Mplus Tutorial

August 2012

Mplus for Windows: An Introduction

Section 1: Introduction

1.1. About this Document

This document introduces you to Mplus for Windows. It is primarily aimed at first time users of Mplus who have prior experience with either exploratory factor analysis (EFA), or confirmatory factor analysis (CFA) and structural equation modeling (SEM). The document is organized into six sections. The first section provides a brief introduction to Mplus and describes how to obtain access to Mplus. The second section briefly reviews SEM assumptions and describes important and useful model fitting features that are unique to Mplus. The third section describes how to get started with Mplus, how to read data from an external data file, and how to obtain descriptive sample statistics. The fourth section explains how to fit exploratory factor analysis models for continuous and categorical outcomes using Mplus. The fifth section of this document demonstrates how you can use Mplus to test confirmatory factor analysis and structural equation models. The sixth section presents examples of two advanced models available in Mplus: multiple group analysis and multilevel SEM. By the end of the course you should be able to fit EFA and CFA/SEM models using Mplus. You will also gain an appreciation for the types of research questions well-suited to Mplus and some of its unique features.

1.2. Introduction to EFA, CFA, SEM and Mplus

Exploratory factor analysis (EFA) is a method of data reduction in which you may infer the presence of latent factors that are responsible for shared variation in multiple measured or observed variables. In EFA each observed variable in the analysis may be related to each latent factor contained in the analysis. By contrast, *confirmatory factor analysis* (CFA) allows you to stipulate which latent factor is related to any given observed variable. *Structural equation modeling* (SEM) is a more general form of CFA in which latent factors may be regressed onto each other. Mplus can fit EFA, CFA, and SEM models.

To effectively use and understand the course material, you should already know how to conduct a multiple linear regression analysis and compute descriptive statistics such as frequency tables using SAS, SPSS, or a similar general statistical software package. You should also understand how to interpret the output from a multiple linear regression analysis. This document also assumes that you are familiar with the statistical assumptions of EFA, CFA, and SEM, and you are comfortable using syntax-based software programs such as SAS. If you do not have experience with CFA or SEM, see our AMOS tutorial for more information about SEM. Finally, you should understand basic Microsoft Windows navigation operations: opening files and folders, saving your work, recalling previously saved work, etc.

1.3. Accessing Mplus

You may access Mplus in one of three ways:

License a copy from Muthén & Muthén for your own personal computer.

Mplus is available to faculty, students, and staff at the University of Texas at Austin via the

STATS Windows terminal server. To use the terminal server, you must obtain an ITS computer account (an IF or departmental account) and then validate the account for Windows NT Services. You then download and configure client software that enables your PC, Macintosh, or UNIX workstation to connect to the terminal server. Finally, you connect to the server and launch Mplus by double-clicking on the Mplus for Windows program icon located in the STATS terminal server program group. Details on how to obtain an ITS computer account, account use charges, and downloading client software and configuration instructions may be found in General FAQ: Connecting to published statistical and mathematical applications on the ITS Windows Terminal Server.

Download the free student version of Mplus from the Muthén & MuthénWeb site for your own personal computer. If your models of interest are small, the free demonstration version may be sufficient to meet your needs. For larger models, you will need to purchase your own copy of Mplus or access the ITS shared copy of the software through the campus network. The latter option is typically more cost effective, particularly if you decide to access the other software programs available on the server (e.g., SAS, SPSS, AMOS, etc.).

1.4. Getting Help with Mplus

If you have difficulties accessing Mplus on the Windows Terminal Server, call the ITS helpdesk at 512-475-9400 or send e-mail to help@its.utexas.edu.

If you are able to log in to the Windows Terminal Server and run Mplus, but have questions about how to use Mplus or interpret output, call the ITS helpdesk at 512-475-9400 to schedule an appointment with an SSC statistical consultant or send e-mail to stats@ssc.utexas.edu.

Important note: Both services are available to University of Texas faculty, students, and staff only. See our Web site at http://ssc.utexas.edu/consulting/ for more details about consulting services, as well as frequently asked questions and answers about EFA, CFA/SEM, Mplus, and other topics. Non-UT and UT Mplus users will find the Muthén & MuthénWeb site to be a useful resource; see the Mplus Discussion forum for frequently-asked questions and answers. You may also post your own questions in this forum.

The *Mplus User's Guide* is available for check out from the PCL general circulation desk. Alternatively, you may order a copy from the Muthén & Muthén Web site.

Section 2: Latent Variable Modeling Using Mplus

2.1. Overview of SEM Assumptions for Continuous Outcome Data

Before specifying and running a latent variable model, you should give some thought to the assumptions underlying latent variable modeling with continuous outcome variables. Several of these assumptions are:

- A theoretical basis for model specification

- A reasonable sample size
- Identified model equations
- Complete data or appropriate handling of incomplete data
- Continuously and normally distributed endogenous variables

These assumptions apply equally to all EFA and CFA/SEM software programs. The details of these assumptions can be found in our AMOS tutorial, but they may be summarized as follows: Recommendations for sample size vary depending upon the complexity of the specified model, but typical figures range from 5 to 15 cases per estimated parameter with overall sample size preferred to exceed $N = 200$ cases. Furthermore, any model you consider should have a theoretical basis, and substantive inferences should be drawn based upon your ability to rule out alternative explanations for findings, rather than on statistical considerations alone.

Like AMOS, Mplus features Full Information Maximum Likelihood (FIML) handling of missing data, an appropriate, modern method of missing data handling that enables Mplus to make use of all available data points, even for cases with some missing responses. For more details on missing data handling methods, including FIML, see General FAQ: Handling missing or incomplete data and AMOS FAQ: Handling Missing Data using AMOS. One added missing data handling feature that is unique to Mplus is its ability to generate model modification indices for databases that are incomplete.

2.2. Categorical Outcomes and Categorical Latent Variables

Where Mplus diverges from most other SEM software packages is in its ability to fit latent variable models to databases that contain ordinal or dichotomous outcome variables. Note that Mplus will not yet fit models to databases with nominal outcome variables that contain more than two levels. Nonetheless, the ability to fit models to variables that contain ordinal and dichotomous categorical outcome variables is very useful. Furthermore, Mplus will fit *latent class analysis* (LCA) models that contain categorical latent variables and fit *mixture models* that generate expected classifications of observations based upon the characteristics of your specified model.

2.3. Should you use Mplus?

Should you use Mplus to perform EFA, CFA, and SEM analyses on your data? In order to facilitate rapid access to both simple and complex latent variable models, the Mplus developers have built a streamlined set of data import and model specification commands. All Mplus commands are specified using command syntax, though a syntax generator is under development at the time of this writing. If you are not comfortable with reading data and specifying statistical models using command syntax, Mplus may not be the optimal choice for you. On the other hand, if you prefer to work with command syntax when you use statistical software programs or you do

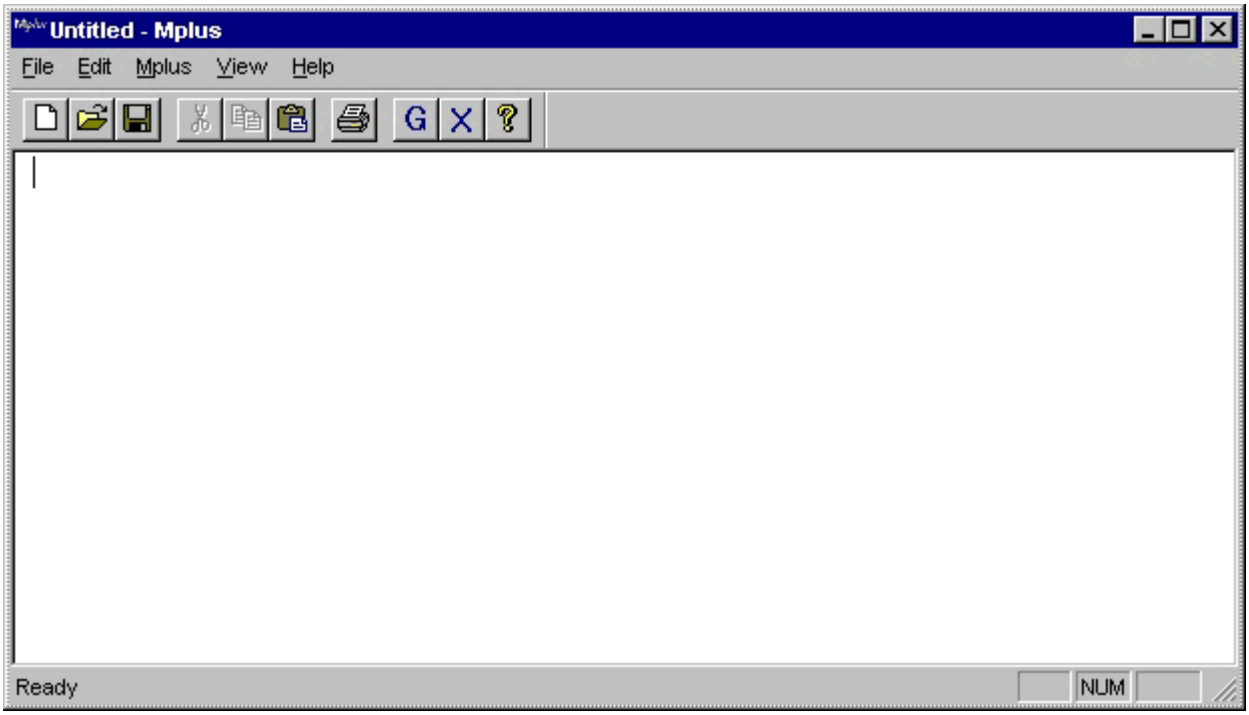
not mind learning software syntax to perform data analysis, you will probably find it useful to learn Mplus. This is particularly true when you consider some of the features unique to Mplus:

- The ability to build models with dichotomous and ordered categorical outcome variables
- The capacity to build models that contain categorical latent variables
- Optimal full information maximum likelihood (FIML) missing data handling for both exploratory as well as CFA and SEM models
- Modification index output, even when you invoke FIML missing data handling
- The ability to fit multilevel or hierarchical CFA and SEM models

Section 3: Using Mplus

3.1. Launching Mplus

If you are using a personal or demonstration copy of Mplus, locate the *Mplus* entry in the *Program Files* subsection of the Microsoft Windows *Start* menu.. If you are using the STATS Windows terminal server, locate the *Mplus for Windows* icon in the Citrix Program Neighborhood and double-click on it to launch Mplus. You will then be prompted to enter your Windows NT Services account name, your password, and the domain name, WNT. After you have entered this information, click **OK** to launch Mplus. Once you have launched Mplus, you will see the following window appear on your computer's desktop:



3.2. The Input and Output Windows

The window shown above is the *input window*. You write Mplus syntax in this window to read the data to be analyzed and to specify your model of interest. You then save your Mplus syntax and select *Run Mplus* from the ***Mplus*** menu to submit your syntax to the Mplus engine for processing:

Mplus
    Run Mplus

Note: If you are using Mplus on the STATS terminal server, do not save your work to the default Mplus directory. Instead, save your work on one of the client disk drives or your allocated WNTDISK server space. The server space is mounted as drive U and has the advantage of being available to you whenever you log in to the terminal server. There is, however, a nominal fee associated with using this space for file storage. You may also save your files on a local disk drive. Each local drive is preceded by a $ (e.g., $C:) in the list of available disk drives shown in the ***Save As*** menu option in the ***File*** menu.

Once Mplus has finished processing your command syntax, it replaces the input window with the *output window*. The output window first displays your Mplus syntax. Below the Mplus syntax are the Mplus model results. If there is an error in your Mplus syntax or you want to modify your Mplus syntax in any way (e.g., to fit a different model to the data), you must return to the appropriate command file by selecting that file's name from the ***File*** menu's list of recently-accessed files. That action returns the input window's contents to the screen and you can then modify the previous commands, save the modified command file, and run Mplus once again to obtain new output.

3.3. Reading Data and Outputting Sample Statistics

After you have launched Mplus, you may build a command file. There are nine Mplus commands: **TITLE**, **DATA** (required), **VARIABLE** (required), **DEFINE**, **SAVEDATA**, **ANALYSIS**, **MODEL**, **OUTPUT**, and **MONTECARLO**. The most commonly used Mplus commands are described in this document. According to the *Mplus User's Guide*, "The Mplus commands may come in any order. The **DATA** and **VARIABLE** commands are required for all analyses. All commands must begin on a new line and must be followed by a colon. Semicolons separate command options. There can be more than one option per line. The records in the input setup must be no longer than 80 columns. They can contain upper and/or lower case letters and tabs." (page 1).

A description of the Mplus defaults appears in Mplus FAQ: Mplus Defaults. You should review these defaults carefully and be sure that you understand them fully prior to analyzing data with Mplus.

The first Mplus syntax to appear in the command file is typically a **TITLE** command. The **TITLE** command allows you to specify a title that Mplus will print on each page of the output

file.

Following the **TITLE** command is the **DATA** command. The **DATA** command specifies where Mplus will locate the data, the format of the data, and the names of variables. At present, Mplus will read the following file formats: tab-delimited text, space-delimited text, and comma-delimited text. The input data file may contain records in free field format or fixed format. If you are using data stored in another form (e.g., SAS, SPSS, or Excel), you will need to convert it to one of the formats with which Mplus can work before you read it into Mplus. See our FAQs or information on how to convert common statistical data file formats to plain, comma-delimited, or tab-delimited text files.

The next command is the **VARIABLE** command. The **VARIABLE** command names the columns of data that Mplus reads using the **DATA** command.

Following the **VARIABLE** command is the **ANALYSIS** command. The **ANALYSIS** command tells Mplus what type of analysis to perform. Many analysis options are available; a number of these are shown in the examples that appear in this document.

Consider the following example database: In 1939 Karl Holzinger and Francis Swineford administered 26 aptitude tests to 145 students in the Grant-White School. Of the 26 tests, six are used here: visual perception, cubes, lozenges, paragraph comprehension, sentence completion, and word meaning. An additional variable, gender, is included in the database, but not used in this example. This database is available in SPSS format as one of the example datasets used by the AMOS SEM software package. AMOS and its example program files and datasets are available on the STATS terminal server; a free student version of AMOS containing this database may be downloaded from the Smallwaters Corporation Web site. The SPSS file's name is *grant.sav*. You can download this file in tab-delimited text format HERE. Then you can write the following Mplus syntax to read the data from the file.

```
TITLE:          Grant-White School:  Summary Statistics

DATA:           FILE IS U:\Projects\Documentation\Mplus\grant.dat ;
                FORMAT IS free ;

VARIABLE:
                NAMES ARE   visperc
                            cubes
                            lozenges
                            paragrap
                            sentence
                            wordmean
                            gender ;

                USEVARIABLES ARE    visperc
                                    cubes
                                    lozenges
                                    paragrap
```

```
                                    sentence
                                    wordmean ;

    ANALYSIS:       TYPE = basic ;
```

In this sample program, the **DATA** command uses the **FILE** subcommand to tell Mplus where to locate the relevant data file. In this case, the file's location is U:\Projects\Documentation\Mplus\grant.dat. The **FORMAT** subcommand uses the default ***free*** option to let Mplus know that the data points appear in order in the data file with the data points separated by commas, tabs, or spaces. Alternatively, you can use FORTRAN format statements to read data when data are in fixed columns. FORTRAN-formatted input is recommended for large databases because it is more efficient than the default ***free*** data field input; see the Mplus manual for a detailed description of how to specify FORTRAN input formats.

The next command shown is the **VARIABLE** command. The **VARIABLE** command uses the **NAMES** subcommand to list the variables contained in the Grant-White database. While it is possible to have more than one variable name on a row of the command file, this example lists the variables with one variable per line because the appearance of the variable names in the command file is easy to read. Because Mplus allows variable names to have a maximum width of eight characters, the variable name "paragraph" is shortened to paragraph.

Following the **NAMES** subcommand is the **USEVARIABLES** subcommand. **USEVARIABLES** enables you to specify a particular subset of variables to be used in the data analysis. A similar subcommand, **USEOBS**, allows you to select subsets of cases to be used in a particular analysis. For example, if you wanted to limit the analysis to female participants, you could include the subcommand

```
    USEOBS gender EQ 1 ;
```

where a *gender* value of 1 designated female cases in the database.

The **ANALYSIS** command specifies the **TYPE** of analysis to be performed by Mplus. In this example the type is ***basic***. The basic model type does not have Mplus fit any model to the sample data; instead Mplus will compute sample statistics only. Using basic as the analysis type is useful during the initial phase of building your command file because you can use the Mplus sample statistics output to compare Mplus results to results you obtained using SAS, SPSS, Excel, or other statistical software programs to verify that Mplus is reading your input data correctly.

It is worth noting that Mplus has many default settings that enable you to write compact syntax, which results in brief command files. Once you understand the Mplus defaults fully, you may take advantage of them to write shorter command files. For instance, the first example shown above may be simplified:

```
TITLE:          Grant-White School:  Summary Statistics
```

DATA: **FILE IS U:\Projects\Documentation\Mplus\grant.dat ;**

VARIABLE:
NAMES ARE visperc
cubes
lozenges
paragrap
sentence
wordmean
gender ;

USEVARIABLES ARE visperc - wordmean ;

ANALYSIS: **TYPE = *basic* ;**

The **FORMAT is free** statement has been omitted because the default format is free-field data input. The **USEVARIABLES** statement also shows a handy Mplus feature, the variable list option. The variable list option enables you to conveniently refer to a list of variables using a dash to separate the first and last variables in the contiguous series of variables.

The output from the basic analysis appears below. Although Mplus initially returns a copy of the input command file, that portion of the output has been omitted here in the interest of saving space.

```
SUMMARY OF ANALYSIS
Mplus VERSION 1.04
PAGE    2
Holzinger and Swineford Grant-White School Summary Statistics


Number of groups                                 1
Number of observations                         145

Number of y-variables                            6
Number of x-variables                            0
Number of continuous latent variables            0

Observed variables in the analysis
   VISPERC     CUBES       LOZENGES    PARAGRAP    SENTENCE
WORDMEAN


Estimator                                       ML
Maximum number of iterations                  1000
Convergence criterion                     .500D-04

Input data file(s)
```

```
  U:\Projects\Documentation\Mplus\grant.dat

Input data format  FREE


RESULTS FOR BASIC ANALYSIS


     SAMPLE STATISTICS


            Means/Intercepts/Thresholds
                  1             2             3
4             5
               ________      ________      ________      ________
________
      1        29.579        24.800        15.966         9.952
18.848


            Means/Intercepts/Thresholds
                  6
               ________
      1        17.283


            Covariances/Correlations/Residual Correlations
               VISPERC       CUBES         LOZENGES      PARAGRAP
SENTENCE
               ________      ________      ________      ________
________
 VISPERC       47.801
 CUBES         10.012        19.758
 LOZENGES      25.798        15.417        69.172
 PARAGRAP       7.973         3.421         9.207        11.393
 SENTENCE       9.936         3.296        11.092        11.277
21.616
 WORDMEAN      17.425         6.876        22.954        19.167
25.321


            Covariances/Correlations/Residual Correlations
               WORDMEAN
               ________
 WORDMEAN      63.163
```

Mplus initially identifies the number of groups and observations in the analysis, followed by the number of X (predictor) and Y (outcome) variables and the sample (input) covariances, variances, and means. Once you have verified that these values are correct, you can turn your attention to fitting your model(s) of interest. The next section continues with the same example database, but describes how to perform an exploratory factor analysis of the continuous variables

in the Grant-White database using Mplus.

Section 4: Exploratory Factor Analysis

4.1. Exploratory Factor Analysis with Continuous Variables

Once you have read the data into Mplus and verified that the sample statistics show that the data have been read correctly, you can perform exploratory factor analysis using Mplus by altering the **ANALYSIS** command as follows:

ANALYSIS: TYPE = *efa* 1 2 ;
ESTIMATOR = *ml* ;

This syntax instructs Mplus to perform an exploratory factor analysis of the Grant-White database. ***Efa*** tells Mplus to perform an exploratory factor analysis. The 1 and 2 following the ***efa*** specification tells Mplus to generate all possible factor solutions between and including 1 and 2. In this instance, one and two factor solutions will be produced by the analysis. Finally, the **ESTIMATOR = *ml*** option has Mplus use the maximum likelihood estimator to perform the factor analysis and compute a chi-square goodness of fit test that the number of hypothesized factors is sufficient to account for the correlations among the six variables in the analysis. This optional specification overrides the default unweighted least-square (***uls***) estimator.

If your data are not joint multivariate normally distributed, you may want to replace the ***ml*** with either the ***mlm*** or ***mlmv*** estimators. One useful feature of Mplus is its ability to handle non-normal input data. Recall that the default ***ml*** estimator assumes that the input data are distributed joint multivariate normal. If you have reason to believe that this assumption has not been met and your sample is reasonably large (e.g., $N = 200$), you may substitute ***mlm*** or ***mlmv*** in place of ***ml*** on the **ESTIMATOR =** line. The ***mlm*** option provides a mean-adjusted chi-square model test statistic whereas the ***mlmv*** option produces a mean and variance adjusted chi-square test of model fit. SEM users who are familiar with Bentler's EQS software program should also note that the ***mlm*** chi-square test and standard errors are equivalent to those produced by EQS in its **ML;ROBUST** method.

You may also add the **OUTPUT** command following the **ANALYSIS** command. The **OUTPUT** command is used to specify optional output. For this example the keyword ***sampstat*** tells Mplus to include sample statistics as part of its printed output.

OUTPUT: *sampstat* ;

Mplus produces the sample correlations, eigenvalues, and the chi-square test of the one factor model to the sample data. As you can see from the results, shown below, the chi-square test is statistically significant, so the null hypothesis that a single factor fits the data is rejected; more factors are required to obtain a non-significant chi-square. Since the chi-square test is sensitive to sample size (such that large samples often return statistically significant chi-square values) and non-normality in the input variables, Mplus also provides the *Root Mean Square Error of*

Approximation (*RMSEA*) statistic. The RMSEA is not as sensitive to large sample sizes. According to Hu and Bentler (1999), RMSEA values below .06 indicate satisfactory model fit. The RMSEA yielded a result of .162, which was consistent with the chi-square result in suggesting that the one factor model does not fit the data adequately.

```
CONTINUOUS VARIABLE CORRELATION MATRIX
              VISPERC       CUBES         LOZENGES      PARAGRAP
SENTENCE
              ________      ________      ________      ________
________
 VISPERC
 CUBES           .326
 LOZENGES        .449          .417
 PARAGRAP        .342          .228          .328
 SENTENCE        .309          .159          .287          .719
 WORDMEAN        .317          .195          .347
.714           .685


              Grant-White School: Exploratory Factor Analysis


           EXPLORATORY ANALYSIS WITH  1 FACTOR(S) :



           EIGENVALUES FOR SAMPLE CORRELATION MATRIX
                 1             2             3
4             5
              ________      ________      ________      ________
________
      1         3.009         1.225          .656
.530           .311


           EIGENVALUES FOR SAMPLE CORRELATION MATRIX
                 6
              ________
      1          .270


           EXPLORATORY ANALYSIS WITH  1 FACTOR(S) :
           CHI-SQUARE VALUE              43.241
           DEGREES OF FREEDOM                9
           PROBABILITY VALUE              .0000

           RMSEA (ROOT MEAN SQUARE ERROR OF APPROXIMATION) :
           ESTIMATE (90 PERCENT C.I.) IS   .162 (  .115   .212)
           PROBABILITY RMSEA LE  .05 IS     .000
```

Mplus next produces the estimated factor loadings and error variances. Notice that the *visperc*,

cubes, and *lozenges* factor loadings are low relative to the other factor loadings displayed below.

```
ESTIMATED FACTOR LOADINGS
                   1

              ________
 VISPERC         .415
 CUBES           .272
 LOZENGES        .415
 PARAGRAP        .865
 SENTENCE        .818
 WORDMEAN        .827


            ESTIMATED ERROR VARIANCES
               VISPERC       CUBES         LOZENGES      PARAGRAP
SENTENCE

               ________      ________      ________      ________
________
                 .828          .926          .828
.252           .330

               ________
      1           .316
```

The estimated correlation matrix is the correlation matrix reproduced by Mplus under the assumption that a single factor is sufficient to explain the sample correlations. From the model fit results shown above, this is not the case, so it is not surprising that this implied or model-based correlation matrix differs substantially from the sample correlation matrix reported above.

```
ESTIMATED CORRELATION MATRIX
               VISPERC       CUBES         LOZENGES      PARAGRAP
SENTENCE

               ________      ________      ________      ________
________
 VISPERC         1.000
 CUBES            .113         1.000
 LOZENGES         .172          .113         1.000
 PARAGRAP         .359          .235          .359         1.000
 SENTENCE         .339          .223          .340
.708          1.000
 WORDMEAN         .343          .225          .343
.715           .677


               WORDMEAN
               ________
 WORDMEAN        1.000
```

The residuals matrix represents the difference between the sample correlation matrix and the

implied correlation matrix. As noted above, since the model did not fit the observed data particularly well, there are some values in this matrix that are non-trivial in size. In particular, the *cubes-visperc*, *lozenges-visperc*, and *lozenges-cubes* residual values are high relative to the other values in the matrix.

```
        RESIDUALS OBSERVED-EXPECTED
              VISPERC       CUBES         LOZENGES      PARAGRAP
SENTENCE
              ________      ________      ________      ________
________
 VISPERC          .000
 CUBES            .213          .000
 LOZENGES         .276          .304          .000
 PARAGRAP        -.017         -.007         -.031          .000
 SENTENCE        -.030         -.063         -.053
.011           .000
 WORDMEAN         -.026         -.030          .004
.000           .009


            RESIDUALS OBSERVED-EXPECTED
               WORDMEAN
               ________
 WORDMEAN         .000
```

The Root Mean Square Residual (RMR) is another descriptive model fit statistic. According to Hu and Bentler (1999), RMR values should be below .08 with lower values indicating better model fit. The value of .1225 shown below for the one factor solution indicates unacceptably poor model fit.

```
 ROOT MEAN SQUARE RESIDUAL IS         .1225
```

In short, the one factor solution was a poor fit to the data. In particular, the model did not account well for the correlations among the *visperc*, *cubes*, and *lozenges* variables. What about the two factor solution? Mplus reports the two factor solution following the single factor model. The chi-square test of model fit is non-significant, indicating that the null hypothesis that the model fits the data cannot be rejected (the model fits the data well). This finding is corroborated by the RMSEA: Its estimate is zero; it's 90% confidence interval has an upper bound value of .055, which is below the Hu and Bentler (1999) recommended cutoff value of .06. The RMSEA estimate and its upper bound confidence interval value should both fall below .06 to ensure satisfactory model fit.

```
EXPLORATORY ANALYSIS WITH  2 FACTOR(S) :


            EXPLORATORY ANALYSIS WITH  2 FACTOR(S) :
            CHI-SQUARE VALUE               1.079
```

```
          DEGREES OF FREEDOM                     4
          PROBABILITY VALUE                  .8976

          RMSEA (ROOT MEAN SQUARE ERROR OF APPROXIMATION) :
          ESTIMATE (90 PERCENT C.I.) IS    .000 (  .000   .055)
          PROBABILITY RMSEA LE  .05 IS      .944
```

For exploratory factor analysis solutions with two or more factors, Mplus reports *varimax rotated loadings* and *promax rotated loadings*.Varimax loadings assume the two factors are uncorrelated whereas promax loadings allow the factors to be correlated. Directly below the promax loadings is the factor intercorrelation matrix.

In this example the two factors are correlated .480. With even a modest correlation among the two factors, you should choose to interpret the promax rotated loadings. The loadings show that the *visperc*, *cubes*, and *lozenges* variables load onto the first factor whereas the remaining variables load onto the second factor.

```
VARIMAX ROTATED LOADINGS
                    1             2

 VISPERC         .547          .250
 CUBES           .550          .092
 LOZENGES        .728          .196
 PARAGRAP        .241          .830
 SENTENCE        .174          .816
 WORDMEAN        .247          .788


          PROMAX ROTATED LOADINGS
                 1             2

 VISPERC         .540          .112
 CUBES           .585         -.063
 LOZENGES        .755         -.001
 PARAGRAP        .046          .841
 SENTENCE       -.025          .846
 WORDMEAN        .063          .794


          PROMAX FACTOR CORRELATIONS
                 1             2

      1         1.000
      2          .480         1.000
```

Mplus next reports estimated error variances for each observed variable, the estimated correlation matrix, and the residual correlation matrix. Notice that unlike the preceding one factor solution, this dual factor solution's estimated correlation matrix is very close in value to

the original sample correlation matrix. Accordingly, the residual correlation matrix has all values close to zero and the RMR value of .0092 is well below the Hu and Bentler (1999) recommended cutoff of .08.

```
           ESTIMATED ERROR VARIANCES
              VISPERC       CUBES         LOZENGES      PARAGRAP
SENTENCE
              ________      ________      ________      ________
________
      1          .638          .689          .431
.253           .304


           ESTIMATED ERROR VARIANCES
              WORDMEAN
              ________
      1          .318


           ESTIMATED CORRELATION MATRIX
              VISPERC       CUBES         LOZENGES      PARAGRAP
SENTENCE
              ________      ________      ________      ________
________
 VISPERC        1.000
 CUBES           .324         1.000
 LOZENGES        .448          .419         1.000
 PARAGRAP        .339          .209          .338         1.000
 SENTENCE        .299          .170          .286          .719
1.000
 WORDMEAN        .332          .208          .334
.714           .686


           ESTIMATED CORRELATION MATRIX
              WORDMEAN
              ________
 WORDMEAN       1.000


           RESIDUALS OBSERVED-EXPECTED
              VISPERC       CUBES         LOZENGES      PARAGRAP
SENTENCE
              _______       ________      ________      ________
________
 VISPERC         .000
 CUBES           .002          .000
 LOZENGES        .001         -.002          .000
 PARAGRAP        .002          .019         -.010          .000
```

```
 SENTENCE          .010          -.011           .000
.000            .000
 WORDMEAN          -.015          -.013           .013
.001           -.001


            RESIDUALS OBSERVED-EXPECTED
               WORDMEAN

               _________
 WORDMEAN          .000


 ROOT MEAN SQUARE RESIDUAL IS          .0092
```

This example assumes that the Grant-White database is complete. In other words, there are no missing cases in the Grant-White database. What if some cases had missing values? Often databases have cases with incomplete data. The next section describes a feature unique to Mplus: exploratory factor analysis of a database with incomplete cases.

4.2. Exploratory Factor Analysis with Missing Data

Suppose you altered the Grant-White database so that cases with *visperc* scores that exceed 34 have missing *cubes* scores and that cases with *wordmean* scores of 10 or below have missing *sentence* values. In this instance the missing cubes and sentence completion data are said to be *missing at random* (MAR) because the patterns of missing data are explainable by the values of other variables in the database, visual perception and word meaning. Ordinarily, if you do not specify a missing data analysis in Mplus, Mplus performs *listwise* or *casewise* deletion of cases with any missing data. That is, any case with one or more missing data points is omitted entirely from analyses. However, for exploratory factor analysis, confirmatory factor analysis, and structural equation modeling with continuous variables, Mplus features a missing data option that outperforms the default listwise deletion method. The optional method that offers superior performance is called full information maximum likelihood (FIML); details on FIML can be found in General FAQ: Handling missing or incomplete Data and in AMOS FAQ: Handling missing data using AMOS.

Regardless of whether you choose to use FIML or listwise data deletion to handle missing data, if you have missing data in your input database, you must tell Mplus how the missing values for each variable are represented in the database. You use the MISSING subcommand of the **VARIABLE** command to accomplish this task. In this example, missing values for cubes and sentence are represented by -9, so the **MISSING** subcommand reads:

MISSING ARE *all* (-9) ;

The ***all*** keyword tells Mplus that all variables in the analysis use -9 to represent missing values. If your database contains blanks to represent missing values, you may use the specification

MISSING = *blank* ;

Similarly, you may use

MISSING ARE . ;

if your database contains period symbols to represent missing values. Other missing value specifications are available; see the *Mplus User's Guide* for specifics.

If you insert the **MISSING** syntax into the previous exploratory factor analysis program and specify that Mplus use the newly-created database that contains cases with missing values, grant-missing.dat, Mplus will perform listwise deletion of the cases with incomplete data. The Mplus command file follows:

TITLE: Grant-White School: EFA with Missing Data

DATA: FILE IS U:\Projects\Documentation\Mplus\grant-missing.dat ;

VARIABLE:

NAMES ARE visperc
cubes
lozenges
paragrap
sentence
wordmean
gender ;

USEVARIABLES ARE visperc - wordmean ;

MISSING ARE *all* (-9) ;

ANALYSIS: TYPE = *efa* 1 2;
ESTIMATOR = *ml* ;

Selected output from the analysis appears below.

```
Grant-White School: Exploratory Factor Analysis with Missing Data

SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                          79

Number of y-variables                            6
Number of x-variables                            0
Number of continuous latent variables            0
```

Notice that Mplus considers the database to contain 79 usable cases rather than the original 145 cases.

```
           EXPLORATORY ANALYSIS WITH  1 FACTOR(S) :
```

```
          CHI-SQUARE VALUE                14.651
          DEGREES OF FREEDOM                   9
          PROBABILITY VALUE                .1009

          RMSEA (ROOT MEAN SQUARE ERROR OF APPROXIMATION) :
          ESTIMATE (90 PERCENT C.I.) IS   .089 (  .000   .169)
          PROBABILITY RMSEA LE  .05 IS     .199
```

The one factor solution also fits the database for the 79 usable cases. This finding stands in direct contrast to the example in the previous section where all 145 cases had complete data and the one factor model was rejected. Clearly the reduction of *N* from 145 to 79 has resulted in a substantial loss of statistical power to reject false hypotheses.

Fortunately, you can use Mplus's FIML missing data handling option to rectify the problem. Add the keyword ***missing*** to the **TYPE** subcommand of the **ANALYSIS** command, like this:

ANALYSIS: TYPE = *missing efa* 1 2 ;
ESTIMATOR = *ml* ;

Run the analysis and consider the results, shown below.

```
Grant-White School: Exploratory Factor Analysis with Missing Data

SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                         145

Number of y-variables                            6
Number of x-variables                            0
Number of continuous latent variables            0
```

Mplus now uses all 145 cases in its computations.

```
SUMMARY OF DATA

     Number of patterns           4

COVARIANCE COVERAGE OF DATA

Minimum covariance coverage value    .100


     PROPORTION OF DATA PRESENT


           Covariance Coverage
              VISPERC       CUBES         LOZENGES      PARAGRAP
SENTENCE
              ________      ________      ________      _________
```

```
 VISPERC        1.000
 CUBES           .697          .697
 LOZENGES       1.000          .697         1.000
 PARAGRAP       1.000          .697         1.000         1.000
 SENTENCE        .821          .545          .821
.821           .821
 WORDMEAN       1.000          .697         1.000
1.000           .821
```

Mplus further recognizes that there are four distinct patterns of missing data contained in the database and it displays the amount of data used to generate each input covariance for the analysis. From the missing data coverage matrix, you can see that the *cubes-sentence* covariance has the lowest coverage with just under 55% of cases available to build the covariance. Mplus requires a minimum coverage value of 10% per covariance, though you can override this default if you wish.

```
EXPLORATORY ANALYSIS WITH  1 FACTOR(S) :
           CHI-SQUARE VALUE              29.732
           DEGREES OF FREEDOM                9
           PROBABILITY VALUE              .0005

           RMSEA (ROOT MEAN SQUARE ERROR OF APPROXIMATION) :
           ESTIMATE (90 PERCENT C.I.) IS   .126 (  .078   .178)
           PROBABILITY RMSEA LE  .05 IS     .007
```

Unlike the example that used listwise deletion of cases with missing data, the chi-square test of model fit for the one factor solution rejects the one factor model. Using FIML missing data handling, you conclude that one factor is not sufficient to explain the pattern of correlations among the six input variables, just as you did in the first example from the preceding section where Mplus used the complete database containing 145 cases. As with the complete dataset, the two factor solution fits the data well using the FIML method with the incomplete dataset:

```
            EXPLORATORY ANALYSIS WITH  2 FACTOR(S) :
           CHI-SQUARE VALUE                .578
           DEGREES OF FREEDOM                4
           PROBABILITY VALUE              .9655

           RMSEA (ROOT MEAN SQUARE ERROR OF APPROXIMATION) :
           ESTIMATE (90 PERCENT C.I.) IS   .000 (  .000   .000)
           PROBABILITY RMSEA LE  .05 IS     .982
```

4.3. Exploratory Factor Analysis with Categorical Outcomes

So far, the examples shown here contained continuous outcomes. If you have observed outcome variables that have ten or fewer categories, and the variables' responses are dichotomous or ordered categories, you may elect to have Mplus treat these variables as categorical indicators. This type of model is often sensible for analyzing Likert scale items because while the items

themselves typically are coarsely categorized on a 1 to 5 or 1 to 7 scale, the items often attempt to measure an individual's standing on a continuous underlying unobserved variable.

For the purposes of illustration, suppose that you recode each variable into a replacement variable where all six variables' values at the median or below are assigned a categorical value of 1.00 and all values above the median assigned a value of 2.00. Mplus recodes the lowest value to zero with subsequent values increasing in units of 1.00. While the two underlying latent factors remain continuous, the six categorical observed variables' response values are now ordered dichotomous categories. To analyze the modified database using Mplus, you may use the syntax that appeared in the initial exploratory factor analysis example, with the following modifications, and the new data file that contains the categorical variables, grantcat.dat, as shown below.

```
TITLE:     Grant-White School: EFA with categorical outcomes

DATA:      FILE IS U:\Projects\Documentation\Mplus\grantcat.dat ;

VARIABLE:

  NAMES ARE  viscat
             cubescat
             lozcat
             paracat
             sentcat
             wordcat ;

  USEVARIABLES ARE  viscat - wordcat ;

  CATEGORICAL ARE viscat - wordcat ;

ANALYSIS:    TYPE = efa 1 2;
             ESTIMATOR = wlsmv ;

OUTPUT:      sampstat ;
```

First, you must change the names of the variables in the **NAMES** and **USEVARIABLES** subcommands of the **DATA** command. Next, you tell Mplus which variables are categorical with the **CATEGORICAL** subcommand of the **DATA** command, like this:

CATEGORICAL ARE vizcat ... wordcat ;

You should also change the **ESTIMATOR** option for the **ANALYSIS** command. The default is unweighted least-squares (***uls***), which is fast and is useful for exploratory work, but a more optimal choice for categorical outcomes, based on the work of Muthén, DuToit, and Spisic (1997), is weighted least-squares with mean and variance adjustment, ***wlsmv***.

ANALYSIS: TYPE = *efa* 1 2;
ESTIMATOR = *wlsmv* ;

Selected output from the analysis appears below. Notice that the categorical nature of the data precludes computation of the descriptive model fit statistics such as the RMSEA, though Mplus does produce the familiar chi-square test of overall model fit.

```
EXPLORATORY ANALYSIS WITH  2 FACTOR(S) :
CHI-SQUARE VALUE                2.823
DEGREES OF FREEDOM                  4
PROBABILITY VALUE                .5875
```

The chi-square result for the two factor model is not significant, which indicates that two factors are sufficient to explain the intercorrelations among the six observed variables. The varimax and promax rotated factor loadings appear below. The pattern and values obtained from this analysis are consistent with the results of the first exploratory factor analysis of the completely continuous data discussed previously.

```
                VARIMAX ROTATED LOADINGS
                1             2
           ________      ________
VISCAT        .571          .332
CUBESCAT      .700          .117
LOZCAT        .667          .244
PARACAT       .473          .642
SENTCAT       .235          .847
WORDCAT       .206          .858


          PROMAX ROTATED LOADINGS
                1             2
           ________      ________
VISCAT        .559          .159
CUBESCAT      .777         -.137
LOZCAT        .698          .022
PARACAT       .347          .550
SENTCAT       .005          .876
WORDCAT      -.031          .899


          PROMAX FACTOR CORRELATIONS
                1             2
           ________      ________
     1       1.000
     2        .557         1.000
```

Although Mplus does not produce the RMSEA descriptive model fit statistic for categorical

outcomes, it does output the standardized root mean residual, RMR:

```
ROOT MEAN SQUARE RESIDUAL IS          .0310
```

The value of .031 suggests an excellent fit of the two factor model to the observed data.

There are several notes worth keeping in mind when you perform exploratory factor analysis with categorical outcome variables.

Although one or more of the observed variables may be categorical, any latent variables in the model are assumed to be continuous (this is a property of the exploratory factor analysis model; confirmatory factor analysis models with categorical latent variables may be fit as *mixture* models using Mplus; see the *Mplus User's Guide* for more information about mixture models).

FIML missing data handling is not available with the analysis of categorical outcomes.

The analysis specification and interpretation of the output is the same whether one, a subset, or all observed variables are categorical.

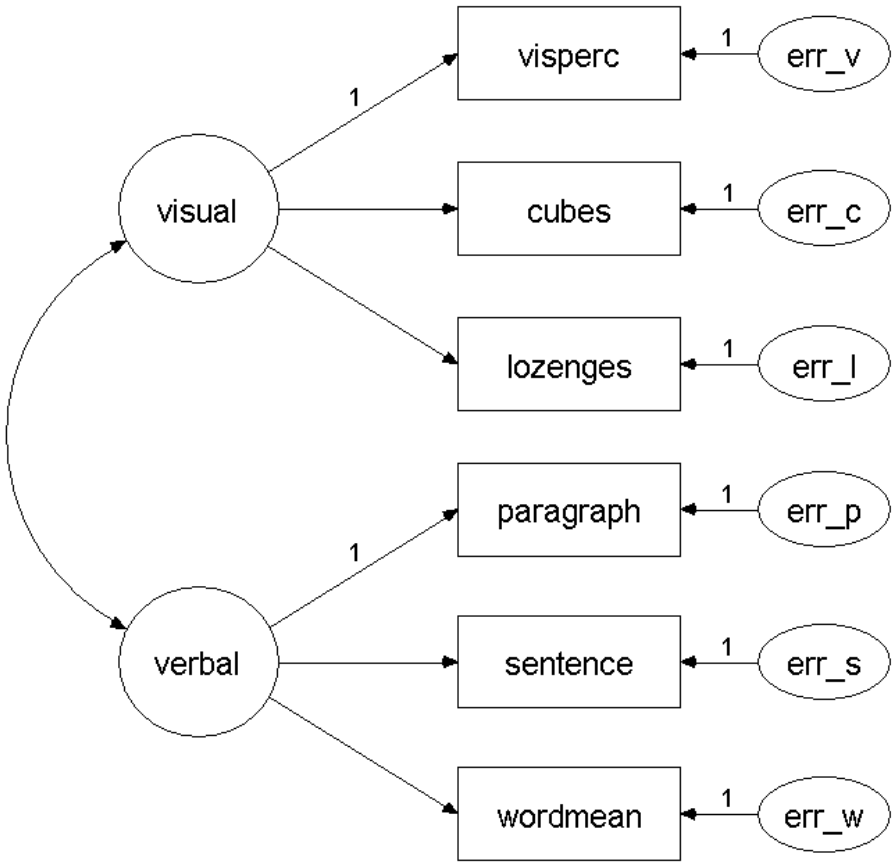
Categorical observed variables may be dichotomous or ordered polytymous (i.e., ordered categorical outcomes of more than two levels), but nominal level observed variables with more than two categories may not be used in the analysis as outcome variables.

Sample size requirements are somewhat more stringent than for continuous variables; typically you want a minimum of 200 cases (preferably more) to perform any analysis with categorical outcome variables.

Keeping these considerations in mind, Mplus provides a convenient mechanism to perform an exploratory factor analysis of dichotomous and ordered categorical responses. Since many exploratory factor analyses are performed on Likert scale items that contain ordered categories, Mplus is a useful tool for the exploration of the factor structure of these instruments.

Section 5: Confirmatory Factor Analysis and Structural Equation Models

The examples in the preceding section demonstrate how you can use Mplus to fit exploratory factor analysis models to the Grant-White database. What if you had an a priori hypothesis that the visual perception, cubes, and lozenges variables belonged to a single factor whereas the paragraph, sentence, and word meaning variables belonged to a second factor? The diagram shown below illustrates the model visually.



You can test this hypothesized factor structure using *confirmatory factor analysis*, as shown in the next section.

5.1. Confirmatory Factor Analysis with Continuous Variables

To set up a confirmatory factor analysis, return to the first exploratory factor analysis example of Section 4. You may use the same Mplus syntax as that appearing in example 1 of Section 4, with the following modifications to the ending section of the command file:

ANALYSIS: TYPE = *general* ;

MODEL:
visual BY visperc@1 cubes lozenges ;
verbal BY paragrap@1 sentence wordmean ;
visual WITH verbal ;

OUTPUT: *standardized sampstat* ;

The ***general*** analysis type tells Mplus that you are fitting a general structural equation model rather than specific model such as an exploratory factor analysis. The model is general in the sense that you must define what parameters are estimated; all other parameters are assumed to be fixed. In the exploratory factor analysis context, Mplus already knows the specifics of that model, so specifying the model is handled automatically by Mplus. By contrast, in the confirmatory factor analysis and structural equation modeling context each hypothesized model is unique, so you must tell Mplus how the model is constructed. The **MODEL** command allows you to specify the parameters of your model.

The first line of the **MODEL** command shown above defines a latent factor called *visual*. The **BY** keyword (an abbreviation for "measured by") is used to define the latent variables; the latent variable name appears on the left-hand side of the **BY** keyword whereas the measured variables appear on the right-hand side of the **BY** keyword. It has three observed indicator variables: *visperc*, *cubes*, and *lozenges*. Similarly, in the second line of the **MODEL** command a latent factor called *verbal* has three indicators: *paragrap*, *sentence*, and *wordmean*. The third line of **MODEL** command uses the **WITH** keyword to correlate the *visual* latent factor with the *verbal* latent factor (note: this model is the same as AMOS Example 20-2r).

The *visperc* and *paragrap* variables are each followed by **@1**. The **@** sign tells Mplus to fix the factor loading (regression weight) of the *visual-visperc* relationship to the value that follows the **@**, 1.00. Similarly, the *verbal-paragrap* relationship is also fixed to 1.00. The reason you fix these two parameters is to provide a scale for the visual and verbal latent variables' variances. If you ever need to supply starting values for a particular parameter in Mplus, you can specify its number after an asterisk, like this: **sentence*.5**. Omitting the asterisks when you do not specify starting values is the default. Note that each variable is separated from the other variables in the analysis by at least one space.

Finally, the **OUTPUT** command contains an added keyword, ***standardized***. This option instructs Mplus to output standardized parameter estimate values in addition to the default unstandardized values. Selected output from the analysis appears below.

```
Grant-White School: Confirmatory Factor Analysis

SUMMARY OF ANALYSIS

Number of groups                                 1
Number of observations                         145

Number of y-variables                            6
Number of x-variables                            0
Number of continuous latent variables            2

Observed variables in the analysis
   VISPERC     CUBES       LOZENGES    PARAGRAP    SENTENCE
WORDMEAN

Continuous latent variables in the analysis
   VISUAL      VERBAL
```

The summary of analysis information tells you that there are six continuous observed variables in the analysis and two latent factors, *visual* and *verbal*. Mplus then displays the input covariance matrix generated from the six observed variables:

```
     SAMPLE STATISTICS


           Covariances/Correlations/Residual Correlations
              VISPERC       CUBES         LOZENGES      PARAGRAP
SENTENCE
              ________      ________      ________      ________
________
 VISPERC       47.801
 CUBES         10.012        19.758
 LOZENGES      25.798        15.417        69.172
 PARAGRAP       7.973         3.421         9.207        11.393
 SENTENCE       9.936         3.296        11.092        11.277
21.616
 WORDMEAN      17.425         6.876        22.954        19.167
25.321

           Covariances/Correlations/Residual Correlations
              WORDMEAN
              ________
 WORDMEAN      63.163
```

Mplus next reports the results of fitting the hypothesized model to the sample data.

```
THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                      3.663
          Degrees of Freedom             8
          P-Value                   .8861

Loglikelihood

          H0 Value              -2575.128
          H1 Value              -2573.297

Information Criteria

          Number of Free Parameters             13
          Akaike (AIC)                    5176.256
          Bayesian (BIC)                  5214.954
```

```
          Sample-Size Adjusted BIC        5173.817
            (n* = (n + 2) / 24)

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                        .000
          90 Percent C.I.                 .000   .046
          Probability RMSEA <= .05        .957
```

As was the case for the exploratory factor analysis of these data, Mplus reports the chi-square goodness-of-fit test and the RMSEA descriptive model fit statistic. The chi-square test of model fit is not significant and the RMSEA value is well below the value of .06 recommended by Hu and Bentler (1999) as an upper boundary, so you can conclude that the proposed model fits the data well. Mplus also reports the *Akaike Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC). These are descriptive indexes of model fit that you can use to compare the goodness of model fit of two or more competing models. Smaller values indicate better model fit.

Mplus also outputs the unstandardized coefficients (*Estimates* in the output), the standard errors (abbreviated *S.E.* in the output), the estimates divided by their respective standard errors (*Est./S.E.*), and two standardized coefficients for each estimated parameter in the model (*Std* and *StdYX*). The estimate divided by the standard error tests the null hypothesis that the parameter estimate is zero in the population from which you drew your sample. An unstandardized estimate divided by its standard error may be evaluated as a *Z* statistic, so values that exceed +1.96 or fall below -1.96 are significant below $p = .05$.

```
MODEL RESULTS

                   Estimates     S.E.  Est./S.E.    Std     StdYX

 VISUAL   BY
    VISPERC            1.000     .000       .000    4.358     .632
    CUBES               .542     .116      4.658    2.360     .533
    LOZENGES           1.392     .272      5.112    6.064     .732

 VERBAL   BY
    PARAGRAP           1.000     .000       .000    2.920     .868
    SENTENCE           1.309     .115     11.352    3.821     .825
    WORDMEAN           2.247     .197     11.402    6.560     .828

 VISUAL   WITH
    VERBAL             6.784    1.720      3.943     .533     .533
```

In this example, each of the estimated parameters has an estimate to standard error ratio greater than +1.96, so each factor loading is statistically significant, as well as the correlation between the *visual* and *verbal* latent factors ($Z = 3.943$). The variance components of the two factors, shown in the output appearing below, are also statistically significant, indicating that the amount of variance accounted for by each factor is significantly different from zero.

Each unstandardized estimate represents the amount of change in the outcome variable as a function of a single unit change in the variable causing it. In this example, you assume that the latent variables, in addition to some measurement error (shown below), are responsible for the scores on the six observed variables. For instance, for each single unit change in the *verbal* latent factor, *sentence* scores increase by 1.309 units.

Different measures often have different scales, so you will often find it useful to examine the standardized coefficients when you want to compare the relative strength of associations across observed variables that are measured on different scales. Mplus provides two standardized coefficients. The first, labeled *Std* on the output, standardizes using the latent variables' variances whereas the second type of standardized coefficient, *StdYX*, standardizes based on latent and observed variables' variances. This standardized coefficient represents the amount of change in an outcome variable per standard deviation unit of a predictor variable. In this output, you can see clearly that the standardized coefficients of *paragrap*, *sentence*, and *wordmean* are larger than those of *visperc*, *cubes*, and *lozenges*. This finding suggests that the *verbal* latent factor does a better job at explaining the shared variance among *paragrap*, *sentence*, and *wordmean* than does the *visual* latent factor for its three indicator variables, *visperc*, *cubes*, and *lozenges*.

This assertion is corroborated by the residual variances output by Mplus. The standardized coefficients for the first three indicators are larger than those for the remaining three indicators.

```
Residual Variances
Grant-White School: Confirmatory Factor Analysis

                  Estimates     S.E.  Est./S.E.    Std     StdYX

    VISPERC          28.485    4.739      6.011   28.485     .600
    CUBES            14.050    1.978      7.105   14.050     .716
    LOZENGES         31.933    7.269      4.393   31.933     .465
    PARAGRAP          2.791     .584      4.775    2.791     .247
    SENTENCE          6.869    1.164      5.900    6.869     .320
    WORDMEAN         19.695    3.385      5.819   19.695     .314

 Variances
    VISUAL           18.989    5.582      3.402    1.000    1.000
    VERBAL            8.525    1.376      6.196    1.000    1.000

R-SQUARE

    Observed
    Variable  R-Square

    VISPERC       .400
    CUBES         .284
    LOZENGES      .535
    PARAGRAP      .753
    SENTENCE      .680
    WORDMEAN      .686
```

Finally, the r-square output illustrates that only modest amounts of variance are accounted for in the first three indicators whereas much larger amounts of variance are accounted for in the final three indicators. As is the case with exploratory factor analysis of continuous outcome variables, you may want to use the ***mlm*** or ***mlmv*** estimators in lieu of the default ***ml*** estimator if your input data are not distributed joint multivariate normal by using the **ESTIMATOR =** option on the **ANALYSIS** command. The ***mlm*** option provides a mean-adjusted chi-square model test statistic whereas the ***mlmv*** option produces a mean and variance adjusted chi-square test of model fit; both options also induce Mplus to produce robust standard errors displayed in the model results table that are used to compute *Z* tests of significance for individual parameter estimates. An added advantage of the ***mlm*** option is that its chi-square test and standard errors are equivalent to those produced by EQS in its **ML;ROBUST** method. Muthén and Muthén have placed formulas on their Web site that allow you to use ***mlm***-produced chi-square values in nested model comparisons.

5.2. Handling Missing Data

It is often the case that you have missing data in the context of confirmatory factor analysis and structural equation modeling. Using Mplus, you can employ the optimal Full Information Maximum Likelihood (FIML) approach to handling missing data that was described above in the section Exploratory Factor Analysis with Missing Data in Section 4. Consider once again the same modified database, grant-missing.dat, containing incomplete cases that was used in the earlier exploratory factor analysis with missing data. As in the previous example, define the missing value code to be -9 for all variables using the **MISSING** subcommand in the **VARIABLE** command, copy the **MODEL** syntax from the previous confirmatory factor analysis example into the Mplus input window, and then modify the **ANALYSIS** command so that it reads as follows:

ANALYSIS:
TYPE = *general missing h1* ;

The ***missing*** keyword alerts Mplus to activate the FIML missing data handling feature. The additional ***h1*** keyword tells Mplus to output the chi-square goodness-of-fit test in addition to the typical summary statistics, missing data pattern information, parameter estimates, and standard errors obtained in an analysis. Mplus requires that you specify the ***h1*** keyword because large models with many missing data patterns can take a long time to converge. If this describes your situation, you may want to omit the ***h1*** option on the **TYPE =** line to verify that you have specified your model correctly before invoking the ***h1*** option to produce the chi-square test of model fit. If you elect to remove the ***h1*** option from the **ANALYSIS TYPE =** command, be sure to omit the ***sampstat*** option from the **OUTPUT** line, as well. If ***sampstat*** is included on the **OUTPUT** line, Mplus automatically assumes the ***h1*****ANALYSIS** option and computes the chi-square test of model fit, even if ***h1*** is not included on the **ANALYSIS TYPE =** line.

The command syntax for the missing data model is thus:

TITLE: Grant-White School: CFA with missing data

DATA: FILE IS U:\Projects\Documentation\Mplus\grant-missing.dat ;

VARIABLE:

NAMES ARE visperc
cubes
lozenges
paragrap
sentence
wordmean
gender ;

USEVARIABLES ARE visperc - wordmean ;

MISSING ARE *all* (-9) ;

ANALYSIS: TYPE = *general missing h1* ;

MODEL:
visual BY visperc@1 cubes lozenges ;
verbal BY paragrap@1 sentence wordmean ;
visual WITH verbal ;

OUTPUT: *standardized sampstat* ;

The chi-square test of model fit for the confirmatory factor analysis with missing data shows that the hypothesized model fit the data well:

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                      2.777
          Degrees of Freedom             8
          P-Value                   .9476

Loglikelihood

          H0 Value              -2376.312
          H1 Value              -2374.923

Information Criteria

          Number of Free Parameters             19
          Akaike (AIC)                    4790.623
          Bayesian (BIC)                  4847.181
          Sample-Size Adjusted BIC        4787.058
```

```
            (n* = (n + 2) / 24)

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                    .000
          90 Percent C.I.             .000   .011
          Probability RMSEA <= .05    .982
```

The Mplus parameter estimates, standard errors, and standardized parameter estimates are similar to those found in the preceding confirmatory factor analysis example. The only substantial difference is the inclusion of an additional section that contains means and intercepts for the latent factors and observed variables. These means and intercepts are required to be estimated by the FIML missing data handling procedure, but are otherwise not a part of the tested model.

MODEL RESULTS

	Estimates	S.E.	Est./S.E.	Std	StdYX
VISUAL BY					
VISPERC	1.000	.000	.000	4.377	.635
CUBES	.469	.127	3.679	2.051	.473
LOZENGES	1.373	.294	4.673	6.010	.725
VERBAL BY					
PARAGRAP	1.000	.000	.000	2.914	.866
SENTENCE	1.187	.114	10.376	3.460	.821
WORDMEAN	2.247	.206	10.888	6.547	.827
VISUAL WITH					
VERBAL	7.014	1.800	3.896	.550	.550
Residual Variances					
VISPERC	28.354	5.037	5.629	28.354	.597
CUBES	14.589	2.340	6.234	14.589	.776
LOZENGES	32.642	7.938	4.112	32.642	.475
PARAGRAP	2.824	.627	4.507	2.824	.250
SENTENCE	5.781	1.070	5.401	5.781	.326
WORDMEAN	19.872	3.578	5.554	19.872	.317
Variances					
VISUAL	19.158	5.859	3.270	1.000	1.000
VERBAL	8.493	1.393	6.099	1.000	1.000
Intercepts					
VISPERC	29.579	.572	51.673	29.579	4.291
CUBES	24.616	.421	58.431	24.616	5.678
LOZENGES	15.965	.689	23.184	15.965	1.925
PARAGRAP	9.952	.279	35.620	9.952	2.958
SENTENCE	19.054	.366	52.057	19.054	4.522
WORDMEAN	17.283	.658	26.274	17.283	2.182

Finally, Mplus produces the r-square values for the observed variables. Once again, these are similar to those obtained from the original database with complete cases.

```
R-SQUARE

    Observed
    Variable  R-Square

    VISPERC        .403
    CUBES          .224
    LOZENGES       .525
    PARAGRAP       .750
    SENTENCE       .674
    WORDMEAN       .683
```

If you elect to use Mplus's FIML approach to handling missing data, be aware that the only available estimator is the maximum likelihood option, ***ml***. If you suspect that your data are non-normally distributed, remember that the chi-square test of model fit may be affected by the non-normality problem. Depending on the severity of the non-normality problem and the amount of missing data you have, you may want to explore other ways of handling the missing data problem prior to performing analyses using Mplus; see General FAQ: Handling missing or incomplete data or schedule an appointment with a consultant for details on these alternative methods of handling missing data.

5.3. Confirmatory Factor Analysis with Categorical Outcomes

Confirmatory factor analysis with dichotomous and polytomous categorical outcomes, or confirmatory factor analysis with mixed categorical and continuous outcomes is also possible using Mplus. Recall the grantcat.dat database used in the example Exploratory Factor Analysis with Categorical Outcomes in Section 4. Using the same database that replaces the six continuous observed variables with dichotomous variables, you can use the confirmatory factor analysis syntax from the example Confirmatory Factor Analysis With Continuous Variables with the following modifications.

First, add the **CATEGORICAL ARE vizcat ... wordcat ;** statement to the **DATA** command. Mplus will now treat the six observed variables as categorical in the analysis. The entire command syntax is shown here.

```
TITLE:     Grant-White School: CFA with categorical outcomes

DATA:      FILE IS U:\Projects\Documentation\Mplus\grantcat.dat ;

VARIABLE:

  NAMES ARE  viscat
             cubescat
```

lozcat
paracat
sentcat
wordcat ;

USEVARIABLES ARE viscat - wordcat ;

CATEGORICAL ARE viscat - wordcat ;

ANALYSIS: TYPE = *general* ;

MODEL:
visual BY viscat@1 cubescat lozcat ;
verbal BY paracat@1 sentcat wordcat ;
visual WITH verbal ;

OUTPUT: *sampstat standardized* ;

Selected results from the analysis appear below.

```
Chi-Square Test of Model Fit

          Value                      7.463*
          Degrees of Freedom             6**
          P-Value                    .2800

*  The chi-square value for MLM, MLMV, WLSM and WLSMV cannot be used
for
   chi-square difference tests.

** The degrees of freedom for MLMV and WLSMV are estimated according
to
   formula 109 (page 281) in the Mplus User's Guide.
```

The chi-square test of model fit is once again non-significant, suggesting that the specified model fits the data adequately. The default estimator for models that contain categorical outcomes is the mean and variance-adjusted weighted least-squares method, ***wlsmv***. Optional estimators you may choose are weighted least-squares (***wls***) and mean-adjusted weighted least-squares (***wlsm***). As is the case in the exploratory factor analysis of categorical data example, there are no descriptive model fit statistics produced by Mplus when it analyzes categorical outcomes. Mplus also produces a note alerting you not to use the MLMV, WLSM, and WLSMV chi-square values in nested model comparisons (the warning about the MLM chi-square is not relevant as long as you use the formulas shown on the Mplus Web site for nested model MLM chi-square comparisons when you use the MLM estimator in the analysis of continuous outcomes). You should not use the MLM estimator for the analysis of intrinsically categorical outcome variables.

Mplus then outputs the model results:

```
MODEL RESULTS

                    Estimates     S.E.  Est./S.E.    Std     StdYX

 VISUAL   BY
    VISCAT             1.000     .000       .000    .729     .729
    CUBESCAT           .831      .212      3.922    .606     .606
    LOZCAT             .975      .230      4.248    .710     .710

 VERBAL   BY
    PARACAT            1.000     .000       .000    .814     .814
    SENTCAT            1.058     .134      7.920    .861     .861
    WORDCAT            1.038     .127      8.154    .844     .844

 VISUAL   WITH
    VERBAL             .397      .087      4.592    .670     .670

 Variances
    VISUAL             .531      .162      3.273   1.000    1.000
    VERBAL             .662      .117      5.661   1.000    1.000

 Thresholds
    VISCAT$1           .095      .104       .913    .095     .095
    CUBESCAT$1         .271      .105      2.571    .271     .271
    LOZCAT$1          -.043      .104      -.415   -.043    -.043
    PARACAT$1          .009      .104       .083    .009     .009
    SENTCAT$1          .183      .105      1.743    .183     .183
    WORDCAT$1          .043      .104       .415    .043     .043
```

This output is similar to that of a confirmatory factor analysis with continuous outcomes, with one notable exception: Mplus now produces *threshold* information for each categorical variable. A threshold is the expected value of the latent variable or factor at which an individual transitions from a value of 0 to a value of 1.00 on the categorical outcome variable when the continuous underlying latent variable's score is zero. There are only two categorical values for each outcome variable, so there is only one threshold per variable. For any categorical outcome variable with *K* levels, Mplus will output *K-1* threshold values. For example, a five-point Likert scale item would contain four threshold values. The first threshold would represent the expected value at which an individual would be most likely to transition from a value of 0 to a value of 1.00 on the Likert outcome variable. The second threshold would represent the expected value at which an individual would be most likely to transition from a value of 1.00 to a value of 2.00 on the outcome variable, and so on through the fourth threshold, which represents the expected value at which an individual would transition from 3.00 to 4.00 on the outcome variable.

Finally, Mplus produces the r-square table output. The r-square values are computed for the continuous latent variables underlying the categorical outcome variables rather than the actual outcome variables as is the case in analyses that contain continuous outcome variables. Note that

the r-square values for the categorical outcomes cannot be interpreted as the proportion of variance explained as is the case in the analysis of continuous outcomes. Therefore, examining the sign and significance of the estimated coefficients shown in the model results table above is generally more informative than interpreting r-square values.

```
R-SQUARE

    Observed  Residual
    Variable  Variance  R-Square

    VISCAT        .469      .531
    CUBESCAT      .633      .367
    LOZCAT        .495      .505
    PARACAT       .338      .662
    SENTCAT       .259      .741
    WORDCAT       .287      .713
```

The r-square table's residual variance output is, however, useful for computing expected probabilities. You can use threshold and coefficient information shown above with the residual variance information from the r-square table to compute the expected probability of case having a value of 0 or 1.00. Consider the following formula for computing the conditional probability of a Y = 0 response given the factor eta.:

P(Y_ij = 0|eta_ij) = F[(tau_j - lambda_j*eta_i)*(1/square root of theta_jj)]

where:

eta is the factor's value
F is the cumulative normal distribution function
tau is the measured item's threshold
lambda is the item's factor loading
theta is the residual variance of the measured item

Suppose you want to obtain the estimated probability for *sentcat* = 0 at *eta* = 0. Using the formula, shown above, you can compute this value:

P(Y_ij|eta_ij) = F[(.183 - 0)*(1/square root of .259)]
= F[.183*1.9649437]

= F[.3595847]

You can look up the value of .3595847 in a *Z* table in a statistics textbook, or you can supply the computed value of .3595847 to the **PROBNORM** function in SAS to obtain the correct probability value. The **PROBNORM** function returns the value from a cumulative normal distribution for the inputted value. A simple SAS program such as the one shown below enables you to obtain the final expected probability value of .64.

DATA one ;

```
    p = PROBNORM(.3595847) ;
RUN ;

PROC PRINT DATA = one ;
RUN ;
```

You may substitute other values of eta and lambda to obtain different expected probability values. In general, the same cautions and limitations that were discussed above in the section Exploratory Factor Analysis with Categorical Variables section also apply to the analysis of categorical outcomes in the confirmatory factor analysis and structural equation modeling contexts. In addition, the following point is worth considering:

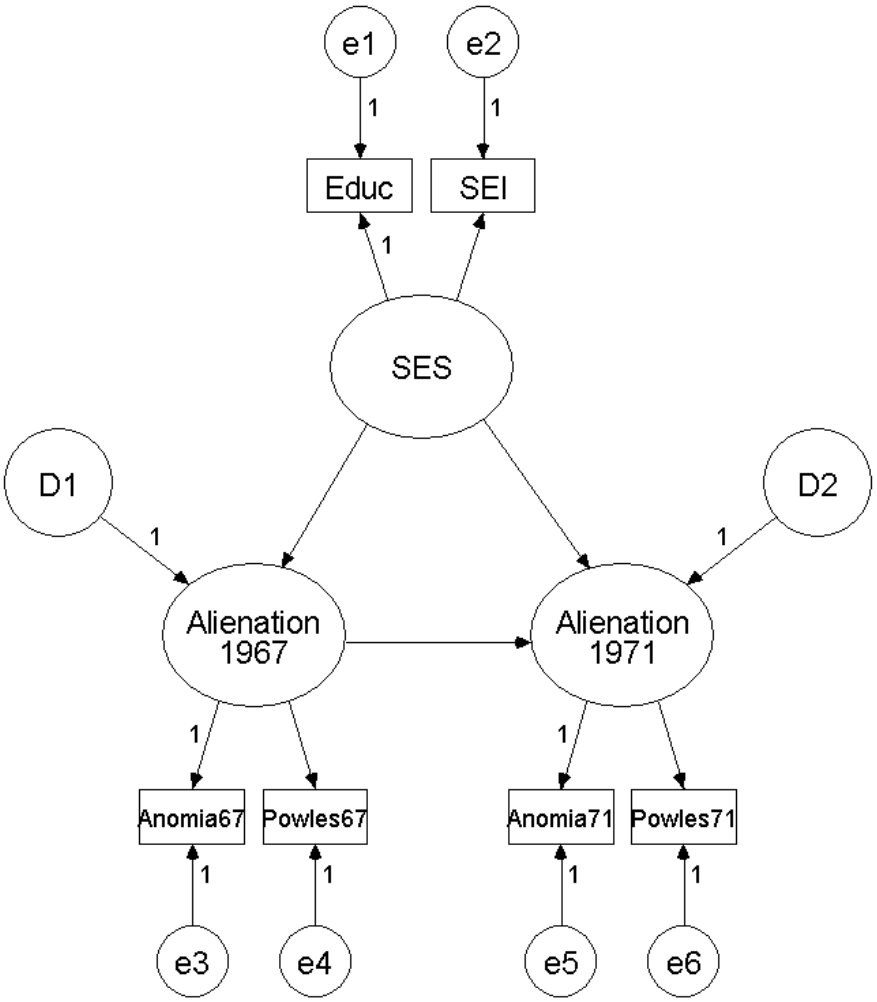
Do not list independent (exogenous) categorical variables in the **CATEGORICAL** statement. Instead, create dummy variables (i.e., variables with values of 0 and 1 representing group membership status) and include them in the model as predictors, *or* create a multiple group analysis based upon category membership as described in the Multiple Group Analysis section of this document.

5.4. Structural Equation Modeling with Continuous Outcomes

In addition to exploratory and confirmatory factor analysis, you may use Mplus to fit structural equation models that feature causal relationships among latent variables. A ubiquitous example of a structural equation model is that of the impact of socioeconomic status (SES) on alienation in 1967 and 1971. A study conducted by Wheaton, Muthén, Alwin, and Summers (1977) fit several structural equation models to a database of 932 research participants. The database contained the following observed, continuous variables:

Educ - Education level
SEI - Socioeconomic index
Anomia67 - Anomie in 1967
Anomia71 - Anomie in 1971
Powles67 - Powerlessness in 1967
Powles71 - Powerlessness in 1971

One of the fitted structural equation models features a latent factor, *SES*, that influences *Educ* and *SEI* scores. The *SES* latent variable in turn influences two additional latent variables: *Alien67* and *Alien71*. *Alien67* represents self-perceived alienation in 1967 and it influences responses on the anomie and powerlessness variables measured in 1967. Similarly, *Alien71* represents self-perceived alienation in 1971 and it influences responses on the anomie and powerlessness variables measured in 1971. *SES* influences both *Alien67* and *Alien71* and *Alien67* also influences *Alien71*. A figure of this model appears below, as well as on page 31 of our AMOS tutorial.



The dataset, wheaton-generated.dat, used in the AMOS tutorial example, is also used here in the analysis that follows:

TITLE: Wheaton et al. Example 1: Full SEM

DATA:
FILE IS U:\Projects\Documentation\Mplus\wheaton-generated.dat ;

VARIABLE:

```
NAMES ARE  educ
           sei
           anomia67
           powles67
           anomia71
           powles71 ;

USEVARIABLES ARE    educ - powles71 ;

ANALYSIS:  TYPE = general ;

MODEL:     ses BY educ@1 sei ;
           alien67 BY anomia67@1 powles67 ;
           alien71 BY anomia71@1 powles71 ;

           alien67 ON ses ;
           alien71 ON ses alien67 ;

OUTPUT:    standardized sampstat ;
```

The syntax for this analysis is similar to that of the confirmatory factor analysis example shown in section 5.1 above. The only noteworthy difference is the use of the **ON** keyword in the **MODEL** command to specify the regression relationships among the latent variables; the **WITH** keyword is used to specify correlations or covariances among variables. In this example, the *alien67* latent variable is regressed on the *SES* latent variable. Similarly, the *alien71* latent variable is regressed on both the *SES* and *alien67* latent variables. The model fit statistics appear below:

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                     76.184
          Degrees of Freedom            6
          P-Value                  .0000
```

...output deleted...

RMSEA (Root Mean Square Error Of Approximation)

```
          Estimate                   .112
          90 Percent C.I.            .090   .135
          Probability RMSEA <= .05   .000
```

The statistically significant chi-square test of absolute model fit coupled with the poor RMSEA fit statistic value suggest that this model may need some modification before it fits the data well.

The model fit and r-square tables appear below.

```
MODEL RESULTS

                   Estimates     S.E.  Est./S.E.    Std     StdYX

 SES      BY
    EDUC               1.000     .000       .000    2.420     .784
    SEI                 .592     .043     13.694    1.433     .683

 ALIEN67  BY
    ANOMIA67           1.000     .000       .000    2.929     .816
    POWLES67            .823     .038     21.734    2.409     .793

 ALIEN71  BY
    ANOMIA71           1.000     .000       .000    2.989     .843
    POWLES71            .825     .039     21.305    2.465     .778

 ALIEN67  ON
    SES                -.759     .062    -12.235    -.627    -.627

 ALIEN71  ON
    SES                -.172     .064     -2.689    -.139    -.139
    ALIEN67             .710     .056     12.609     .696     .696

 Residual Variances
    EDUC               3.677     .416      8.839    3.677     .386
    SEI                2.345     .172     13.651    2.345     .533
    ANOMIA67           4.301     .364     11.807    4.301     .334
    POWLES67           3.422     .260     13.150    3.422     .371
    ANOMIA71           3.637     .369      9.849    3.637     .289
    POWLES71           3.951     .289     13.681    3.951     .394
    ALIEN67            5.201     .495     10.516     .606     .606
    ALIEN71            3.352     .382      8.781     .375     .375

 Variances
    SES                5.854     .557     10.515    1.000    1.000


R-SQUARE

    Observed
    Variable  R-Square

    EDUC          .614
    SEI           .467
    ANOMIA67      .666
    POWLES67      .629
    ANOMIA71      .711
    POWLES71      .606
```

```
 Latent
Variable  R-Square

ALIEN67       .394
ALIEN71       .625
```

There are several noteworthy features of these tables. First, the model results table contains residual variance estimates for the *alien67* and *alien71* latent variables. These variables are predicted by the *SES* latent variable, so it makes sense that the residual or unexplained variance is due to factors other than *SES* in the model. Because *SES* is not predicted by any other variables, its variance is estimated independently and is shown in the *Variances* section of the model results table. The path coefficients from *SES* to *alien67*, from *SES* to *alien71*, and from *alien67* to *alien71* and their associated standard errors, tests of significance, and standardized coefficients also appear in the same table.

The r-square table contains r-square values for each of the predicted latent variables, *alien67* and *alien71*, as well as the observed variables. Taken as a whole, these results suggest that the model is capturing the observed variables' variances fairly well, though the prediction of alienation in 1967 is somewhat weak as is the variance accounted for in the *SEI* variable. The model may be modified, however. When all variables are continuous, Mplus can print modification indices that can provide an empirical basis to aid your decision to free additional paths, means, intercepts, or variance components to be estimated in your model. A *modification index* provides the expected drop in model fit chi-square if a parameter that is currently not free is in fact allowed to be estimated. As always, theory should be your first guide in the decision to modify your model. To request modification indices, add the following keywords to the **OUTPUT** line:

modindices (4)

The number shown in the parentheses is the amount of chi-square reduction necessary for Mplus to print any given modification index. The critical chi-square statistic is 3.84 for 1 degree of freedom at $p = .05$, so this example sets the cutoff to print modification indices at 4.00. If you do not specify a cutoff value, Mplus supplies 10.00 as the default value. The modification indices from this model appear below.

```
MODEL MODIFICATION INDICES

Minimum M.I. value for printing the modification index     4.000

                          M.I.     E.P.C.  Std E.P.C.  StdYX E.P.C.

WITH Statements

POWLES67 WITH EDUC         8.381     -.574      -.574        -.061
ANOMIA71 WITH EDUC         5.626      .533       .533         .049
ANOMIA71 WITH ANOMIA67    62.098     2.091      2.091         .164
ANOMIA71 WITH POWLES67    48.629    -1.546     -1.546        -.144
POWLES71 WITH ANOMIA67    54.470    -1.693     -1.693        -.149
POWLES71 WITH POWLES67    41.262     1.233      1.233         .128
```

In addition to the raw modification index value (*M.I.*), Mplus also prints the understandardized expected parameter change (*E.P.C.*) and standardized versions of the expected parameter change.

You can draw several immediate conclusions about the model from this table. First, the largest raw modification indicies are associated with correlating the residuals of the anomie and powerlessness variables, indicating that freeing these parameters to be estimated will result in the largest improvement in model fit. Second, the StdYX expected parameter change values are comparable with each other because they are standardized coefficients. The largest of these is the correlation of *anomia67* with *anomia71* (.164). The next largest value is the correlation of *anomia67* with *powles71* (-.149). However, you must ask yourself, "Is this modification theoretically sensible and meaningful?" about any modification you plan to undertake. You can make a case for correlating *anomia67* and *anomia71*, and *powles67* and *powles71*, because these measures are identical instruments measured on the same people at two different time points. It is conceivable that some method or instrument variance is shared across time on the same measurement instruments, but not across two distinct measurement instruments.

With this information, suppose you change the **MODEL** command to add two residual covariances via the **WITH** statement: *anomia67* with *anomie71*, and *powles67* and *powles71*. The Mplus syntax for the **MODEL** command appears below.

```
MODEL:    ses BY educ@1 sei ;
          alien67 BY anomia67@1 powles67 ;
          alien71 BY anomia71@1 powles71 ;

          alien67 ON ses ;
          alien71 ON ses alien67 ;

          anomia67 WITH anomia71 ;
          powles67 WITH powles71 ;
```

Consider the result of this modification on the model fit statistics.

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                      7.826
          Degrees of Freedom             4
          P-Value                   .0978
```

...output deleted...

```
RMSEA (Root Mean Square Error Of Approximation)

          Estimate                    .032
          90 Percent C.I.             .000   .065
          Probability RMSEA <= .05    .782
```

The chi-square test of overall model fit is not significant and the RMSEA value is well below the recommended .06 cutoff that indicates good model fit, so you conclude that your modified model fits the data well (the value of .065 for the upper bound of the 90 percent confidence interval for the RMSEA suggests that the model could be improved even more if you wished to pursue further model modifications). If you use them properly, model modification indices are a powerful tool in your analytic toolbox. The following points about model modification indices are worth considering:

Model modification should always be informed by theory.

The more modifications you perform on any given model, the more likely the results are to be sample specific (i.e., results won't generalize to new samples).

Mplus model modification indices are available when you use full information maximum likelihood (FIML) to handle missing data.

Mplus model modification indices are not available for models that contain categorical outcome variables. Instead, request ***tech2*** on the **OUTPUT** to obtain unstandardized first order derivatives that may be used as approximate guides for modification of models containing categorical outcomes.

Section 6: Advanced Models

Although Mplus can fit many standard models and it contains some useful features lacking in other SEM programs at the time of this writing (e.g., FIML missing data handling with exploratory factor analysis, modification indices with FIML missing data handling for structural equation and confirmatory factor analysis models), Mplus advanced modeling features are its most distinctive trademark. A full treatment of Mplus's advanced modeling features is beyond the scope of this tutorial, but several representative examples appear below.

6.1. Multiple Group Analysis

Recall the first confirmatory factor analysis example that features data from 145 students from the Grant-White School contained in the data file **grant.dat**. 72 of those students are male whereas 73 students are female. Suppose you decide to investigate the equality of the factor structure across the two groups of students. You can use Mplus to perform one or more *multiple group analyses* in which the parameters of your choosing are stipulated to be equal across the two groups of children. For instance, suppose you wanted to test the equality of the factor loading *and* factor variances and covariance values for males and females. The Mplus command file shown below performs this test.

```
TITLE: Grant-White School: Multiple Group CFA

DATA:
  FILE IS U:\Projects\Documentation\Mplus\grant.dat ;
```

VARIABLE:

NAMES ARE visperc
cubes
lozenges
paragrap
sentence
wordmean
gender ;

USEVARIABLES ARE visperc - wordmean ;

GROUPING = gender (1=males 2=females);

ANALYSIS: TYPE = *mgroup* ;

MODEL: visual BY visperc@1 cubes lozenges ;
verbal BY paragrap@1 sentence wordmean ;
visual (1) ;
verbal (2) ;
visual WITH verbal (3) ;

OUTPUT: *standardized sampstat* ;

Several new elements of this program are immediately apparent. First, the **GROUPING =** option for the **VARIABLE** command tells Mplus which variable in the database contains the information about group membership. For each value of the grouping variable, you supply a name that Mplus uses to define separate groups in the analysis. The **ANALYSIS** command contains an ***mgroup*** keyword that lets Mplus know you are specifying a multiple group analysis. Use the **GROUPING =** option for raw data; use the ***mgroup*ANALYSIS** keyword when you input summary data such as covariance matrices for each group. Both multiple group specification methods are included in this example for illustrative purposes, though only the **GROUPING =** option is required to run the command file because you input raw data.

By default Mplus assumes that the following specified parameter estimates are equal across multiple groups:

Factor loadings

Intercepts of continuous outcome variables

Thresholds of categorical outcome variables

That is, any model that contains factor loadings, intercepts, or thresholds will assume their estimates are identical across the multiple groups contained in the analysis. For instance, in this example the four specified factor loading values are assumed to be equal across the two groups. By contrast, parameter estimates that are not specified in the **MODEL** statement are allowed to vary across the groups. In this analysis each of the residual variances of the six observed variables will differ across the two groups.

The factor's variances and covariances are not assumed to be equal across the two groups by default, so you can equate the parameter estimate values across the two groups by using Mplus equality constraints. You can specify which parameters you want to be held equal across the two groups by assigning a number in parentheses to each set of equal parameters. For example, in the program shown above, you assigned the **visual** factor variance a value of **(1)**. Mplus thus estimates a single factor loading common to both groups.

The output from the analysis appears below.

```
SUMMARY OF ANALYSIS

Number of groups                                 2

Grant-White School: Multiple Group CFA

Number of observations
   Group MALES                                  72
   Group FEMALES                                73

Number of y-variables                            6
Number of x-variables                            0
Number of continuous latent variables            2
```

...output deleted...

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                     22.346
          Degrees of Freedom            23
          P-Value                   .4994
```

...output deleted...

```
RMSEA (Root Mean Square Error Of Approximation)

          Estimate                     .000
          90 Percent C.I.              .000   .093
```

Mplus initially reports the number of groups and the number of cases within each group. Though not shown here in the interests of conserving space, Mplus also displays the sample statistics for each group separately. Since the obtained chi-square model fit statistic (22.346) is smaller than its degrees of freedom (23) and the RMSEA is well below the cutoff value of .06, you conclude the model fits the data very well. One possible exception to this interpretation arises from the RMSEA upper bound value of .093, which exceeds the .06 cutoff recommended by Hu and Bentler (1999). Overall, however, the equality of factor loadings and factor variance-covariance structure for boys and girls appears to be a reasonable assumption.

The model results table output by Mplus features the factor loadings, factor variances, factor intercorrelations, and residuals variances for each group. Notice that the factor loadings' unstandardized regression coefficients and standard errors are identical for the boys' group and the girls' group. The variances of the *visual* and *verbal* factors are also identical across the two samples, as is the covariance between the two factors. By contrast, the residual variance estimates are not the same for the two groups because these parameters were not listed in the model specification.

```
MODEL RESULTS

                   Estimates     S.E.  Est./S.E.    Std     StdYX

Group MALES

 VISUAL   BY
    VISPERC            1.000     .000       .000    4.339     .612
    CUBES               .555     .116      4.780    2.407     .527
    LOZENGES           1.384     .263      5.262    6.005     .703

 VERBAL   BY
    PARAGRAP           1.000     .000       .000    2.865     .881
    SENTENCE           1.312     .116     11.344    3.759     .844
    WORDMEAN           2.272     .200     11.363    6.511     .825

 VISUAL   WITH
    VERBAL             6.896    1.698      4.060     .555     .555

 Residual Variances
    VISPERC           31.503    6.807      4.628   31.503     .626
    CUBES             15.047    2.926      5.142   15.047     .722
    LOZENGES          37.000    9.806      3.773   37.000     .506
    PARAGRAP           2.366     .694      3.408    2.366     .224
    SENTENCE           5.727    1.387      4.127    5.727     .288
    WORDMEAN          19.950    4.513      4.421   19.950     .320

 Variances
    VISUAL            18.827    5.476      3.438    1.000    1.000
    VERBAL             8.210    1.321      6.217    1.000    1.000

Group FEMALES

 VISUAL   BY
    VISPERC            1.000     .000       .000    4.339     .648
    CUBES               .555     .116      4.780    2.407     .558
    LOZENGES           1.384     .263      5.262    6.005     .767

 VERBAL   BY
    PARAGRAP           1.000     .000       .000    2.865     .858
    SENTENCE           1.312     .116     11.344    3.759     .796
```

```
    WORDMEAN           2.272      .200     11.363    6.511      .827

 VISUAL   WITH
    VERBAL             6.896    1.698      4.060     .555      .555

 Residual Variances
    VISPERC           26.004    5.695      4.566   26.004      .580
    CUBES             12.834    2.482      5.171   12.834      .689
    LOZENGES          25.191    7.820      3.221   25.191      .411
    PARAGRAP           2.947     .816      3.609    2.947      .264
    SENTENCE           8.164    1.795      4.549    8.164      .366
    WORDMEAN          19.614    4.749      4.130   19.614      .316

 Variances
    VISUAL            18.827    5.476      3.438    1.000    1.000
    VERBAL             8.210    1.321      6.217    1.000    1.000


R-SQUARE

Group MALES

    Observed
    Variable  R-Square

    VISPERC       .374
    CUBES         .278
    LOZENGES      .494
    PARAGRAP      .776
    SENTENCE      .712
    WORDMEAN      .680

Group FEMALES

    Observed
    Variable  R-Square

    VISPERC       .420
    CUBES         .311
    LOZENGES      .589
    PARAGRAP      .736
    SENTENCE      .634
    WORDMEAN      .684
```

It is worth noting that you can constrain parameters to be equal for a single group analysis in Mplus by assigning two or more parameters listed within the **MODEL** command a unique number, much as you did in the example shown above. It is therefore possible to impose between and within-groups constraints simultaneously using Mplus.

You can also impose equality constraints or custom model specifications within specific groups

in a multiple group analysis by referring to the group's name. For instance, if you wanted to equate the residual variances for the six variables for males only, you could modify the model statement to read as follows:

```
MODEL:      visual BY visperc@1 cubes lozenges ;
            verbal BY paragrap@1 sentence wordmean ;
            visual (1) ;
            verbal (2) ;
            visual WITH verbal (3) ;

MODEL males:    visperc - wordmean (4)  ;
```

This model constrains the residual variance values of the six observed variables for males to be equal, but the females' residual variances are allowed to remain unique for each measured variable.

For more information on multiple group analysis, including cautionary notes regarding multiple group analysis, see AMOS FAQ: Multiple group analysis.

6.2. Multilevel Models

Investigators often draw data from sources that feature a *hierarchical* or *multilevel* structure such as students nested within classrooms, patients residing in hospitals, children grouped within a family, individuals grouped within couples, etc. In recent years, specialized software such as HLM and MLWin have been developed to fit regression and related-models (e.g., ANOVA, ANCOVA, MANOVA, and MANCOVA) to such databases because many statistical software packages such as SPSS and SAS assume every observation is independent of the observations that precede and follow it (some exceptions to this general rule are the MIXED procedure in SAS and the LISREL multilevel module, both of which may be used to fit multilevel regression models). In situations where individuals are members of some type of larger aggregate or cluster (e.g., families, couples, classrooms), this *independence assumption* can be and often is violated. Violations of the independence assumption can seriously degrade the results from an analysis conducted on multilevel data.

Although specialized software products such as HLM and related programs permit multilevel regression analyses, Mplus features a latent variable-based approach to multilevel modeling that has the following benefits:

Assessment of overall model fit using the usual maximum-likelihood chi-square test statistic when cluster sizes are equal, as well as the MLM and MLMV robust estimator options when cluster sizes are not equal (the default estimator is ***mlm***).

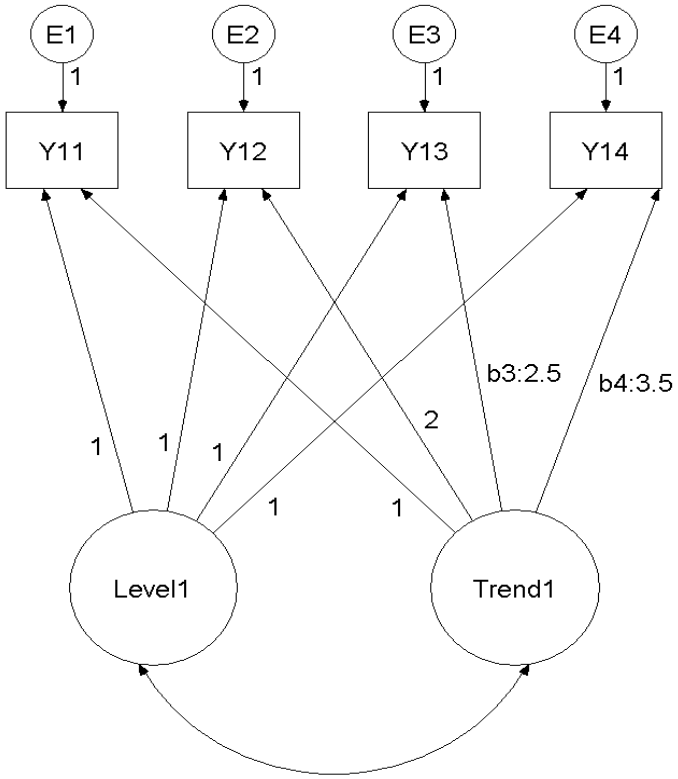
Latent variables in the analysis with the concomitant purging of measurement errors.

The construction and testing of measurement models.

Automatic sorting of the input data and construction of the appropriate between and within-groups covariance matrices used in the analysis.

Specification of parallel process models in which multiple sets of repeatedly-measured variables are analyzed, with each set having its own growth parameters.

Latent growth factors may predict other variables and may in turn be predicted by other variables in the model.



Separate model specifications are permissible for each level of the analysis.

Mplus accounts for the effect of a single clustering variable by calculating two separate covariance matrices, a between cluster matrix and a pooled within-cluster covariance matrix. Taken together, these matrices represent the total variation among the observed variables included in the model. Mplus may also be used to address the issue of cluster-sampled (i.e., non-random sampled) data using a similar mechanism. Fortunately, as noted above, you need only supply Mplus with the input data and the name of the clustering variable. Mplus handles data sorting and computation of the appropriate input matrices internally.

An example of a multilevel latent growth analysis appears below. It is based on a more complex example that can be found on the Mplus Web site. See Muthén (1997) for related examples. The data are also available for download. In this example, Y11 through Y14 are observed variables, E1 through E4 are residual variance estimates, Level1 is the random intercept, and Trend1 is a random slope variable.

In the model diagram, the level or intercept variable is linked to each observed variable via fixed coefficients of 1.00. The trend or slope latent variable's first two coefficients are fixed at 1.00 and 2.00, respectively, followed by two free parameters, *b3* and *b4*. The two free parameters have start values of 2.5 and 3.5, respectively. The level and trend are allowed to correlate. The Mplus model specification appears next:

```
TITLE: Multilevel latent growth model (based on Mplus example program)

DATA:
      FILE IS u:\projects\documentation\mplus\comp.dat;
      FORMAT is 3f8 f8.4 8f8.2 3f8 2f8.2

VARIABLE:
NAMES ARE     g1 g2 cluster g3
                    y11-y14
                    y21-y24
                    x1-x5;
 USEOBS = (x1 EQ 1 AND g1 EQ 2);
 MISSING = ALL (999);
 USEVAR =  y11-y14 ;
 CLUSTER = cluster;

DEFINE:     y11 = y11/5;
               y12 = y12/5;
               y13 = y13/5;
               y14 = y14/5;

ANALYSIS:  TYPE = twolevel;

MODEL:
                %BETWEEN%

                level1b BY y11-y14@1;
                trend1b BY y11@0 y12@1 y13*2.5 y14*3.5;

               [y11-y14@0];
               [level1b-trend1b];

                level1b WITH trend1b ;

                %WITHIN%

                level1w BY y11-y14@1;
                trend1w BY y11@0 y12@1 y13*2.5 y14*3.5;

                level1w WITH trend1w ;
```

OUTPUT: ***sampstat standardized ;***

In the interests of conserving space, this program makes use of several Mplus shortcuts. First, the **DATA** command illustrates the use of the FORTRAN **FORMAT** statement to read the variables from the large data file efficiently, as recommended by the Mplus manual. The **USEOBS** command limits the observations to the subset of cases of interest for this analysis.

The first multilevel analysis command is the **CLUSTER** command. The **CLUSTER** command identifies which variable in the database denotes group or cluster membership. In this example, the variable's name is *cluster*. Following the **CLUSTER** command is the **DEFINE** command. **DEFINE** allows you to rescale the observed variables so that Mplus is more likely to converge when it fits the multilevel model to the database (multilevel models often have more difficultly converging than single-level models).

The **ANALYSIS** command defines the type of analysis as ***twolevel***. This option tells Mplus that you are fitting a two-level model to the data. At present, Mplus can only fit multilevel models with a single clustering variable, though Mplus can fit some three-level models if you consider the third level of the model to consist of equally-spaced repeated measurements of the observed variables. As mentioned previously, you may use ***ml***, ***mlm***, or ***mlmv*** as estimator options for multilevel models. If you select the ***ml*** estimator, Mplus produces RMSEA model fit statistics in addition to the familiar chi-square test of model fit. Use the ***ml*** estimator option only if cluster sizes are equal and it is reasonable to assume joint multivariate normality of the model residuals; otherwise, use the default ***mlm*** estimator or the optional ***mlmv*** estimator.

The **MODEL** command contains the model specification statements for the between and within-cluster components of the model. The between-cluster model specification is listed under the **%BETWEEN%** subcommand. Notice that any mean and intercept structure specifications occur here; these occur at the between level only. The **%WITHIN%** subcommand then lists the model specification for the within-cluster model for individuals in the dataset.

The output from this analysis appears below, with some output deleted in the interest of conserving space. The first displayed output is the summary of data, which displays the number of clusters and the ID numbers contained within clusters of a given size. For instance, two clusters contain seven cases each. These clusters are cluster number 103 and cluster number 132.

```
SUMMARY OF DATA

     Number of clusters          50

       Size (s)    Cluster ID with Size s

          2           114
          3           136
          6           304
          7           103    132
          9           102    109
         10           305
```

```
          11          111
          14          134
          15          116     106
          16          118     138     110     105
          17          101     128
          18          133     131     122
          19          303     124     146
          20          147     137     307
          21          129     141     145
          22          144     127     142     143
          23          139     308
          24          119
          25          120     121     112     123
          26          140
          27          301     108     117
          29          135
          34          104
          35          115
          40          302
          41          309

     Average cluster size  19.609
```

Mplus also displays the intraclass correlations of the observed variables. The intraclass correlation assesses the level of variance in the observed variable that is attributable to membership in its cluster. Even small intraclass correlations suggest the need for a multilevel analysis. In this analysis, the amount of variance attributable to cluster membership ranges from 15% to 20%, suggesting that a multilevel analysis is required.

```
     Estimated Intraclass Correlations for the Y Variables

                Intraclass              Intraclass
Intraclass
     Variable  Correlation   Variable  Correlation   Variable
Correlation

     Y11            .206      Y12           .150
Y13           .167
     Y14            .165
```

The overall test of model fit is satisfactory, as is the RMSEA information.

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                      7.561*
          Degrees of Freedom             4
          P-Value                   .1087
```

The model results appear below. The results are divided by level. Mplus first outputs the results for the between-cluster portion of the model:

```
MODEL RESULTS

                    Estimates     S.E.  Est./S.E.    Std     StdYX

Between Level

 LEVEL1B  BY
    Y11                1.000     .000       .000     .687     .923
    Y12                1.000     .000       .000     .687     .914
    Y13                1.000     .000       .000     .687     .842
    Y14                1.000     .000       .000     .687     .764

 TREND1B  BY
    Y11                 .000     .000       .000     .000     .000
    Y12                1.000     .000       .000     .027     .036
    Y13                2.432     .173     14.026     .065     .080
    Y14                3.458     .256     13.519     .092     .103

 LEVEL1B  WITH
    TREND1B             .038     .011      3.369    2.077    2.077

 Residual Variances
    Y11                 .082     .031      2.668     .082     .148
    Y12                 .016     .013      1.264     .016     .029
    Y13                 .005     .010       .509     .005     .007
    Y14                 .065     .028      2.337     .065     .080

 Variances
    LEVEL1B             .472     .087      5.450    1.000    1.000
    TREND1B             .001     .003       .282    1.000    1.000

 Means
    LEVEL1B           10.557     .114     92.953   15.368   15.368
    TREND1B             .522     .046     11.427   19.561   19.561

 Intercepts
    Y11                 .000     .000       .000     .000     .000
    Y12                 .000     .000       .000     .000     .000
    Y13                 .000     .000       .000     .000     .000
    Y14                 .000     .000       .000     .000     .000
```

Mplus then displays the corresponding model results for the within-cluster level of the model:

```
Within Level

 LEVEL1W  BY
    Y11                1.000     .000       .000    1.447     .897
```

```
   Y12                1.000      .000       .000     1.447      .863
   Y13                1.000      .000       .000     1.447      .785
   Y14                1.000      .000       .000     1.447      .689

TREND1W  BY
   Y11                 .000      .000       .000      .000      .000
   Y12                1.000      .000       .000      .193      .115
   Y13                2.709      .826      3.281      .524      .284
   Y14                4.237     1.417      2.991      .820      .390

LEVEL1W  WITH
   TREND1W             .082      .033      2.466      .294      .294

Residual Variances
   Y11                 .507      .052      9.791      .507      .195
   Y12                 .516      .038     13.567      .516      .183
   Y13                 .580      .045     12.885      .580      .171
   Y14                 .943      .167      5.646      .943      .214

Variances
   LEVEL1W            2.093      .109     19.199     1.000     1.000
   TREND1W             .037      .027      1.390     1.000     1.000
```

Though this analysis produced similar findings for the between and within-cluster components of the model, this is not always the case. It is often the case that you will need different model specifications for the between versus the within-cluster sections of the model's specification.

It is also worth noting that despite the congruence between the within and the between-cluster components of this model, if you fit the model as a single level model (using the ***mlm*** estimator option), you obtain the following results:

```
MODEL RESULTS

                  Estimates     S.E.  Est./S.E.    Std     StdYX

LEVEL    BY
   Y11                1.000      .000       .000     1.606      .903
   Y12                1.000      .000       .000     1.606      .870
   Y13                1.000      .000       .000     1.606      .797
   Y14                1.000      .000       .000     1.606      .708

TREND    BY
   Y11                 .000      .000       .000      .000      .000
   Y12                1.000      .000       .000      .227      .123
   Y13                2.451      .130     18.812      .556      .276
   Y14                3.496      .195     17.901      .793      .350

LEVEL    WITH
   TREND               .124      .031      4.000      .341      .341
```

Residual Variances					
Y11	.582	.055	10.593	.582	.184
Y12	.528	.044	12.071	.528	.155
Y13	.565	.045	12.614	.565	.139
Y14	1.061	.112	9.443	1.061	.206
Variances					
LEVEL	2.580	.137	18.881	1.000	1.000
TREND	.051	.012	4.317	1.000	1.000
Means					
LEVEL	10.557	.057	184.473	6.572	6.572
TREND	.517	.032	15.958	2.280	2.280
Intercepts					
Y11	.000	.000	.000	.000	.000
Y12	.000	.000	.000	.000	.000
Y13	.000	.000	.000	.000	.000
Y14	.000	.000	.000	.000	.000

Although the chi-square model fit test for this model indicates the model fits the data well (chi-square = 3.697 with 3 DF, $p = .295$), you can see that all variance estimates are statistically significant. This finding does not take into account the non-independence of individuals who are grouped within the same cluster; it thus stands in contrast to the more appropriate multilevel model that shows a non-significant variance component for the trend latent variable on both the between and within-cluster levels.

The following notes are worth considering before you specify a multilevel model and fit it to your data using Mplus.

On occasion, you may need to supply starting values to Mplus to obtain a solution that converges. Assigning reasonable starting values to variance estimates may be helpful. Another approach that often yields satisfactory starting values is to fit a single-level model to the entire sample, ignoring clustering; take the parameter estimates from that model and supply them as the starting values for the multilevel model.

Each analysis should have at least 30 to 50 clusters.

Variables measured at the group or cluster level (e.g., family size) may only be used at that level of the analysis.

Variables measured at the individual or within-cluster level exist at both levels of the analysis and need to be considered in both the between and within-cluster model specifications.

FIML missing data handling is not available for multilevel models; missing data issues must be resolved prior to the multilevel analysis.

References

Hu, L., & Bentler, P.M. (1999). Cutoff criteria in fix indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6(1)*, 1-55.

Muthén, B. (1997). Latent variable modeling with longitudinal and and multilevel data. In A. Raftery (ed.), Sociological Methodology 1997 (pp. 453-480). Boston: Blackwell Publishers.

Muthén, B., du Toit, S.H.C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Accepted for publication in *Psychometrika*.

Muthen, L.K. and Muthen, B.O. (1998). *Mplus User's Guide*. Los Angeles: Muthen & Muthen.

Wheaton, B., Muthén, B., Alvin, D., & Summers, G. (1977). Assessing reliability and stability in panel models. In D.R. Heise (Ed.): *Sociological Methodology.* San Francisco: Jossey-Bass.