Multidimensional Item Response Theory Parameter Estimation With Nonsimple Structure Items
Holmes Finch
Applied Psychological Measurement 2011 35: 67 originally published online 30 September 2010
DOI: 10.1177/0146621610367787

The online version of this article can be found at:
http://apm.sagepub.com/content/35/1/67

Published by:

SAGE
http://www.sagepublications.com

Additional services and information for Applied Psychological Measurement can be found at:

Email Alerts: http://apm.sagepub.com/cgi/alerts
Subscriptions: http://apm.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav
Citations: http://apm.sagepub.com/content/35/1/67.refs.html
Estimation of multidimensional item response theory (MIRT) model parameters can be carried out using the normal ogive with unweighted least squares estimation with the normal-ogive harmonic analysis robust method (NOHARM) software. Previous simulation research has demonstrated that this approach does yield accurate and efficient estimates of item discrimination and difficulty across a variety of conditions. However, these studies have been limited primarily to the case of simple structure, where each item is associated only with one of the latent traits underlying the data. The current simulation study seeks to extend this earlier work by comparing NOHARM and unidimensional IRT-based estimates of difficulty and discrimination values for items that do not conform to simple structure, that is, are associated with more than one latent trait. The outcomes of interest were relative bias and standard error values for parameter estimates under a variety of conditions, including the degree of nonsimple structure present. Results demonstrate that both bias and standard error tended to be larger for items that do not conform to simple structure than for those that do, but that the degree of such differences was influenced by factors such as correlation between the latent traits, sample size and distribution of the latent trait, among others.
items measure two reading constructs, such as passage comprehension and word meaning, a standard IRT model will not be appropriate for estimating item and person parameters, unless all of the items measure the same combination of the latent traits. In response to this problem, psychometricians developed the compensatory multidimensional IRT (MIRT) model for linking multiple latent traits with an item response (e.g., Reckase, 1985). The 3PL MIRT model can be expressed as

\[
P(U_i = 1 | \theta_j) = c_i + \frac{1 - c_i}{1 + e^{-1.7 \sum a_{ij}(\theta_j - b_j)}}
\]

where

- \(\theta_j\) = examinee ability for latent trait \(j\),
- \(c_i\) = pseudoguessing parameter for item \(i\),
- \(a_{ij}\) = discrimination parameter for item \(i\) on latent trait \(j\), and
- \(b_j\) = difficulty parameter for item \(i\).

This MIRT model takes the same general form as the unidimensional three-parameter logistic (3PL) model, reflecting the link between the latent ability and the probability of a correct item response through an item’s discrimination, difficulty, and pseudochance parameters. However, the MIRT model can be used to better understand the traits being measured in a multidimensional context by providing an estimate of each latent trait, as well as an item’s discrimination for each of these, and the overall difficulty of the item. Reckase (2007) noted that three areas in measurement where the MIRT model has received the most attention is the analysis of test content, equating, and computer adaptive testing.

**Confirmatory Factor Analysis for Item Responses**

Mathematically, there is a direct link between the MIRT model and factor analysis for item response data (e.g., Knol & Berger, 1991; McDonald, 1997; Reise, Widaman, & Pugh, 1993; Takane & de Leeuw, 1987). This correspondence between the models allows for a straightforward conversion of parameters from one to the other, thus making factor analysis a viable option for MIRT parameter estimation. The confirmatory factor analysis (CFA) model takes the form

\[
X = \Lambda \xi + \delta,
\]

where

- \(X\) = observed indicator variables (i.e., items),
- \(\Lambda\) = factor loadings,
- \(\xi\) = latent variable(s), and
- \(\delta\) = uniqueness for observed indicator variables.

This CFA model corresponds to the 2PL IRT model, in which pseudoguessing is not included. These CFA parameters can be easily transformed into item parameters (McDonald, 1999). Item discrimination \(a\) and difficulty \(b\) can be expressed as:

\[
a_i = \frac{\lambda_i}{\sqrt{1 - \lambda_i^T \varphi \lambda_i}},
\]

where

- \(\lambda_i\) = factor loading vector for item \(i\), and
- \(\varphi\) = covariance matrix of factors.
\[ b_i = \frac{-\tau_i}{\sqrt{1 - \lambda_i^2 \varphi_i}}, \]  

where \( \tau_i \) equals the threshold for item \( i \).

There are a number of methods available for estimating factor analysis models for item response data, including marginal maximum likelihood (Bock, Gibbons, & Muraki, 1988), weighted least squares (B. Muthén, du Toit, & Spisic, 1997), robust maximum likelihood (Satorra & Bentler, 1994), and unweighted least squares (ULS; McDonald, 1997), among others. Several of these have been investigated in previous studies in terms of estimation accuracy for item and person parameters and will be discussed in more detail below. However, it can be mentioned here that this prior work has found that the ULS method of estimation generally performs well in terms of MIRT parameter estimation accuracy and ability to reach convergence across a variety of sample sizes and test lengths.

This ULS estimation can be used in either an exploratory or confirmatory factor analysis framework with existing software, such as the normal ogive harmonic analysis robust method (NOHARM). In the confirmatory case, items are explicitly associated with the appropriate latent variable by the researcher, based presumably on prior knowledge regarding the latent structure of the instrument. It is this approach that will be used in the current study. The model underlying these factor models estimated with ULS, first introduced by McDonald (1967) for use with dichotomous data, is based on the normal ogive. It allows for the inclusion of multiple latent traits, each of which has its own discrimination parameter, and one difficulty parameter value per item. There is no pseudoguessing parameter estimated by this model, though if desired, an estimate for \( c \) can be obtained from standard unidimensional IRT software such as BILOGMG (Zimowski, Muraki, Mislevy, & Bock, 2003) and then provided to the NOHARM software as fixed values. When this is done, the resulting estimates of the threshold and factor loading parameters are adjusted for the pseudoguessing values.

Prior Research Investigating MIRT Model Estimation

The use of the CFA model to obtain MIRT parameter estimates has been investigated previously when data conform to a simple structure solution; that is, each item was associated with only one latent variable, even when multiple such traits were present. Several of these studies compared ULS estimation with marginal maximum likelihood estimation (MMLE) as carried out in the TESTFACT software (Bock et al., 2004). This earlier work found that ULS was able to reproduce MIRT item responses as well as or better than MMLE, particularly for small samples (Gosz & Walker, 2002; Knol & Berger, 1991; Miller, 1991). Tate (2003) expanded on these earlier studies by examining the parameter recovery of ULS, MMLE, and robust weighted least squares (RWLS) estimation using MPlus (L. K. Muthén & Muthén, 2006). He found that the RWLS method had difficulty in correctly estimating item difficulty or discrimination in the presence of pseudoguessing. Furthermore, Tate also demonstrated that ULS generally recovered item parameter values well in the CFA case, except when the underlying model involved a second-order factor on which four first-order factors loaded. Finch (2010) extended Tate’s work by including a broader array of simulation conditions, and examining more outcomes including actual parameter estimation bias and standard errors of the parameters produced by RWLS and ULS. This study reported similar results to Tate in regard to the negative impact of pseudoguessing on RWLS parameter estimates. However, little estimation bias was found for ULS when the pseudoguessing parameter was first estimated by BILOG and then provided to
NOHARM as fixed values. Both methods exhibited greater parameter estimation bias—particularly for difficulty—when the latent traits were not normally distributed, and a higher correlation between the latent traits was associated with greater estimation bias when the latent traits were skewed, for both methods. Across virtually all simulated conditions, the estimates obtained from ULS exhibited less bias and comparable standard errors to those from RWLS carried out in MPlus.

Taken together, the results of these studies suggest that ULS typically performed as well as, if not better than, other approaches (i.e., RWLS, TESTFACT) for modeling MIRT data. Therefore, it would seem that in the MIRT model situation where items exhibit simple structure, NOHARM is an excellent candidate for item parameter estimation.

Current Study
The goal of the current study was to assess the accuracy and precision of item difficulty and discrimination parameter estimation using ULS with NOHARM in the MIRT model context when some items do not exhibit simple structure. As discussed above, prior work has established that ULS is a promising method for MIRT parameter estimation in the simple structure condition where each item is associated with only one of the latent traits that are being measured. However, little or no work has been published assessing parameter estimation when some of the items are associated with multiple factors. This is an important issue, as in practice there is no reason to assume that items strictly conform to perfect simple structure, but rather may be associated with multiple latent traits to differing degrees. Thus, the primary purpose of this study was to ascertain the degree to which ULS was able to correctly recover these item parameter values for such nonsimple structure items under a variety of data conditions. A secondary goal was to determine whether the presence of nonsimple structure items affected item parameter estimation of pure simple structure items. Finally, the degree of nonsimple structure was manipulated in order to determine whether this factor affected parameter estimation. ULS as carried out by NOHARM was selected as the focus of this study because of the prior research demonstrating it to be as good as, and often better than, other methods for estimating MIRT modeling in terms of parameter estimation accuracy and precision, particularly in the presence of pseudoguessing. Following is a discussion of the simulation methodology used in this study, including a more detailed description of how the data were generated and which conditions were manipulated.

Method
The research questions regarding the accuracy and precision of item discrimination and difficulty parameter estimates in the nonsimple structure case were addressed using a Monte Carlo simulation study. All of the simulation conditions described below were completely crossed, with 500 replications per combination. The variables manipulated in this study are described below. A factorial analysis of variance (ANOVA) was used to identify main effects and interactions of the manipulated variables that contributed significantly to the estimation bias in the nonsimple structure items.

Method of Estimation
Three methods of parameter estimation were used in this study. First, a MIRT model was estimated with ULS using the NOHARM software described previously. A confirmatory factor model was estimated by using a target matrix that explicitly linked each item with the latent trait
with which it was known to be associated in the population. Using a target matrix, the nonsimple structure items were associated with both latent traits, whereas the simple structure items were associated only with one of the traits. Where appropriate, $c$ parameters were estimated using BILOG and given to NOHARM as fixed values for use in the estimation of item difficulty and discrimination. Because this was a confirmatory factor model, the issue of factor indeterminacy associated with exploratory factor analysis was not a concern here. With exploratory analyses, the factor solution from one sample to the next may not be aligned in the same fashion. For example, a researcher might conduct an exploratory factor analysis with 30 items for one sample and find that Items 1 through 15 load on Factor 1 whereas Items 16 through 30 load on Factor 2. If they were to collect data from another sample using the same items, a factor analysis might assign Items 1 through 15 to Factor 2 and Items 16 through 30 to Factor 1. Although the factor number is different in the two analyses, the actual results are comparable because the items group in the same fashion. This factor indeterminacy is not an issue for confirmatory factor analysis, such as that conducted here, because the items are associated with a specific factor through the use of a target loading matrix.

In addition to ULS, two different approaches to estimation based on BILOG were also used. The inclusion of these unidimensional approaches to the problem was based on what may happen in practice if a true MIRT model–based method such as ULS were not used. In the first BILOG approach, the potential multidimensionality in the data was ignored and all the items were fit assuming a single latent trait. This approach would be analogous to a researcher being unaware of the possibility that two dimensions underlie their data and thus fit a standard unidimensional model to the items. The second method based on BILOG involved dividing the items into two sets based on the latent trait that they belong to; that is, each item is grouped with the latent trait with which it is believed to be primarily associated. In the case of complex structure, two of the items were associated with the first latent trait and two with the second. After they are divided appropriately, item parameter estimates were obtained separately for the two groups of items. In this situation, the researcher would know that two latent traits underlie the data and thus would obtain estimates separately for items associated with the two latent traits.

**Item Parameters**

The data were generated from the compensatory MIRT model using item parameter values that were similar to those previously reported in Finch (in press), so as to replicate this earlier work which focused only on simple structure items. The means and standard deviations along with the minimum and maximum values of the distributions used for each parameter, which are described subsequently, were drawn from parameter values on a large scale reading examination given to students in a southern state. This test length is typical of such proficiency exams in this state. Specifically, the discrimination parameters were drawn from a normal distribution with a mean of 0.9657, a standard deviation of 0.3161, a minimum possible value of 0.70 and a maximum possible value of 2.00, whereas the difficulty parameter was generated from the standard normal distribution. The pseudoguessing values (where appropriate) were drawn from the uniform distribution with values ranging from 0.1178 and 0.3580. For every replication, 30 items were simulated from a two-dimensional latent structure, each associated with 15 items. For each of the latent traits, 13 of these items were generated based on simple structure; that is, they had a nonzero discrimination value for only one of the traits. In what will be referred to as the semicomplex structure condition, two additional items per trait (four items total) were generated so that they loaded primarily on one trait and had a smaller, nonzero, discrimination value on the other trait. This smaller discrimination parameter was generated from the normal
distribution with a mean of 0.35, or exactly half the size of the value on the primary latent trait, a standard deviation of 0.15, a minimum possible value of 0.1 and a maximum value of 0.6. In the second nonsimple structure condition, which is referred to as a complex structure, the discrimination parameters of the four nonsimple structure items (two per latent trait) were generated from the same distribution for both factors, with the mean of 0.9657, standard deviation of 0.3161, minimum value of 0.70, and maximum of 2.00. In summary, semicomplex structure was designed to reflect the case where an item was primarily, but not exclusively, associated with one latent trait, whereas the complex structure reflected the situation in which an item measured two latent traits with approximately equal strength.

**Number of Examinees**

The number of examinees was simulated to be 250, 500, 1,000, or 2,000. These values were selected to examine model performance from very small (for testing settings) to fairly large sample size conditions, and are similar to those used in a prior study examining the estimation accuracy in the simple structure condition (Finch, in press). In addition, a broad range of number of examinees has been used in other previous studies using ULS, from as low as 100 (Maydeu-Olivares, 2001; Parshall, Kromrey, Chason, & Yi, 1997) to as many as 2,000 (Batley & Boss, 1993; Tate, 2003).

**Intertrait Correlation**

The two latent traits were simulated to be correlated at 0, .3, .5, or .8. These values reflect the range of potential conditions from no or relatively low correlation (0 and .3) to a very large correlation (.8). Prior research has used a similarly wide range of values, including Batley and Boss (1993) with values of 0, .25, and .5; Miller (1991) with \( r = 0 \) or .5; Gosz and Walker (2002) with .5, .75, and .9; Tate (2003) with .6; Maydeu-Olivares (2001) with .5; and Finch, Habing, and Huynh (2003) with 0, .3, .5, .8, and .95; and Finch (in press) with 0, .3, .5, and .8. To replicate conditions from these earlier studies, the four values listed above were selected.

**Distribution of Latent Traits**

Prior simulation studies examining the impact of nonnormal data on IRT parameter estimation in the unidimensional case (Abdel-fattah, 1994; De Ayala & Sava-Bolesta, 1999; Stone, 1992) have consistently found that the distribution of the latent trait has an impact on accuracy of item parameter estimates. In addition, it has been shown that the distribution of latent traits affects the performance of item factor analysis (Finch, in press; Finch et al., 2003). For this reason, the distributions of the latent trait parameters were manipulated in this study. They were generated either from a standard normal distribution (referred to as normal below) or from a distribution with skewness of −1.5 and kurtosis of 3.0 (referred to as nonnormal). These values have been shown to affect the estimation accuracy of ULS in the simple structure condition (Finch, in press), and will thus be used here to ascertain their impact in the nonsimple structure case. To maintain the desired intertrait correlations previously discussed, the methodology outlined by Headrick and Sawilowsky (1999) was used to create the skewed distributions.

**Pseudoguessing**

The multidimensional data were simulated either with (M3PL) or without (M2PL) pseudoguessing present in order to determine the impact of the underlying model on estimation accuracy in...
the nonsimple structure condition. Previous research has demonstrated that although some meth-
ods of MIRT item parameter estimation are influenced by the presence of pseudoguessing in the
data, ULS as carried out by NOHARM is not, generally speaking, in the presence of simple struc-
ture (Tate, 2003). The reason for this robustness to model type is apparently because NOHARM
allows for the inclusion of previously obtained pseudoguessing estimates in model estimation.
For this study, BILOGMG (Zimowski, Muraki, Mislevy, & Bock, 2003) was used to obtain es-
timates of the pseudoguessing values that were in turn provided to the NOHARM software in the
M3PL case.

Type of Latent Structure

As already discussed, data were generated from two model types: (a) semicomplex structure, in
which two items per latent trait were primarily associated with that trait, but had a nonzero dis-
 crimination value with the other, and (b) complex structure, in which the four items had discrim-
ination parameter values that were strongly associated with both latent traits. In addition, the
other 26 items (13 per latent trait) were generated to reflect pure simple structure, being associ-
ated with only one of the traits. Although the case of only two latent traits with an equal number
of items per dimension is somewhat simplistic, this study represents one of the first examinations
of parameter estimation in the nonsimple structure case. Given the relative novelty of this study,
it was deemed important to maintain a simple design in order that results might be clear and
definitive. It is recognized, however, that future studies should focus on more complex (and
more realistic) designs.

Outcome Measures

Relative bias of the difficulty and discrimination parameter estimates was calculated as
\[
\sum \left( \frac{\hat{\theta} - \theta}{\theta} \right),
\]
where \( \hat{\theta} \) was the sample estimate for the parameter in question and \( \theta \) was the known
population value. Note that this is an absolute value of the difference. The empirical standard
error of the parameter estimates was calculated as
\[
\sqrt{\frac{\sum (\hat{\theta} - \theta)^2}{n-1}},
\]
where \( \bar{\theta} \) was the mean of the parameter estimates and \( n \) was the number of replications per combination of study conditions,
or 500 in this case.

For the calculation of estimation bias and standard error, the decision regarding the actual
parameter values against which to compare the parameter estimates is crucial. In the case of
the simple structure items, this decision was straightforward because each item was only asso-
ciated with one of the latent traits. However, each of the nonsimple structure items was associ-
ated with both latent traits to a greater or lesser extent. Thus, the criterion against which the
estimates were compared could be associated with only the primary latent trait, both latent traits
or the multidimensional composite of the latent traits. For the purposes of this study, the focus
was on how well the item parameters were estimated for the primary or target latent trait. The
target latent trait was defined as the one for which the discrimination value was simulated to
be larger in the semicomplex condition. Thus, for example, Items 14 and 15 both were simulated
to have larger discrimination values for the first latent trait in the semicomplex condition, making
it the primary dimension for these items. For each replication, the discrimination and difficulty
parameter estimates for items 14 and 15 were compared with the generating parameter values
associated with the first latent trait in calculating the values of bias and standard error for
both ULS and the unidimensional analyses. The primary latent trait was defined in the same
way (as the one for which the item had the largest discrimination value in the semicomplex condition) for both types of nonsimple structure simulated here.

The parameters for the primary latent trait were selected as the criterion for comparison because they reflected the case in which a researcher believes that each item is primarily associated with one latent dimension. Such a situation might arise if an exam intended to measure reading comprehension included some items about a particular historical context. Unbeknownst to the researcher, a second dimension beyond reading (i.e., knowledge of this specific period in history) might affect the responses to these items. However, the primary focus of the items remains reading so that the researcher is primarily interested in how well the items discriminate on the reading trait, as well as how difficult they are in terms of reading. The results of this study are intended to inform researchers about the potential consequences of modeling such nonsimple structure in different ways.

Results

Discrimination Parameter

A factorial ANOVA was used to identify the main effects and interactions of the manipulated factors that were significantly related to the level of bias in the nonsimple structure items. A separate analysis was run for each of these four items, and it was found that the statistically significant terms affecting estimation bias were identical across them. Therefore, ANOVA results are reported only for the first nonsimple structure item. However, mean bias is reported for all of the simple structure items. Note that relative bias was similar across the simple structure items and the correlation between this outcome and magnitude of the parameter value was near 0. Therefore, it seemed reasonable to combine results for these items. Two interactions were found to be statistically significant: type of nonsimple structure by distribution of the latent traits ($\eta^2 = .423$) and sample size by underlying model ($\eta^2 = .547$). In addition, the main effect for correlation between the latent traits was also found to be statistically significant ($\eta^2 = .492$). The significant main effects and other interactions were all subsumed by these interactions and thus are not discussed here.

Table 1 contains the average relative bias for the ULS and two types of BILOG estimates for the terms sample size and underlying model. Perhaps the most notable outcome in this table is the relatively higher positive bias values for the BILOG estimates as compared with ULS, for the nonsimple structure items. The greatest bias occurred for the estimates in which the nonsimple structure of the item responses was completely ignored (NB1). However, even when items associated with the two latent traits were modeled separately (NB2), the level of bias was still greater than when the multidimensional structure of the data was correctly modeled with NOHARM (NN). Relative bias for all of the estimation methods was higher for the nonsimple structure items than for those that exhibited simple structure in the population. The ULS estimates generally displayed somewhat lower bias than either of the BILOG methods in the simple structure case, with a greater difference with smaller sample sizes. In terms of the standard errors of the discrimination parameter estimates, in many of the simulated conditions the NB1 method had the highest values, although this was not always the case. The standard errors for the combined BILOG approach were also lower for the simple structure items in most cases. Finally, all of the methods studied here displayed somewhat higher standard errors in the M3PL case, particularly for the smaller sample-size conditions.

The results in Table 2 indicate that bias was higher in the semicomplex condition for the nonsimple structure items. This result was most notable for both BILOG conditions (NB1, NB2), although it was present for ULS as well. In addition, when the data were not normally distributed,
all methods had greater relative bias, with NB1 exhibiting the highest values. Across nonsimple structure conditions, ULS displayed the least amount of bias, whereas for the normal distribution and simple structure items, relative bias was similar for the three methods. The standard errors were higher in the nonsimple structure case for NB1 across all conditions, and for NB2 under complex structure. The ULS standard errors did not display any particular pattern with respect to the type of structure, although they were smaller when the latent trait was normally distributed. In terms of the correlation among the latent traits, higher values were associated with greater bias for all three methods of estimation. In addition, for NB2 a stronger correlation between the latent traits was associated with a larger standard error.

### Difficulty Parameter

As with the discrimination parameter estimates, an ANOVA was used to identify the main effects and interactions of the manipulated factors that were significantly related to the difficulty...
parameter bias for the nonsimple structure items. And as was true for the discrimination parameter values, results for these four items were very similar in terms of the significant terms. For this reason, only ANOVA results for the first nonsimple structure items are reported here. The interaction of correlation between the latent traits by type of nonsimple structure by distribution of the latent trait was the highest-order significant term ($p = .035$, $\eta^2 = .328$). The only other term found to be statistically significant, other than those subsumed in this interaction, was the main effect of the underlying model ($p < .001$, $\eta^2 = .190$).

To determine whether there was a relationship between the magnitude of the population difficulty parameter and the amount of bias, Pearson correlation coefficients were calculated between the actual item difficulty and the relative bias for each of the simulated conditions. Results showed that there were no significant relationships between these two variables, with a Pearson’s $r$ of .03.

The relative bias in difficulty parameter estimates for the three estimation methods by correlation, type of item, and distribution of the latent traits appears in Table 3. As was true for item discrimination, bias was higher for the nonsimple structure items, although in this case the bias was generally negative rather than positive (not reflected in the table, which includes absolute values of bias), suggesting an underestimation of item difficulty. In addition, difficulty bias was greater when the data were not normally distributed, regardless of the estimation method used. In the simple structure case, the relative bias displayed by estimates obtained from ULS and the two BILOG approaches were very comparable across conditions. With respect to the nonsimple structure items, the relative bias was higher for the NB1 approach for both the complex and semicomplex data structures. Furthermore, when the data were not normally distributed, NN and NB2 exhibited comparable levels of bias and in the normal case NB2 had somewhat higher levels than did NN, particularly for higher correlations among the latent traits.

<table>
<thead>
<tr>
<th>$r$</th>
<th>T</th>
<th>D</th>
<th>SN</th>
<th>SB1</th>
<th>SB2</th>
<th>NN</th>
<th>NB1</th>
<th>NB2</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>Complex</td>
<td>Nonnormal</td>
<td>.17/1.02</td>
<td>.17/1.15</td>
<td>.18/1.11</td>
<td>.23/1.28</td>
<td>.33/1.04</td>
<td>.23/1.15</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>.02/.94</td>
<td>.03/1.18</td>
<td>.02/.93</td>
<td>.03/1.24</td>
<td>.11/1.06</td>
<td>.03/1.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Semicomplex</td>
<td>Nonnormal</td>
<td>.15/1.00</td>
<td>.16/1.07</td>
<td>.16/1.06</td>
<td>.21/1.04</td>
<td>.31/1.00</td>
<td>.22/1.10</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>.02/.94</td>
<td>.02/8.4</td>
<td>.02/8.8</td>
<td>.02/9.7</td>
<td>.08/9.4</td>
<td>.03/1.03</td>
<td></td>
</tr>
<tr>
<td>.30</td>
<td>Complex</td>
<td>Nonnormal</td>
<td>.16/1.01</td>
<td>.16/1.06</td>
<td>.17/1.19</td>
<td>.24/1.28</td>
<td>.38/9.5</td>
<td>.24/1.17</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>.04/96</td>
<td>.03/1.24</td>
<td>.04/.94</td>
<td>.02/1.27</td>
<td>.15/8.9</td>
<td>.06/1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Semicomplex</td>
<td>Nonnormal</td>
<td>.14/1.02</td>
<td>.16/1.36</td>
<td>.16/1.13</td>
<td>.20/1.06</td>
<td>.35/9.9</td>
<td>.21/1.04</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>.02/93</td>
<td>.02/9.9</td>
<td>.02/9</td>
<td>.02/9.8</td>
<td>.11/8.8</td>
<td>.06/9.0</td>
<td></td>
</tr>
<tr>
<td>.50</td>
<td>Complex</td>
<td>Nonnormal</td>
<td>.18/1.03</td>
<td>.18/1.26</td>
<td>.18/1.13</td>
<td>.28/1.32</td>
<td>.42/7.2</td>
<td>.30/8.0</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>.02/.94</td>
<td>.03/1.03</td>
<td>.04/95</td>
<td>.01/1.26</td>
<td>.18/8.1</td>
<td>.07/8.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Semicomplex</td>
<td>Nonnormal</td>
<td>.16/9.9</td>
<td>.14/1.18</td>
<td>.18/1.19</td>
<td>.22/1.07</td>
<td>.38/9.2</td>
<td>.27/1.11</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>.04/95</td>
<td>.04/9.6</td>
<td>.04/9.1</td>
<td>.02/1.00</td>
<td>.16/7.8</td>
<td>.07/8.0</td>
<td></td>
</tr>
<tr>
<td>.80</td>
<td>Complex</td>
<td>Nonnormal</td>
<td>.15/1.01</td>
<td>.14/1.02</td>
<td>.16/1.01</td>
<td>.39/1.43</td>
<td>.46/8.2</td>
<td>.34/8.4</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>.02/98</td>
<td>.03/9.5</td>
<td>.01/9.4</td>
<td>.03/3.3</td>
<td>.21/7.3</td>
<td>.07/7.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Semicomplex</td>
<td>Nonnormal</td>
<td>.15/1.02</td>
<td>.15/1.05</td>
<td>.17/1.09</td>
<td>.29/1.09</td>
<td>.40/8.5</td>
<td>.33/8.1</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
<td>.01/1.00</td>
<td>.01/9.6</td>
<td>.02/9.2</td>
<td>.02/1.04</td>
<td>.15/7.7</td>
<td>.08/7.6</td>
<td></td>
</tr>
</tbody>
</table>

Note: $r$ = correlation between latent traits; $T$ = type of item; $D$ = distribution of latent traits; SN = simple structure for NOHARM; SB1 = simple structure for BILOG with combined estimates; SB2 = simple structure for BILOG with separate estimates; NN = nonsimple structure NOHARM; NB1 = nonsimple structure for BILOG with combined estimates; NB2 = nonsimple structure for BILOG with separate estimates.
ULS and both BILOG techniques typically displayed lower standard errors when the latent traits were normally distributed, regardless of the correlation between them or the type of non-simple structure. In addition, for the nonsimple structure items, estimates associated with both BILOG methods had declines in the standard error as the correlation between the latent traits increased, regardless of the type of structure or distribution. On the other hand, the ULS standard errors increased somewhat concomitantly with increasing correlation values. The ULS-based standard errors were higher for the nonsimple structure items, although this result was not in evidence for BILOG.

Relative bias and standard errors based on sample size and underlying model appear in Table 4. Based on the results of the ANOVA, sample size was not significantly related to bias whereas the underlying model was. For the ULS-based estimates, relative bias was somewhat higher in the M3PL case for both simple and nonsimple structure items, whereas there were essentially no differences in bias for the BILOG estimates across models. The standard errors were slightly higher in the M3PL case across methods and type of item, except for NB1.

**Recovery of Interfactor Correlations and Pseudoguessing Parameters**

Although not a primary focus of this study, the ability of ULS to accurately recover the correlations between the latent traits was investigated as was the ability of BILOG to estimate the pseudoguessing value that was subsequently given to NOHARM as a fixed value. Both values play a role in the estimation of the item discrimination and difficulty parameters, and thus themselves warrant some attention. Across the conditions simulated in this study, ULS was able to recover intertrait correlations with somewhat greater accuracy in the complex condition (mean bias of 0.02) as compared to the semicomplex (mean bias of 0.06). Results of ANOVA indicated that the bias was not influenced by the population value of the correlation, sample size, type of model or distribution of the latent traits, nor any interaction of these factors. Table 5 contains the bias values for the two types of items (complex and semicomplex) by sample size, type of model, interfactor correlation and distribution of the latent traits.

In terms of the pseudoguessing parameter, BILOG produced increasingly biased estimates as the correlation between the latent trait increased. Bias ranged from 0.01 to 0.054 across levels of the correlation. This result was consistent with the increasing bias demonstrated for the difficulty

<table>
<thead>
<tr>
<th>Table 4. Absolute Value of Relative Bias/Standard Error of Difficulty Parameter Estimates by Sample Size and Underlying Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>250</td>
</tr>
<tr>
<td>500</td>
</tr>
<tr>
<td>1,000</td>
</tr>
<tr>
<td>2,000</td>
</tr>
</tbody>
</table>

Model:
- 2: simple structure for NOHARM; SB1 = simple structure for BILOG with combined estimates; SB2 = simple structure for BILOG with separate estimates; NN = nonsimple structure NOHARM; NB1 = nonsimple structure for BILOG with combined estimates; NB2 = nonsimple structure for BILOG with separate estimates.

Note: SN = simple structure for NOHARM; SB1 = simple structure for BILOG with combined estimates; SB2 = simple structure for BILOG with separate estimates; NN = nonsimple structure NOHARM; NB1 = nonsimple structure for BILOG with combined estimates; NB2 = nonsimple structure for BILOG with separate estimates.
Discussion

It is hoped that the results of this study will provide useful information for researchers interested in the estimation of item parameters for the MIRT model. Prior research has shown that the ULS approach to parameter estimation used here is often optimal in the context of true simple structure items when compared with a variety of alternatives such as MML and RWLS (Finch, in press; Gosz & Walker, 2002; Tate, 2003). Although these earlier studies focused on the case of pure simple structure in which individual items were associated with only one of the latent traits present in the data, the goal of the current research was to extend this work to the case where some items were associated with more than one of the latent variables. It could be argued that this represents a more realistic situation in many real-world applications where the MIRT model would be appropriate (Zhang & Stone, 2008). Therefore, these results should provide further guidance to practitioners interested in modeling MIRT data, and aware of the likelihood that at least some items on their scale will be associated with more than one of the latent traits being measured.

Based on the results presented above, it appears that parameter estimates obtained using the MIRT model exhibited lower levels of bias in both discrimination and difficulty parameter estimates for items that do not exhibit simple structure when two latent traits are present than did unidimensional estimation. Indeed, for either BILOG approach, bias for both parameters was clearly an issue, with higher difficulty bias when items loaded equally on two latent traits and higher discrimination bias when they loaded more on one latent trait than another. There was also more bias in the ULS-based estimates for the nonsimple structure items, but relative to the BILOG approaches, it was much lower.

Table 5. Absolute Value of Relative Bias of Interfactor Correlation Estimates From Unweighted Least Squares by Sample Size, Underlying Model, Population Correlation Value, and Distribution of the Latent Traits

<table>
<thead>
<tr>
<th></th>
<th>Complex</th>
<th>Semicomplex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>.063</td>
<td>.018</td>
</tr>
<tr>
<td>500</td>
<td>.059</td>
<td>.019</td>
</tr>
<tr>
<td>1,000</td>
<td>.058</td>
<td>.021</td>
</tr>
<tr>
<td>2,000</td>
<td>.061</td>
<td>.023</td>
</tr>
<tr>
<td>Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.057</td>
<td>.020</td>
</tr>
<tr>
<td>3</td>
<td>.062</td>
<td>.021</td>
</tr>
<tr>
<td>Correlation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.058</td>
<td>.022</td>
</tr>
<tr>
<td>.3</td>
<td>.060</td>
<td>.022</td>
</tr>
<tr>
<td>.5</td>
<td>.062</td>
<td>.019</td>
</tr>
<tr>
<td>.8</td>
<td>.061</td>
<td>.020</td>
</tr>
<tr>
<td>Distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonnormal</td>
<td>.063</td>
<td>.024</td>
</tr>
<tr>
<td>Normal</td>
<td>.058</td>
<td>.017</td>
</tr>
</tbody>
</table>

Note: ULS = unweighted least squares; N = sample size; M = model;
In terms of direction, all three methods demonstrated positive bias for the discrimination estimates, suggesting that this parameter was overestimated across conditions. In other words, a researcher using these results would assume that his or her item was better at discriminating among examinees than it actually was. On the other hand, item discrimination parameters were consistently underestimated when the latent traits were nonnormal. It is important to note in this regard that the type of nonnormality simulated here involved negative skewness, corresponding to the case where most examinees had fairly high ability scores but a few had much lower ones. To ascertain whether this underestimation was associated with the direction of skewness, a small number of simulations were conducted in which the −1.5 skewness used here was replaced with a +1.5. The resulting difficulty estimates for all three methods were positively skewed, suggesting that the direction of bias for this parameter was indeed associated with the direction of skewness.

**Implications for Practice**

The results presented above have implications for practitioners working with multidimensional IRT data in which one or more items do not exhibit simple structure. Perhaps most importantly, ignoring the multidimensional structure will almost surely result in biased discrimination and difficulty parameter estimates. Although the current study does not investigate the impact of ignoring the structure on the estimation of examinee ability, this would certainly be a potential area of concern and focus of future research, given the results presented here. Indeed, the parameter estimate bias described above for the case where multidimensionality was ignored echoes prior work in this area (Ansley & Forsyth, 1985; Reckase, 1985; Reckase et al., 1986; Camilli et al., 1995). In the worst cases, relative bias exceeded 0.40, or more than one third the actual value of the parameter itself. Although separating the items based on the known latent structure and then using a unidimensional IRT model resulted in lower bias than when simply ignoring the multidimensionality, it did not generally produce estimates as accurate as those based on an actual MIRT model estimated using ULS, particularly in the case of item discrimination. Thus, one major implication of this study for practitioners is that they consider using the MIRT model when it is known that a measure is multidimensional, particularly when nonsimple structure items are present.

A second implication of this study for practitioners is that parameter estimation for nonsimple structure items will generally not be as accurate as for simple structure items, even for ULS. The exception to this result is item difficulty when the latent traits are both normally distributed and ULS is used for estimation, in which case bias for both types of nonsimple structure items was very similar to that of the simple structure. Furthermore, for skewed latent traits, the direction of item difficulty bias seems to be associated with the direction of skewness. In this study, the latent traits were simulated with a negative skew, and difficulty was consistently underestimated in that case. However, additional simulations with a positive skew resulted in positively biased difficulty estimates at approximately the same magnitude reported here.

When it is known (or suspected) that some items on a scale exhibit a pattern similar to the semicomplex structure simulated here, practitioners should be cognizant of the potential for greater discrimination parameter estimate bias regardless of the estimation method used. For example, in the semicomplex case the bias in ULS discrimination estimates for the primary latent trait was negative, whereas for the secondary latent trait the bias was positive. In other words, it would appear that the discrimination parameter estimates for the two latent variables were drawn toward one another, creating biased values for both. In this case, the researcher would need to interpret these estimates with caution, given that they may not be an accurate estimate of discrimination for the item on either latent trait.
Finally, the results of this study point more generally to the benefits of using the appropriate model when it is known that data are multidimensional in nature. Although in some instances the relative bias for the unidimensional estimation based on separate item sets was not much greater than that of the MIRT estimates based on ULS, in no case did this approach work better. Furthermore, it makes the tacit assumption that the researcher is able to correctly divide the items into appropriate groups based on their association with the latent traits being measured. Although such is also true of the confirmatory MIRT model used here, it is possible to use software such as NOHARM in an exploratory fashion to identify probable latent trait–item matches. Even if this were done and then the items divided into groups and estimates provided by BILOG (or some other unidimensional modeling software), these results demonstrate that ULS generally provides less biased estimates for either parameter. It is important to note that regardless of estimation method, when items are misspecified with respect to their associations with the latent traits, none of the methods investigated here would be expected to provide particularly accurate estimates. Prior work in the area of structural equation modeling has demonstrated that model misspecification results in biased parameter estimates and inflated standard errors (French & Finch, 2008). Given the likelihood of more estimation problems in the presence of such item misclassification, the current results can be viewed as a lower bound for bias and standard errors.

Limitations and Future Research

There are limitations to this study that future research should address. First of all, the latent structure simulated here was fairly simple, with only two latent traits and an equal number of items associated with each. Clearly in many real-world situations multiple dimensions may be present and they will not be associated with an equal number of items. The goal of the current study was to further prior work in MIRT parameter estimation, which has traditionally focused on the simple structure case, by examining bias and standard errors in fairly simple, clear conditions. However, future studies should extend the current work by including more than two latent traits with varying numbers of items per dimension.

Future research in this area should also focus on the expansion of the current simulation study to examine other types of nonnormality in the latent traits, as well as the inclusion of more non-simple structure items and different patterns of item-to-latent-trait relationships. Whereas the current work is limited to parameter estimation using ULS via NOHARM, future studies could be expanded to include alternative approaches. However, difficulties inherent in these alternatives in the simple structure case would not likely disappear with nonsimple structure data. Problems with modeling M3PL data using RWLS would continue to be an issue, as would the factor indeterminacy associated with the exploratory factor model used by TESTFACT. There has been work done with the Markov chain Monte Carlo (MCMC) method of estimation for the MIRT model (see Reckase, 2007, for a discussion of these). However, these methods remain somewhat limited in terms of their wider use by practitioners because of the often extended time required for estimates to converge.

A final issue that must be considered when interpreting the results of this study as well as in planning for future research involves the relationship between the correlation between the latent traits and the multidimensional discrimination vector for items, particularly those not exhibiting simple structure. Prior work by Smith and Habing (2007) demonstrated that for items not exhibiting simple structure, the effective correlation in a simulation study may be different from the intended correlation. They demonstrated the impact of this unanticipated correlation on the estimation of several statistics, including Cronbach’s alpha. The reliability estimate was somewhat elevated in comparison to a corrected value in which the simulated correlation was equal to its intended value for the nonsimple structure items. This earlier work must be considered when
interpreting the results of the current study, and future research should incorporate the corrections suggested by Smith and Habing.

**Declaration of Conflicting Interests**

The author(s) declared no conflicts of interests with respect to the authorship and/or publication of this article.

**Funding**

The author(s) received no financial support for the research and/or authorship of this article.

**References**


