $V(y_{ij}^*)=1 + \lambda_i^T cov(\xi)\lambda_i$. This parameterization is more in line with the convention in probit regression model, and it is referred to as the *conditional* parameterization.

As both the marginal and conditional parameterizations fix the scale of y_{ij}^* , to further identify the model, the scale of the latent factors also has to be fixed. Note that scaling y_{ij}^* is only needed for FA, whereas scaling ζ (or θ in IRT) is needed for both FA and IRT. Two widely used scaling conventions, in unidimensional scenario, are to standardize the common factor or to choose a reference indicator. When a multidimensional model is considered (bi-factor structure included), one must standardize all K factors where K refers to the total number of factors. With the choice of reference indicator, at least K items need to be selected as references whose parameters are fixed. In the confirmatory item factor analysis, when the item-factor loading structure is pre-determined, no further constraints are needed to remove rotational indeterminacy. Taken together, Table 1 summarizes four different parameterizations. In *Mplus* (Muthén & Muthén, 1998-2015), the marginal parameterization is notated as “DELTA” parameterization, and the conditional parameterization is notated as “THETA” parameterization.

Table 1. Summary of four parameterizations.

	Reference Indicator	Standardized Factor
Marginal (DELTA)	$\lambda_{ik} = 1, \tau_{ik} = 0$ for $k=1, \dots, K$ $V(y^*) = 1$	$E(\xi_k) = 0, V(\xi_k) = 1$, for $k=1, \dots, K$ $V(y^*) = 1$
Conditional (THETA)	$\lambda_{ik} = 1, \tau_{ik} = 0$ for $k=1, \dots, K$ $V(\varepsilon) = 1$	$E(\xi_k) = 0, V(\xi_k) = 1$, for $k=1, \dots, K$ $V(\varepsilon) = 1$

Note. (1) y^* and ε have no subscript, and it refers to all (i, j) 's.

The conversion between FA and IRT parameterization is well established in unidimensional model (e.g., Takane & de Leeuw, 1987), and multidimensional models

(e.g., McDonald, 1999; Finch, 2010). Following the same derivations in Kamata and Bauer (2008), we arrive at the general conversion formulas for multidimensional models in Table 2. When $K = 1$, the formulas are exactly the same as those in Table 2 of Kamata and Bauer (2008, p.144). In Table 2, $E(\xi_k)$ and $V(\xi_k)$ denote the mean and variance of the k th factor, ξ_k . λ_i denotes the column vector of loading parameters of item i , and $cov(\xi)$ denotes the covariance matrix of the factors. The last row in Table 2 refers to the discrimination-difficulty notation, in which the conversions for a -parameter stay the same as in the slope-intercept parameterization. The conversions for b -parameter are presented, and they are the same for both marginal and conditional parameterizations.

Table 2. Conversion formulas for four factor analysis parameters		
	Reference Indicator	Standardized Factor
Marginal	$a_{ik} = \frac{\lambda_{ik} V(\xi_k)^{1/2}}{\sqrt{1 - \lambda_i^T cov(\xi) \lambda_i}}$ $d_i = \frac{-[\tau_i - \lambda_i^T E(\xi)]}{\sqrt{1 - \lambda_i^T cov(\xi) \lambda_i}}$	$a_{ik} = \frac{\lambda_{ik}}{\sqrt{1 - \lambda_i^T cov(\xi) \lambda_i}}$ $d_i = \frac{-\tau_i}{\sqrt{1 - \lambda_i^T cov(\xi) \lambda_i}}$
Conditional	$a_{ik} = \lambda_{ik} V(\xi_k)^{1/2}$ $d_i = -[\tau_i - \lambda_i^T E(\xi)]$	$a_{ik} = \lambda_{ik}$ $d_i = -\tau_i$
Discrimination-Difficulty Notation (Equations 3 & 4)	$b_i = \frac{\tau_i - \lambda_i^T E(\xi)}{\lambda_{ik} V(\xi_k)^{1/2}}$	$b_i = \frac{\tau_i}{\lambda_{ik}}$

Note. When logistic CDF is used, all above transformation formulas need to be multiplied by a constant 1.7.

Standard Error Conversions

In the FA framework, the limited information WLS is often used. Instead of using the raw response patterns as in MMLE/EM, WLS uses the first-order and second-order marginal proportions obtained from the response contingency tables to facilitate parameter estimation. The

primary idea is to find item parameter values such that they minimize the weighted deviations between the model-implied correlation matrix and the sample tetrachoric correlation matrix. Because WLS usually requires a large sample to precisely estimate a full, optimal weight matrix (Muthén, du Toit, & Spisic, 1997), researchers have suggested using only the diagonals of the weight matrix for estimation, leading to the so-called diagonally weighted (or modified) WLS estimators. Due to this “misspecification” of the weight matrix, the resulting standard errors are biased. As a remedy, the robust standard error via the Huber sandwich estimator is used to correct for specification error (Satorra & Bentler, 1990; Muthén & Muthén, 2015). With high-dimensional models, WLS is much faster than FIML (e.g., Wang, et al., 2016; Wang, Su, & Weiss, 2018).

In this section, we provide the conversion formulas for SE transformation in Table 3. The multivariate delta method (Casella & Berger, 2002) is used to obtain the SE of the transformed FA parameter values when they are transformed to the IRT parameters in Table 2. For item i as an example, $a_i = g(\tau_i, \lambda_i, E(\xi), cov(\xi))$, and the specific form of the function $g(\cdot)$ is provided in Table 2. Then, given the error covariance matrix obtained via the sandwich estimator from a FA model, denoted as Σ_{FA} , the standard error of a_i , the transformed parameter, can be obtained via the multivariate delta method as follows

$$SE(a_i) = \sqrt{\mathbf{\Gamma}_{a_i}^T \times \Sigma_{FA} \times \mathbf{\Gamma}_{a_i}} . \quad (7)$$

In Equation (7), the superscript “T” denotes the matrix transpose, $\mathbf{\Gamma}_{a_i}$ are the first derivative of $a_i = g(\tau_i, \lambda_i, E(\xi), cov(\xi))$ with respect to the model parameters, i.e., $\lambda_i, \tau_i, E(\xi), cov(\xi)$. The error covariance matrix Σ_{FA} is a Q -by- Q matrix output from the WLS estimation procedure, where Q stands for the total number of parameters in a model. A concrete example is given in Table 5 below. The same generic form in (7) applies to d_i and b_i , but their specific forms differ

by the four parameterizations. The generic forms are presented in Table 3 and specific forms of Γ_{a_i} and $\Gamma_{a_i}(\Gamma_{b_i})$ are presented in the Appendix. This multivariate delta method is implemented in R (R core team, 2016), and the script file is available in the online appendix for interested users.

In Table 3, $vec\left(\frac{\partial a_{i1}}{\partial (E(\xi))_s}\right)$ denote the vectorized elements of the first derivatives, which is a K -by-1 vector with the s th element being $\frac{\partial a_{i1}}{\partial (E(\xi))_s}$. Similarly, taking the first derivative of a_{i1} with respect to the lower-triangular elements of factor covariance matrix $cov(\xi)$ and stacking them into a $K \times (K+1)/2$ -by-1 vector results in $vec\left(\frac{\partial a_{i1}}{\partial (cov(\xi))_{st}}\right)$. Here the (st) in the subscript denote the (s, t) th element of the covariance matrix. Please note that although the full size of Γ_{a_i} (or Γ_{a_i}) could be large, there will be a lot of 0's in the vector. That is, for item i , any derivatives with respect to λ_h and τ_h where item $h \neq$ item i are 0 by definition. For the full dimension of the Γ matrix, please refer to Table 5 for a few concrete examples. Otherwise, for the non-zero elements in the Γ matrix, please refer to Table 3.

Table 3. Generic forms of Γ_{a_i} and Γ_{b_i} in Equation (7)

Parameterization	Γ_{a_i}	Γ_{d_i}
Reference indicator	$\begin{bmatrix} \frac{\partial a_{i1}}{\partial \tau_2} & L & \frac{\partial a_{iD}}{\partial \tau_2} \\ M & & M \\ \frac{\partial a_{i1}}{\partial \tau_l} & L & \frac{\partial a_{iK}}{\partial \tau_l} \\ \frac{\partial a_{i1}}{\partial \lambda_{12}} & L & \frac{\partial a_{iK}}{\partial \lambda_{12}} \\ M & & M \\ \frac{\partial a_{i1}}{\partial \lambda_{iK}} & L & \frac{\partial a_{iK}}{\partial \lambda_{iK}} \\ \text{vec}\left(\frac{\partial a_{i1}}{\partial (E(\xi))_s}\right) & L & \text{vec}\left(\frac{\partial a_{iK}}{\partial (E(\xi))_s}\right) \\ \text{vec}\left(\frac{\partial a_{i1}}{\partial (\text{cov}(\xi))_{st}}\right) & L & \text{vec}\left(\frac{\partial a_{iK}}{\partial (\text{cov}(\xi))_{st}}\right) \end{bmatrix}$	$\begin{bmatrix} \frac{\partial d_i}{\partial \tau_2} \\ \vdots \\ \frac{\partial d_i}{\partial \tau_l} \\ \frac{\partial d_i}{\partial \lambda_{12}} \\ \vdots \\ \frac{\partial d_i}{\partial \lambda_{iK}} \\ \text{vec}\left(\frac{\partial d_i}{\partial (E(\xi))_s}\right) \\ \text{vec}\left(\frac{\partial d_i}{\partial (\text{cov}(\xi))_{st}}\right) \end{bmatrix}$
Remarks	<p>Because $\tau_1, \lambda_{11}, \dots, \lambda_{*K}$ are fixed as reference indicators, their relevant terms do not appear in the above matrix. In practice, the “reference indicators” do not have to be the first or first K items and hence “*” denotes the unspecified reference indicator items. For MIRT/bi-factor models, K loadings have to be fixed.</p>	
Standardized factor	$\begin{bmatrix} \frac{\partial a_{i1}}{\partial \tau_1} & L & \frac{\partial a_{iK}}{\partial \tau_1} \\ M & & M \\ \frac{\partial a_{i1}}{\partial \tau_l} & L & \frac{\partial a_{iK}}{\partial \tau_l} \\ \frac{\partial a_{i1}}{\partial \lambda_{11}} & L & \frac{\partial a_{iK}}{\partial \lambda_{11}} \\ M & & M \\ \frac{\partial a_{i1}}{\partial \lambda_{iK}} & L & \frac{\partial a_{iK}}{\partial \lambda_{iK}} \\ \text{vec}\left(\frac{\partial a_{i1}}{\partial (\text{cov}(\xi))_{st}}\right) & L & \text{vec}\left(\frac{\partial a_{iK}}{\partial (\text{cov}(\xi))_{st}}\right) \end{bmatrix}$	$\begin{bmatrix} \frac{\partial d_i}{\partial \tau_2} \\ M \\ \frac{\partial d_i}{\partial \tau_l} \\ \frac{\partial d_i}{\partial \lambda_{12}} \\ M \\ \frac{\partial d_i}{\partial \lambda_{iK}} \\ \text{vec}\left(\frac{\partial d_i}{\partial (\text{cov}(\xi))_{st}}\right) \end{bmatrix}$
Remarks	<p>Because the factor means are fixed as constants, the derivatives with respect to $E(\xi)$ disappear. In addition, although we write $\text{cov}(\xi)$ here, it is really the off-diagonal terms that matter because the variances are fixed as constants as well.</p>	

Note: The generic form of Γ_{b_i} is the same as that of Γ_{d_i} by replacing d_i with b_i .

A Simulation Study

A small scale simulation study was conducted to evaluate if the conversion formulas in Table 3 actually work in terms of computing SEs of the transformed FA parameters when they

are transformed to the equivalent IRT parameterizations (based on Table 2). The models considered were unidimensional two parameter logistic model (2PL), multidimensional 2PL (both simple and complex structures) and bi-factor 2PL. Four parameterizations were included. Please note that we do not intend to compare the transformed SE to the SE obtained directly from the IRT parameterization because the comparison would be confounded by the type of estimators, i.e., sandwich estimator from WLSMV vs. SEM/XPD from FIML. Instead, we intend to show the SEs of the transformed parameter values (obtained from the multivariate delta method) are close to the empirical, “true” SEs from simulations.

Design

For all simulations, logistic CDF is used as the link function in IRT model, and sample size was fixed at 1,000. For the unidimensional 2PL model, test length was fixed at 15. For the between-item multidimensional model, there were 45 items with 15 items measuring each one of the three latent traits separately. For the within-item multidimensional model, there were 45 items that measure multiple latent traits, and each latent trait was measured by 30 items. As an example, θ_1 was measured by items 1~5, 16~25, and 31~45. For the bi-factor model, there were 45 items which measure one general latent trait and one of the three group latent traits. The specific simulation design is shown in Table 4, with details regarding the distributions from which the parameters were simulated. Two hundred replications were conducted per condition². FA models were fitted using *Mplus*³, parameters are estimated by WLS with means and variance adjusted (i.e., WLSMV) along with the sandwich standard errors under all four parameterizations

² We conducted a sensitivity analysis and found that the empirical standard deviation stabilized at around 150 replications, hence 200 is a conservative choice.

³ *Mplus* was chosen because it is a widely used structural equation modeling software package. Other SEM software packages could also be used.

(Finney & DiStefano, 2006; Flora & Curran, 2004). The parameter estimates were transformed using the conversion formula in Table 2, and the SEs were transformed using the conversion formula in Equation (7) and Table 3, with Γ_{a_i} and Γ_{d_i} taking the following dimensions for different models:

Insert Table 4 Here

Table 5. The dimension of the Γ matrix in this study.

Model	UIRT	bi-factor	Between-item MIRT	Within-item MIRT
Γ_{a_i}	30×1	135×4	93×3	138×3
Γ_{d_i}	30×1	135×1	93×1	138×1

Note: $30=15$ (#of items in a test) $\times 2$ (#of parameters per item); $135 = 45$ (# of items in a test) $\times 3$ (#of parameters per item); $93=45$ (# of items in a test) $\times 2$ (#of parameters per item) $+3$ (correlations among factors); $138=30$ (# of items measuring each factor) $\times 3$ (# of factors) $+45$ (# of intercept parameters) $+3$ (correlations among factors). The number of columns in the Γ matrix is consistent with the dimension of the parameter vector, i.e., 1 in UIRT refers to 1 discrimination and 1 intercept parameters per item; 4 and 1 in the bi-factor model refer to 4 discrimination (1 general factor and 3 group factors) and 1 intercept parameters per item; 3 and 1 in both between-item and within-item MIRT refer to 3 discrimination and 1 intercept parameters per item.

Evaluation Criteria

To compare the SE conversions under four parameterizations, average root mean square error (RMSE), average bias, and average relative bias were calculated. The ARMSE is defined by

$$RMSE = \frac{1}{R} \sum_{r=1}^R \sqrt{\frac{1}{I} \sum_{i=1}^I (\hat{\sigma}_i^r - \sigma_i)^2} ,$$

where $\hat{\sigma}_i^r$ is the transformed SE estimate from the delta method for item i in replication r , and σ_i is the empirical standard deviation of transformed parameters across replications and it serves as the true value. Here, σ is used to denote the standard error of a generic parameter, which could be discrimination or threshold parameters. I denotes the total number of items, and R denotes the number of replications. Similarly, the average bias is computed by $bias = \frac{1}{R} \sum_{r=1}^R \left[\frac{1}{I} \sum_{i=1}^I (\hat{\sigma}_i^r - \sigma_i) \right]$,

and the average relative bias computed by $RB = \frac{1}{R} \sum_{r=1}^R \left[\frac{1}{I} \sum_{i=1}^I \frac{(\hat{\sigma}_i^r - \sigma_i)}{\sigma_i} \right]$. For quality control, we

also checked the parameter recovery from the FA parameterization (after transformation) as compared to the true parameter. The purpose is to check the behavior of the conversion formula in Table 2. The average RMSE, bias, and relative bias were also computed for the item parameter estimates.

Results

The results are summarized in Table 6 for unidimensional 2PL, in Table 7 for between-item multidimensional 2PL (M2PL), in Table 8 for within-item M2PL, and in Table 9 for bi-factor 2PL, respectively. In each table, per parameter, the first row is the parameter recovery and the second row refers to the standard error recovery. Note that for the reference indicator option, the reference items are excluded from computing the average bias and average RMSE. For the 2PL and between-item M2PL, results for both the b -parameter (i.e., Equations 3, 4) and the d -parameter (i.e., Equations 1, 2) are included.

Insert Tables 6 to 9 Here

For unidimensional IRT model, the four different parameterizations in the FA framework generated almost identical parameter estimates, which was indicated by the small average bias, relative bias, and RMSE. The standard error estimates in the FA framework, after transformation, aligned well with the “true”, empirical standard error. There was no appreciable difference among the four parameterizations in terms of standard error recovery for a - and d - and b - parameters. Overall, all parameterizations yielded satisfactory standard error recovery.

For the between-item MIRT model, all replications successfully converged but not all replications entered into the final results. In particular, for the marginal-standardized factor

parameterization, there were 17% replications with negative factor loadings, resulting in “NaN” for the transformed SEs (due to negative variance estimates after transformation). Even though these replications were excluded in our results, a closer inspection revealed that the negative factor loadings are due to the flipping (reversing) of the corresponding factor. For instance, when θ_1 was reversed, all items loaded on θ_1 had negative loadings and θ_1 was also correlated negatively with θ_2 and θ_3 . In this case, researchers could either manually multiply the parameters related to θ_1 by -1, or add non-negative constraints on factor loadings during estimation. Please note that the identifiability constraints in Table 1 do not preclude the possibility of factor flipping. Hence, in the marginal-standardized factor parameterization, researchers need to be aware that models could be equivalent up to factor reversing, and only if non-negative constraints are added is that the model strictly identified.

Similarly, 18% replications from the marginal-reference parameterization, and 30.5% replications from the conditional-reference parameterization were excluded due to the observed “NaN” from the SEs of the transformed d -parameters. There is no clear interpretation why the variance of the transformed d -parameters became negative, and one possible reason is that the original sandwich SE estimates of those parameters were relatively high. The results in Table 7 were therefore based on the remaining replications per parameterization. Again, overall, all four parameterizations resulted in acceptable parameter and standard error recovery with no noticeable differences, except the marginal-reference parameterization. In this case, the relative bias of standard error recovery is considerably higher. Even so, the actual bias is still acceptable. This is not surprising because marginal-reference combination requires the most complex transformation. Because the multivariate delta method relies on the first-order approximation of a Taylor series, it introduces some error by ignoring the higher-order terms. Please also note that

the transformation formula for the reference parameterization depends on the estimated mean and covariance of the latent factors. Hence any estimation error in the factor mean and covariance will be carried over in the transformed SEs. The SE recovery of the b -parameters was comparable to that of the d -parameters.

For within-item multidimensional models, we ran 200 replications among which only 131 replications converged properly from *Mplus*. The non-converged cases only happened for the marginal-standardized factor parameterization, and when it occurred, the entire replication was eliminated although the other three parameterizations still yielded converged estimates. Out of these 131 replications, only 63 replications from the marginal-standardized factor parameterization enter final calculation because the other 68 replications produced negative correlation estimates among θ 's, which again distorted the parameter estimates and their standard error estimates. The negative correlation again could be explained by possible reversing of some factors. For the other three parameterizations, all 131 replications were included in the final results.

An elimination criterion was also used for bi-factor model results. Again, out of 200 replications, only 175 converged properly, and non-convergence happened for the marginal and conditional reference indicator parameterizations. For these two parameterizations, the anomaly occurred when the factor mean estimates were extreme (i.e., >20 or <-20). Because the bi-factor model contains the larger number of parameters compared to the within-item or between-item MIRT models, it is unsurprising that fixing one item's parameters may not be enough to fix the scale sometimes, resulting in extreme factor mean estimates. Among these converged replications, only 113 yielded proper parameter and standard error estimates for the marginal and conditional reference indicator parameterizations. The other 62 replications again yielded

extreme factor mean estimates (i.e., absolute values of factor means exceeding 10). For both the within-factor MIRT and bi-factor models, similar trends still continue to hold. That is, all four parameterizations seemed to work equally well, except a few cells, such as the a_2 parameter in the marginal-standardized factor parameterization from the within-item MIRT model, and the a_0 parameter in the marginal-standardized factor parameterization from the bi-factor model. Again, even for these outstanding cells, the actual bias of SE is still acceptable.

In sum, for all four different models considered in the study, the SEs from successful replications are all comparable across the four different parameterizations. However, given the large proportion of unsuccessful replications observed in some cases, we have the following recommendations:

1. UIRT: Any of the four parameterizations is fine.
2. Between-item/Within-item MIRT: The marginal-standardized parameterization may result in factor reversing, which is easy to spot and correct. The marginal-reference and conditional-reference parameterizations sometimes yield invalid SEs of d -parameters after transformation, and future studies need to be conducted to further explore the reasons.
3. Bi-factor model: The marginal-reference and conditional-reference parameterizations may sometimes yield extreme factor means, which lead to either non-convergence or invalid transformed SEs.
4. For multidimensional models in general, the conditional-standardized parameterization is recommended.

A Real Data Example

For illustration purposes, a unidimensional factor analysis model employing each of the four

parameterizations was fit to the National Educational Longitudinal Study (NELS) science test data. The test consists of 25 dichotomous items, and a sample size of 13,487. We randomly sampled 1,000 students for illustration because otherwise, the standard errors of the parameter estimates will all be smaller than .0001, and *Mplus* default output of four decimal places could not capture the small values. The same data set was also input in *flexMIRT* (Cai, 2017) to obtain the IRT parameter estimates.

Table 10 presents the FA model parameter estimates for the NELS science test data, along with the robust SEs. Unsurprisingly, both parameter estimates and their SEs are quite different across the four parameterizations. Then Table 11 presents the transformed values of the item parameter estimates, and the computed SEs of the transformed item parameter estimates. Consistent with our expectation, after transformation, the item parameter estimates are all quite similar and they are also close to the direct IRT parameter estimates. The SEs of the transformed parameter values also look similar and the differences mostly appear in the third decimal place. The SEs from the direct IRT model fitting are not directly comparable because they are obtained from different estimators (sandwich estimator vs. supplemental EM), but the values are still close.

Discussion

In this note, we provide a general conversion formula for transforming the standard errors of item parameter estimates from four different parameterizations in factor analysis framework to the corresponding IRT parameterization. The conversion formula is suitable for a broad family of models, such as unidimensional, correlated-factor, and bi-factor models. This note is motivated by the observation that there is a lack of documentation on computing the standard errors of

transformed parameter values yet there is a need, for instance, to compare the standard error estimates across different studies for meta-analytic purposes.

While previous research that evaluated standard errors (Finch 2010, 2011) comparison between WLS and IRT estimations (such as EM) mainly used empirical standard errors by computing the standard deviation of parameter estimates across replications, the conversion formulas provided in this note offer to directly compute and compare SEs from different parameterizations. A simulation study is conducted to empirically evaluate the performance of the proposed conversion formula. Because the robust standard errors for WLS estimates has been shown to work well (Forero & Maydeu-Olivares, 2009) in the factor analytic framework, the SE of the transformed parameters also showed negligible bias. Even though we did not manipulate any factors (such as test length or sample size) in the simulation design, we considered four different IRT models which encompassed the majority of the applications. The R script crafted for this study is available to interested researchers who need to obtain the standard errors of item parameters from the IRT parameterization when the parameters are obtained in the FA models.

Please note that we did not claim that the transformed SEs from FA models using WLSMV are directly equivalent to those obtained from IRT models via FIML. This is because the sandwich estimator of SEs differs fundamentally from the SEM/XPD methods (limited-information vs. full-information approaches). Instead, we intend to claim that the transformed SEs from WLSMV are appropriate for the parameter estimates that are transformed to the IRT metric.

On a last note, although the SEs from the four different parameterizations in CFA framework can be transformed to the SEs of the item parameters from IRT metric, users need to exercise caution when using the SEs for Wald type of hypothesis testing for DIF and/or item drift

analysis. According to Gonzalez and Griffin (2001), in CFA, “*alternative but equivalent ways to identify a model may yield different standard errors, and hence different Z test for a parameter*”.

This lack of identification invariance of SEs implies that a parameter’s SE, and hence its significance test, can be sensitive to arbitrary choice of identification (i.e., reference indicator vs. standardized factor). Although their conclusion was for the continuous factor analysis, it is likely that the same conclusion also generalizes to categorical CFA due to the nonlinear transformation reflected in Table 2. Future research should look into this issue more closely. It is especially important to check how the Type I error and power of the Wald-based DIF/item drift analysis methods may be affected by the different identification parameterizations of IRT models.

References

- Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.
doi:10.1111/j.1745-3984.1988.tb00308.x
- Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309-329.
doi:10.1348/000711007X249603
- Cai, L. (2017). flexMIRT[®] version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.
- Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2p tables. *British Journal of Mathematical and Statistical Psychology, 59*, 173-194.
doi:10.1348/000711005X66419
- Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221. doi:10.1037/a0023350
- Casella, G., & Berger, R. L. (2002). *Statistical inference*. Australia: Thomson Learning.
- Cheng, Y., & Yuan, K. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika, 75*, 280-291. doi:10.1007/s11336-009-9144-x
- DeMars, C. E. (2012). A comparison of limited-Information and full-information methods in Mplus for estimating item response theory parameters for nonnormal populations. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(4), 610-632.
doi:10.1080/10705511.2012.713272
- Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation

- modeling. *Structural equation modeling: A second course*, 10(6), 269-314.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466-491. doi:10.1037/1082-989X.9.4.466
- Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor-analysis based models. *Applied Psychological Measurement*, 34, 10-26. doi:10.1177/0146621609336112
- Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement*, 35, 67-82.
doi:10.1177/0146621610367787
- Forero, C., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological Methods*, 14, 275-299.
doi:10.1037/a0015825
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57, 423-436. doi:10.1007/BF02295430
- Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every “One” matters. *Psychological Methods*, 6, 258-269. doi:10.1037/1082-989X.6.3.258
- Holzinger K., & Swineford F. (1937). The bi-factor method. *Psychometrika*, 2, 41-54.
doi:10.1007/BF02287965
- Kamata, A., & Bauer, D.J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modelling*, 15(1), 136 - 153.
doi:10.1080/10705510701758406
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA:

Addison-Wesley.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39(3), 479-515.

doi:10.1207/S15327906MBR3903_4

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, 30(1), 55-78. doi:10.1111/j.1745-3984.1993.tb00422.x

Muthén, B.; du Toit, SHC.; Spisic, D. Unpublished technical report. Los Angeles, CA: Muthén & Muthén; 1997. Robust inference using weighted least squared and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Paek, I., & Cai, L. (2014). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling. *Educational and Psychological Measurement*, 74, 58-76. doi:10.1177/0013164413500277

Rao, C.R. (1973). *Linear Statistical Inference and its Applications*. New York: Wiley

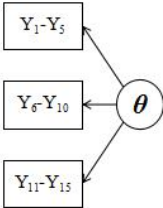
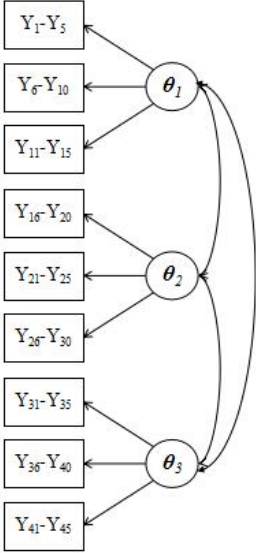
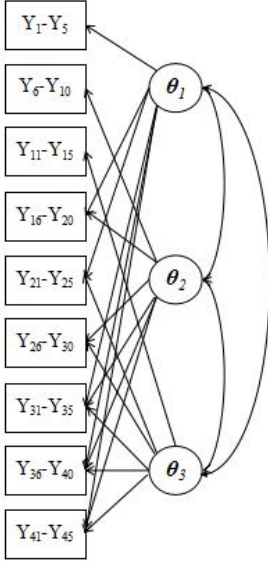
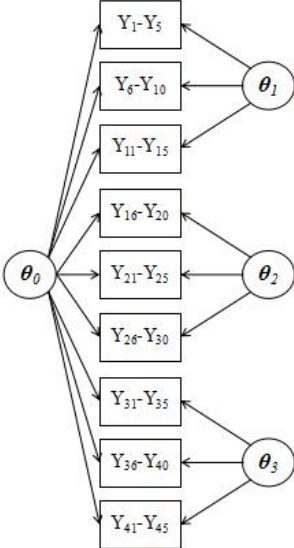
Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional Item Response Theory* (pp. 79-112). Springer, New York, NY.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667-696. doi:10.1080/00273171.2012.715555

Satorra, A., & Bentler, P. (1990). Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics & Data Analysis*, 10, 235-249.

- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408. doi:10.1007/BF02294363
- Team, R. C. (2016). R: A language and environment for statistical computing [Computer software]. Vienna: R Foundation for Statistical Computing.
- Thissen, D., & Wainer, H. (1990). Confidence envelopes for item response theory. *Journal of Educational Statistics*, *15*, 113-128. doi:10.3102/10769986015002113
- Tian, W., Cai, L., Thissen, D., & Xin, T. (2012). Numerical differentiation methods for computing error covariance matrices in item response theory modeling: an evaluation and a new proposal. *Educational and Psychological Measurement*, *73*, 412-439. doi:10.1177/0013164412465875
- Wang, C., Kohli, N., & Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 455-465. doi:10.1080/10705511.2015.1096744
- Wang, C., Su, S., & Weiss, D. J. (2018). Robustness of Parameter Estimation to Assumptions of Normality in the Multidimensional Graded Response Model. *Multivariate behavioral research*, *53*(3), 403-418. doi:10.1080/00273171.2018.1455572
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological methods*, *12*(1), 58. doi:10.1037/1082-989X.12.1.58
- Woods, C. M., Cai, L., Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*, 532-547. doi:10.1177/0013164412464875

Table 4. Summary of Generation Models.

	UIRT	Between-item MIRT	Within-item MIRT	Bi-factor
Item Parameter	$a_i \sim \log N(\mu_a=0, \sigma_a=0.2)$ $d_i \sim U(-2, 2)$	$a_{id} \sim \log N(\mu_a=0, \sigma_a=0.2)$ $d_i \sim U(-2, 2)$	$a_{id} \sim \log N(\mu_a=0, \sigma_a=0.2)$ $d_i \sim U(-2, 2)$	$a_{iGEN} \sim \log N(\mu_a=0, \sigma_a=0.2)$ $a_{iGR} \sim \log N(\mu_a=0, \sigma_a=0.2)$ $d_i \sim U(-2, 2)$
Person Parameter	$\theta_j \sim N(\mu_\theta=0, \sigma_\theta=1)$	$[\theta_i] \sim MVN \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}$	$[\theta_i] \sim MVN \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}$	$\theta_{jGEN} \sim N(\mu_\theta=0, \sigma_\theta=1)$ $\theta_{jGR} \sim N(\mu_\theta=0, \sigma_\theta=1)$
Test Structure				

Note. UIRT = Unidimensional IRT model. For the test structure of bi-factor models, θ_0 is the vector of general factor, θ_1 , θ_2 , and θ_3 are the vectors of group factor

Table 6. Parameter and standard error recovery of Unidimensional IRT Model.

		Marginal						Conditional					
		Standardized			Reference			Standardized			Reference		
		RMSE	bias	RB	RMSE	bias	RB	RMSE	bias	RB	RMSE	bias	RB
<i>a</i>	Par.	.1221	.0056	.0065	.1217	.0048	.0060	.1221	.0056	.0065	.1217	.0048	.0060
	S.E.	.0101	-.0034	-.0293	.0101	-.0036	-.0305	.0101	-.0035	-.0292	.0101	-.0036	-.0305
<i>d</i>	Par.	.0927	-.0061	.0216	.0925	-.0053	.0219	.0927	-.0061	.0216	.0925	-.0052	.0218
	S.E.	.0088	.0010	.0079	.0074	-.0023	-.0252	.0072	-.0022	-.0240	.0079	.0008	.0102
<i>b</i>	Par.	.1634	.0057	.0258	.1669	.0055	.0269	.1634	.0058	.0258	.1668	.0054	.0268
	S.E.	.0331	-.0029	-.0093	.0341	-.0030	-.0089	.0331	-.0029	-.0094	.0341	-.0029	-.0085

Note. For each parameter, the first row is the parameter recovery and the second row refers to SE recovery.

Table 7. Parameter and standard error recovery of between-item MIRT Model.

		Marginal						Conditional					
		Standardized			Reference			Standardized			Reference		
		RMSE	bias	RB	RMSE	bias	RB	RMSE	bias	RB	RMSE	bias	RB
<i>a1</i>	Par.	.1214	.0149	.0160	.1202	.0172	.0185	.1200	.0151	.0162	.1202	.0172	.0186
	S.E.	.0156	.0014	.0201	.0465	.0422	.3759	.0113	.0005	.0084	.0142	-.0002	.0073
<i>a2</i>	Par.	.1191	.0131	.0171	.1193	.0104	.0137	.1191	.0125	.0162	.1192	.0103	.0137
	S.E.	.0125	-.0005	-.0015	.0337	.0295	.2580	.0109	-.0013	-.0064	.0127	-.0017	-.0074
<i>a3</i>	Par.	.1234	.0108	.0107	.1213	.0151	.0146	.1237	.0109	.0109	.1213	.0151	.0146
	S.E.	.0153	-.0014	-.0066	.0289	.0233	.2003	.0110	-.0016	-.0112	.0116	-.0023	-.0154
<i>d</i>	Par.	.0899	-.0046	.0379	.0887	-.0030	.0369	.0892	-.0036	.0353	.0887	-.0030	.0367
	S.E.	.0087	.0046	.0605	.0232	.0077	.0907	.0080	-.0003	-.0018	.0187	-.0000	.0029
<i>b</i>	Par.	.1513	.0050	.0357	.1483	.0072	.0342	.1511	.0038	.0339	.1483	.0071	.0341
	S.E.	.0331	.0004	.0292	.0377	-.0034	-.0282	.0319	-.0021	-.0125	.0371	-.0011	.0009

Note. There are 17% abnormal replications from the marginal-standardization parameterization (i.e., negative loading parameters resulting in “NaN” for the transformed standard errors), 18% replications from the marginal-reference parameterization, and 30.5% from the conditional-reference parameterization. These replications were excluded from the summary.

Table 8. Parameter and standard error recovery of Within-item IRT Model.

		Marginal						Conditional					
		Standardized			Reference			Standardized			Reference		
		RMSE	bias	RB	RMSE	bias	RB	RMSE	bias	RB	RMSE	bias	RB
<i>a</i> ₁	Par.	.1763	.0187	.0202	.1746	.0000	-.0004	.1745	-.0002	-.0006	.1746	-.0001	-.0005
	S.E.	.0219	-.0093	-.0471	.0186	-.0064	-.0358	.0281	.0116	.0585	.0189	-.0085	-.0470
<i>a</i> ₂	Par.	.3066	-.0573	-.0556	.1804	-.0102	-.0095	.1803	-.0104	-.0097	.1803	-.0104	-.0097
	S.E.	.1396	-.1317	-.4373	.0209	-.0090	-.0465	.0298	.0118	.0627	.0217	-.0115	-.0592
<i>a</i> ₃	Par.	.1737	-.0228	-.0231	.1729	-.0053	-.0051	.1729	-.0052	-.0051	.1729	-.0052	-.0051
	S.E.	.0215	-.0060	-.0258	.0190	-.0042	-.0199	.0329	.0165	.0950	.0194	-.0063	-.0313
<i>d</i>	Par.	.1142	.0088	.0026	.1154	.0024	.0064	.1137	.0011	.0071	.1153	.0025	.0062
	S.E.	.0129	.0005	.0108	.0145	.0035	.0299	.0107	.0002	.0059	.0110	.0002	.0063

Note. 131/200 valid replications. Among the valid replications, 34% replications from the marginal-standardization parameterization failed to produce valid transformed SEs due to the negative estimated loadings.

Table 9. Parameter and standard error recovery of bi-factor IRT Model.

		Marginal						Conditional					
		Standardized			Reference			Standardized			Reference		
		RMSE	bias	RB	RMSE	bias	RB	RMSE	bias	RB	RMSE	bias	RB
a_0	Par.	.1322	.0015	.0056	.1317	.0054	.0066	.1322	.0045	.0056	.1316	.0054	.0066
	S.E.	.0681	.0574	.4362	.0334	.0101	.0784	.0117	-.0009	-.0041	.0115	-.0007	-.0032
a_1	Par.	.1539	-.0017	-.0021	.1519	-.0019	-.0023	.1539	-.1730	-.0021	.1519	-.0018	-.0022
	S.E.	.0179	-.0057	-.0304	.0237	.0000	.0098	.0179	-.0056	-.0303	.0189	-.0077	-.0434
a_2	Par.	.1418	.0062	.0074	.1430	.0057	.0066	.1418	.0062	.0073	.1431	.0057	.0066
	S.E.	.0136	-.0033	-.0203	.0461	.0197	.1343	.0136	-.0033	-.0202	.0139	-.0017	-.0068
a_3	Par.	.1444	.0039	.0049	.1456	.0031	.0039	.1444	.0039	.0049	.1456	.0033	.0042
	S.E.	.0126	-.0016	-.0090	.0240	.0063	.0453	.0126	-.0016	-.0091	.0152	-.0018	-.0071
d	Par.	.1037	.0000	.0058	.1037	.0003	.0066	.1037	.0001	.0056	.1036	.0004	.0070
	S.E.	.0095	-.0006	-.0017	.0128	.0000	.0052	.0095	-.0006	-.0017	.0118	.0004	.0078

Note. 175/200 valid replications. Among them, 43.5% replications from both the marginal-reference and conditional-reference parameterizations failed to produce valid transformed SEs due to the extreme factor mean estimates (i.e., absolute values exceeding 10)

Table 10. Factor analysis model parameter estimates for NELS Science data.

Item No.	Marginal				Conditional			
	Standardized Factor		Reference Indicator		Standardized Factor		Reference Indicator	
	λ	τ	λ	τ	λ	τ	λ	τ
1	0.516 (0.038)	-0.516 (0.042)	1	0	0.602 (0.061)	-0.602 (0.05)	1	0
2	0.47 (0.042)	-0.827 (0.045)	0.912 (0.104)	-0.357 (0.08)	0.533 (0.06)	-0.938 (0.054)	0.885 (0.132)	-0.405 (0.085)
3	0.373 (0.042)	-0.479 (0.041)	0.723 (0.099)	-0.106 (0.075)	0.402 (0.053)	-0.516 (0.045)	0.668 (0.113)	-0.114 (0.079)
4	0.399 (0.041)	-0.53 (0.042)	0.773 (0.096)	-0.132 (0.07)	0.435 (0.054)	-0.578 (0.046)	0.722 (0.113)	-0.144 (0.075)
5	0.658 (0.034)	-0.824 (0.045)	1.275 (0.114)	-0.166 (0.092)	0.874 (0.08)	-1.094 (0.068)	1.451 (0.195)	-0.221 (0.118)
6	0.587 (0.037)	-0.852 (0.045)	1.138 (0.108)	-0.265 (0.085)	0.726 (0.071)	-1.053 (0.062)	1.204 (0.163)	-0.328 (0.1)
7	0.394 (0.041)	-0.49 (0.041)	0.763 (0.095)	-0.097 (0.071)	0.428 (0.053)	-0.533 (0.046)	0.711 (0.11)	-0.105 (0.076)
8	0.423 (0.04)	-0.24 (0.04)	0.82 (0.095)	0.183 (0.07)	0.467 (0.053)	-0.265 (0.044)	0.775 (0.114)	0.202 (0.079)
9	0.498 (0.039)	-0.479 (0.041)	0.966 (0.105)	0.019 (0.081)	0.575 (0.06)	-0.552 (0.049)	0.954 (0.14)	0.022 (0.094)
10	0.544 (0.037)	-0.253 (0.04)	1.055 (0.107)	0.291 (0.081)	0.649 (0.063)	-0.302 (0.048)	1.077 (0.152)	0.347 (0.102)
11	0.294 (0.043)	-0.048 (0.04)	0.569 (0.091)	0.246 (0.065)	0.307 (0.049)	-0.05 (0.041)	0.51 (0.094)	0.257 (0.07)
12	0.572 (0.037)	-0.687 (0.043)	1.109 (0.107)	-0.115 (0.085)	0.698 (0.067)	-0.838 (0.056)	1.159 (0.158)	-0.14 (0.101)
13	0.459 (0.043)	-0.732 (0.044)	0.889 (0.105)	-0.274 (0.078)	0.516 (0.062)	-0.824 (0.052)	0.857 (0.132)	-0.308 (0.083)
14	0.679 (0.033)	-0.192 (0.04)	1.315 (0.114)	0.487 (0.09)	0.924 (0.083)	-0.261 (0.055)	1.534 (0.203)	0.663 (0.135)
15	0.438 (0.041)	0.202 (0.04)	0.848 (0.101)	0.639 (0.074)	0.487 (0.056)	0.225 (0.045)	0.808 (0.124)	0.711 (0.092)
16	0.325 (0.042)	-0.07 (0.04)	0.629 (0.096)	0.254 (0.071)	0.343 (0.05)	-0.074 (0.042)	0.569 (0.103)	0.269 (0.078)
17	0.512 (0.039)	0.123 (0.04)	0.992 (0.104)	0.635 (0.077)	0.596 (0.062)	0.143 (0.046)	0.989 (0.142)	0.739 (0.101)
18	0.496 (0.039)	-0.058 (0.04)	0.961 (0.102)	0.438 (0.078)	0.571 (0.059)	-0.066 (0.046)	0.948 (0.136)	0.504 (0.097)
19	0.454 (0.04)	0.136 (0.04)	0.88 (0.101)	0.589 (0.075)	0.509 (0.057)	0.152 (0.045)	0.846 (0.127)	0.662 (0.094)
20	0.282 (0.043)	0.148 (0.04)	0.546 (0.091)	0.43 (0.066)	0.294 (0.049)	0.155 (0.042)	0.488 (0.092)	0.449 (0.072)
21	0.34 (0.042)	0.093 (0.04)	0.658 (0.096)	0.432 (0.071)	0.361 (0.051)	0.099 (0.042)	0.599 (0.104)	0.46 (0.08)
22	0.219 (0.045)	0.285 (0.04)	0.424 (0.092)	0.503 (0.064)	0.224 (0.049)	0.292 (0.041)	0.372 (0.087)	0.515 (0.069)
23	0.214 (0.044)	0.176 (0.04)	0.415 (0.09)	0.39 (0.062)	0.219 (0.048)	0.181 (0.041)	0.364 (0.085)	0.4 (0.067)
24	0.527	0.356	1.022	0.883	0.621	0.419	1.03	1.039

	(0.04)	(0.041)	(0.105)	(0.077)	(0.065)	(0.049)	(0.146)	(0.108)
25	0.297	0.729	0.575	1.025	0.31	0.763	0.515	1.074
	(0.051)	(0.044)	(0.108)	(0.074)	(0.058)	(0.047)	(0.111)	(0.091)

Note. Values in parentheses are standard errors. Under both reference indicator parameterizations, the first item was chosen as the anchor item, of which the intercept and slope were fixed as 0 and 1, respectively. The mean (μ_{ξ}) and variance (σ_{ξ}^2) of factors under both standardized factor parameterizations were fixed as 0 and 1, respectively. Under the marginal-reference parameterization, the estimated value and standard error of μ_{ξ} were 0.516 and 0.042, and the estimated value and standard error of σ_{ξ}^2 were 0.266 and 0.04, respectively. And under the conditional-reference parameterization, the estimated value and standard error of μ_{ξ} were 0.602 and 0.05, and the estimated value and standard error of σ_{ξ}^2 were 0.363 and 0.073, respectively.

Table 11. Transformed IRT parameter estimates and direct estimates of IRT parameters and their corresponding standard errors for NELS Science data.

Item No.	Marginal				Conditional				Direct IRT	
	Standardized Factor		Reference Indicator		Standardized Factor		Reference Indicator		<i>a</i>	<i>d</i>
	<i>a</i>	<i>d</i>	<i>a</i>	<i>d</i>	<i>a</i>	<i>d</i>	<i>a</i>	<i>d</i>		
1	1.024	1.024	1.023	1.024	1.023	1.023	1.024	1.023	1.04	1.02
	(0.086)	(0.092)			(0.108)	(0.076)			(0.12)	(0.09)
2	0.905	1.593	0.906	1.594	0.906	1.595	0.906	1.594	0.98	1.61
	(0.111)	(0.096)	(0.104)	(0.097)	(0.108)	(0.093)	(0.106)	(0.088)	(0.13)	(0.11)
3	0.683	0.878	0.683	0.878	0.683	0.877	0.684	0.877	0.68	0.85
	(0.095)	(0.084)	(0.100)	(0.087)	(0.093)	(0.076)	(0.092)	(0.054)	(0.10)	(0.08)
4	0.740	0.983	0.739	0.984	0.740	0.983	0.740	0.984	0.75	0.96
	(0.099)	(0.085)	(0.098)	(0.080)	(0.093)	(0.076)	(0.092)	(0.080)	(0.10)	(0.08)
5	1.485	1.860	1.484	1.859	1.486	1.860	1.486	1.861	1.70	2.00
	(0.126)	(0.122)	(0.149)	(0.071)	(0.132)	(0.120)	(0.12)	(0.129)	(0.19)	(0.16)
6	1.233	1.789	1.232	1.789	1.234	1.790	1.233	1.790	1.37	1.87
	(0.101)	(0.107)	(0.151)	(0.149)	(0.120)	(0.108)	(0.122)	(0.065)	(0.16)	(0.14)
7	0.729	0.906	0.728	0.907	0.728	0.906	0.728	0.906	0.74	0.89
	(0.098)	(0.085)	(0.097)	(0.096)	(0.093)	(0.076)	(0.086)	(0.079)	(0.10)	(0.08)
8	0.794	0.450	0.793	0.450	0.794	0.451	0.794	0.450	0.78	0.44
	(0.102)	(0.085)	(0.103)	(0.081)	(0.093)	(0.076)	(0.092)	(0.054)	(0.10)	(0.08)
9	0.976	0.939	0.977	0.940	0.978	0.938	0.977	0.939	0.98	0.93
	(0.117)	(0.092)	(0.110)	(0.088)	(0.108)	(0.076)	(0.104)	(0.092)	(0.11)	(0.09)
10	1.102	0.513	1.102	0.513	1.103	0.513	1.103	0.512	1.09	0.52
	(0.091)	(0.091)	(0.124)	(0.088)	(0.108)	(0.076)	(0.104)	(0.046)	(0.11)	(0.08)
11	0.523	0.085	0.522	0.085	0.522	0.085	0.522	0.085	0.49	0.08
	(0.087)	(0.080)	(0.078)	(0.082)	(0.076)	(0.076)	(0.078)	(0.057)	(0.08)	(0.07)
12	1.185	1.424	1.185	1.424	1.187	1.425	1.187	1.424	1.29	1.47
	(0.097)	(0.100)	(0.135)	(0.065)	(0.108)	(0.093)	(0.113)	(0.085)	(0.14)	(0.11)
13	0.878	1.401	0.877	1.402	0.877	1.401	0.878	1.401	0.91	1.40
	(0.108)	(0.093)	(0.101)	(0.100)	(0.108)	(0.093)	(0.101)	(0.085)	(0.11)	(0.10)
14	1.572	0.445	1.569	0.443	1.571	0.444	1.571	0.443	1.56	0.47
	(0.136)	(0.105)	(0.163)	(0.074)	(0.142)	(0.093)	(0.133)	(0.082)	(0.14)	(0.10)
15	0.828	-0.382	0.827	-0.381	0.828	-0.383	0.828	-0.382	0.76	-0.36
	(0.105)	(0.085)	(0.090)	(0.099)	(0.093)	(0.076)	(0.091)	(0.083)	(0.09)	(0.08)
16	0.584	0.126	0.583	0.127	0.583	0.126	0.583	0.125	0.57	0.12
	(0.090)	(0.080)	(0.087)	(0.080)	(0.076)	(0.076)	(0.081)	(0.084)	(0.09)	(0.07)
17	1.013	-0.243	1.012	-0.244	1.013	-0.243	1.013	-0.244	0.97	-0.23

	(0.120)	(0.089)	(0.114)	(0.063)	(0.108)	(0.076)	(0.103)	(0.090)	(0.10)	(0.08)
18	0.971	0.114	0.970	0.113	0.971	0.112	0.971	0.113	0.94	0.12
	(0.116)	(0.088)	(0.101)	(0.043)	(0.108)	(0.076)	(0.093)	(0.081)	(0.10)	(0.08)
19	0.866	-0.259	0.866	-0.257	0.865	-0.258	0.867	-0.260	0.83	-0.24
	(0.107)	(0.086)	(0.092)	(0.097)	(0.093)	(0.076)	(0.096)	(0.089)	(0.09)	(0.08)
20	0.500	-0.262	0.499	-0.263	0.500	-0.264	0.500	-0.264	0.50	-0.25
	(0.086)	(0.079)	(0.089)	(0.074)	(0.076)	(0.076)	(0.087)	(0.076)	(0.08)	(0.07)
21	0.615	-0.168	0.613	-0.167	0.614	-0.168	0.614	-0.169	0.59	-0.16
	(0.091)	(0.081)	(0.089)	(0.076)	(0.093)	(0.076)	(0.090)	(0.055)	(0.09)	(0.07)
22	0.382	-0.497	0.381	-0.495	0.381	-0.496	0.381	-0.495	0.37	-0.47
	(0.082)	(0.078)	(0.083)	(0.067)	(0.076)	(0.076)	(0.087)	(0.064)	(0.08)	(0.07)
23	0.372	-0.306	0.372	-0.306	0.372	-0.308	0.373	-0.307	0.35	-0.29
	(0.082)	(0.078)	(0.082)	(0.067)	(0.076)	(0.076)	(0.081)	(0.076)	(0.08)	(0.07)
24	1.054	-0.712	1.054	-0.711	1.056	-0.712	1.055	-0.712	0.99	-0.68
	(0.124)	(0.092)	(0.118)	(0.076)	(0.108)	(0.076)	(0.108)	(0.094)	(0.10)	(0.08)
25	0.529	-1.298	0.528	-1.296	0.527	-1.297	0.527	-1.299	0.53	-1.26
	(0.107)	(0.083)	(0.101)	(0.106)	(0.093)	(0.076)	(0.096)	(0.067)	(0.09)	(0.08)

Note. Values in parentheses are standard errors. The direct IRT model fitting was conducted using *flexMIRT*.

Appendix A: Γ_{a_i} and Γ_{b_i} in Table 3 for different parameterizations

A1. Marginal-Reference Indicator

The analytic forms for the first derivatives are presented below, the non-presented terms, such as

$vec\left(\frac{\partial a_{i1}}{\partial(E(\xi))}\right)$, is a 0-vector.

$$\frac{\partial a_{ik}}{\partial \lambda_{is}} = \begin{cases} \frac{[1 - \lambda_i^T \text{cov}(\xi) \lambda_i + N_{ik} \lambda_{ik}] V(\xi_k)^{1/2}}{(1 - \lambda_i^T \text{cov}(\xi) \lambda_i)^{3/2}} & s = k \\ \frac{\lambda_{ik} V(\xi_k)^{1/2} N_{is}}{(1 - \lambda_i^T \text{cov}(\xi) \lambda_i)^{3/2}} & s \neq k \end{cases}$$

$$\frac{\partial a_{ik}}{\partial(\text{cov}(\xi))_{st}} = \begin{cases} \frac{\lambda_{ik} V(\xi_k)^{1/2} \lambda_{is} \lambda_{it}}{2 \times (1 - \lambda_i^T \text{cov}(\xi) \lambda_i)^{3/2}} & s = t \neq k \text{ or } s \neq t \\ \frac{\lambda_{ik} V(\xi_k)^{-1/2} (1 - \lambda_i^T \text{cov}(\xi) \lambda_i) + \lambda_{ik}^3 V(\xi_k)^{1/2}}{2 \times (1 - \lambda_i^T \text{cov}(\xi) \lambda_i)^{3/2}} & s = t = k \end{cases}$$

$$\begin{aligned} \frac{\partial d_i}{\partial \lambda_i} &= \frac{E(\xi) \sqrt{1 - \lambda_i^T \text{cov}(\xi) \lambda_i} - [1 - \lambda_i^T \text{cov}(\xi) \lambda_i]^{-1/2} N_i(\tau_i - \lambda_i^T E(\xi))}{1 - \lambda_i^T \text{cov}(\xi) \lambda_i} \\ &= \frac{E(\xi) [1 - \lambda_i^T \text{cov}(\xi) \lambda_i] - N_i(\tau_i - \lambda_i^T E(\xi))}{[1 - \lambda_i^T \text{cov}(\xi) \lambda_i]^{3/2}} \end{aligned}$$

$$\frac{\partial d_i}{\partial \tau_i} = -[1 - \lambda_i^T \text{cov}(\xi) \lambda_i]^{-1/2}$$

$$\frac{\partial d_i}{\partial(\text{cov}(\xi))_{st}} = \frac{-[\tau_i - \lambda_i^T E(\xi)] \lambda_{is} \lambda_{it}}{2 \times (1 - \lambda_i^T \text{cov}(\xi) \lambda_i)^{3/2}}$$

$$\frac{\partial d_i}{\partial(E(\xi))_s} = \frac{\lambda_{is}}{\sqrt{1 - \lambda_i^T \text{cov}(\xi) \lambda_i}}$$

A2. Conditional-Reference Indicator

The analytic forms for the first derivatives are presented as follows. The non-presented terms, such

as $\text{vec}\left(\frac{\partial a_{i1}}{\partial(E(\boldsymbol{\xi}))}\right)$, $\text{vec}\left(\frac{\partial d_i}{\partial(E(\boldsymbol{\xi}))_s}\right)$, and $\text{vec}\left(\frac{\partial d_i}{\partial(\text{cov}(\boldsymbol{\xi}))}\right)$ are a 0-vector.

$$\frac{\partial a_{ik}}{\partial \lambda_{ik}} = V(\xi_k)^{1/2}$$

$$\frac{\partial a_{ik}}{\partial(\text{cov}(\boldsymbol{\xi}))_{kk}} = \frac{1}{2} \lambda_{ik} V(\xi_k)^{-1/2}$$

$$\frac{\partial d_i}{\partial \lambda_i} = E(\xi)$$

$$\frac{\partial d_i}{\partial \tau_i} = -1$$

A3. Marginal-Standardized Factor

The analytic forms for the first derivatives are presented as follows. In all the following equations,

$\text{cov}(\boldsymbol{\xi}) \equiv \text{cor}(\boldsymbol{\xi})$ because the factors are standardized.

$$\frac{\partial a_{ik}}{\partial \lambda_{is}} = \begin{cases} \frac{[1 - \lambda_i^T \text{cov}(\boldsymbol{\xi}) \lambda_i + N_{ik} \lambda_{ik}]}{(1 - \lambda_i^T \text{cov}(\boldsymbol{\xi}) \lambda_i)^{3/2}} & s = k \\ \frac{\lambda_{ik} N_{is}}{(1 - \lambda_i^T \text{cov}(\boldsymbol{\xi}) \lambda_i)^{3/2}} & s \neq k \end{cases}$$

$$\frac{\partial a_{ik}}{\partial(\text{cov}(\boldsymbol{\xi}))_{st}} = \frac{\lambda_{ik} \lambda_{is} \lambda_{it}}{2 \times (1 - \lambda_i^T \text{cov}(\boldsymbol{\xi}) \lambda_i)^{3/2}}$$

$$\frac{\partial d_i}{\partial \lambda_i} = \frac{-\tau_i [1 - \lambda_i^T \text{cov}(\boldsymbol{\xi}) \lambda_i]^{-1/2} N_i}{1 - \lambda_i^T \text{cov}(\boldsymbol{\xi}) \lambda_i} = \frac{-\tau_i N_i}{[1 - \lambda_i^T \text{cov}(\boldsymbol{\xi}) \lambda_i]^{3/2}}$$

$$\frac{\partial d_i}{\partial \tau_i} = -[1 - \lambda_i^T \text{cov}(\boldsymbol{\xi}) \lambda_i]^{-1/2}$$

$$\frac{\partial d_i}{\partial(\text{cov}(\boldsymbol{\xi}))_{st}} = \frac{-\tau_i \lambda_{is} \lambda_{it}}{2 \times (1 - \lambda_i^T \text{cov}(\boldsymbol{\xi}) \lambda_i)^{3/2}}$$

A4. The analytic forms for Γ_{b_i} in Table 3

Reference Indicator

$$\frac{\partial b_i}{\partial \lambda_i} = \frac{-E(\xi_k)\lambda_{ik} - [\tau_i - \lambda_i^T E(\xi)]}{\lambda_{ik}^2 V(\xi_k)^{1/2}}, \text{ and this equation simplifies to } \frac{\partial b_i}{\partial \lambda_i} = \frac{-\tau_i}{\lambda_i^2 V(\xi)^{1/2}} \text{ for 2PL.}$$

$$\frac{\partial b_i}{\partial \tau_i} = \frac{1}{\lambda_{ik} V(\xi_k)^{1/2}},$$

$$\frac{\partial b_i}{\partial (\text{cov}(\xi))_{kk}} = -\frac{1}{2V(\xi_k)^{3/2}} \times \left(\frac{\tau_i - \lambda_i^T E(\xi)}{\lambda_{ik}} \right), \text{ and the first derivative of } b_i \text{ with respect to other}$$

elements in $\text{cov}(\xi)$ are all 0's.

$$\frac{\partial b_i}{\partial (E(\xi))_s} = \frac{-\lambda_{is}}{\lambda_{ik} V(\xi_k)^{1/2}}.$$

Standardized Factor

$$\frac{\partial b_i}{\partial \tau_i} = \frac{1}{\lambda_{ik}}, \text{ and } \frac{\partial b_i}{\partial \lambda_{ik}} = -\frac{\tau_i}{\lambda_{ik}^2}. \text{ The first derivative of } b_i \text{ with respect to all other parameters are 0.}$$

Appendix B: Mplus syntax for bi-factor models

B1. Marginal standardized factor parameterization

```
TITLE: binary bi-factor analysis model BY MARGINAL STANDARDIZED FACTOR
DATA: FILE IS bif.dat;
VARIABLE: NAMES ARE u1-u45;
CATEGORICAL ARE u1-u45;
ANALYSIS: MODEL = NOCOVARIANCES;
TYPE = GENERAL;
ESTIMATOR = WLSMV;
PARAMETERIZATION = DELTA;
      ! DELTA = marginal parameterization; THETA = conditional parameterization.
MODEL: skill0 BY u1* u2-u45;
skill1 BY u1* u2-u15;
skill2 BY u16* u17-u30;
skill3 BY u31* u32-u45;
[skill1@0];
skill1@1;
[skill2@0];
skill2@1;
[skill3@0];
skill3@1;
[skill0@0];
skill0@1;
OUTPUT: TECH1, TECH3,TECH4;
```

B2. Marginal reference indicator parameterization

```
TITLE: binary bi-factor analysis model BY MARGINAL REFERENCE INDICATOR
DATA: FILE IS bif.dat;
VARIABLE: NAMES ARE u1-u45;
CATEGORICAL ARE u1-u45;
ANALYSIS: MODEL = NOCOVARIANCES;
TYPE = GENERAL;
ESTIMATOR = WLSMV;
PARAMETERIZATION = DELTA;
      ! DELTA = marginal parameterization; THETA = conditional parameterization.
MODEL: skill0 BY u1* u2@1 u3-u45;
skill1 BY u1 u2-u15;
skill2 BY u16 u17-u30;
skill3 BY u31 u32-u45;
[u1$1@0];
```

[u2\$1@0];
[u16\$1@0];
[u31\$1@0];
[skill0*];
[skill1*];
[skill2*];
[skill3*];
OUTPUT: TECH1, TECH3,TECH4;