Structural Equation Modeling with Mplus

Workshop for the Early Head Start Research Consortium

Jason T. Newsom, Ph.D.
Associate Professor
Institute on Aging
School of Community Health
Portland State University

September 16-17, 2004
Kansas City

newsomj@pdx.edu
www.ioa.pdx.edu/newsom

Schedule

Thursday, Sept. 16.

8:30-9:00 Introductions and overview
9:00-11:00 Preparing data sets for Mplus, Mplus basic syntax, path analysis, indirect effects tests, confirmatory factor analysis basics
11:00-12:00 Breakout to analysis teams to work on the above topics
12:00-1:30 Lunch
1:30-3:30 More on CFAs, fit indices, model modification, full structural models, nonnormal data
3:30-5:00 Breakout to analysis teams to work on the above topics

Friday, Sept. 17

8:30-11:00, alternative estimation, missing data, longitudinal modeling issues, latent growth curve analysis
11:00-12:30 Breakout to analysis teams to work on the above topics

Table of Contents

Mplus Overview

Mplus Capabilities

Mplus is a general structural equation modeling (SEM) package capable of the commonly used analyses such as:

- confirmatory factor analysis
- path analysis
- full structural models (path analysis with latent variables—a combination of path analysis and confirmatory factor analysis)
- multi-group structural models
- latent growth curve analysis
- estimation methods for non-normal data
- missing data estimation

In addition to these standard SEM features. Mplus has special capabilities that not all SEM packages have:

- exploratory factor analysis (including EFA with dichotomous items)
- analysis of dichotomous measured variables
- analysis of ordinal measured variables
- multinomial logistic, poisson, and probit regression models
- latent class analysis (confirmatory factor analysis with categorical latent variables)
- mixture modeling (structural models with categorical latent variables)
- multilevel regression (hierarchical linear models)
- multilevel structural equation models (for hierarchically structured data)
- estimation procedures for sampling weights, clustered sampling, and stratified sampling designs

Syntax Basics

Types of Files

Like SPSS and SAS, Mplus has three basic types of files. Each of these file types is really just ASCII file format, which is sometimes convenient for emailing or pasting into another program.

Data file: contains the data. This is an ASCII or text file and the extension .dat is commonly used but not required. It should contain no character data and missing data symbols are limited.

Input file: stores syntax that specifies the model that you would like to test (.inp extension is used in Mplus 3).

Output file: contains the results for the model tested (.out extension).

Overview of Program File Sections

There are seven commonly used sections (referred to as "Commands") of Mplus input files that I will review.[1] Each command name is followed by a colon and functions as a section heading that contains one or more statements. Each individual statement is followed by a semicolon (*caution*: omitting the semicolon will lead to syntax errors).

In all examples, I will use upper case letters to distinguish Mplus commands or statements from example content. In practice, however, commands and statements can be in lower or upper case or mixed cases.

Title: This *optional* section simply gives any title which you wish to give your model. The first line of the title is printed at the top of each output page, but the title can be longer than one line.

```
TITLE:  The most important model of my life;
```

Data: This required command gives information about the data file location, its format, the type of data (e.g., raw data, covariance matrix), and the number of groups. By default the format is "free" or unspecified, but fixed format, in which columns widths are specified, can also be used. By default, Mplus assumes you are reading raw data.

```
DATA:   FILE=c:\jason\mplus\ehs1.dat;
```

Variable: The required variable section is used for giving the names of variables in the data set, selecting variables, and specifying missing data.

```
VARIABLE:  NAMES = x1 x2 x3 x4 y1 y2 y3;
```

[1] There are three additional Mplus commands: SAVEDATA, MONTECARLO, and PLOT which are less often used and I will not have time to discuss.

```
USEVARIABLES = x1 x2 x3;
MISSING = x1-y3(-5);
```

Define: This is an optional section that computes new variables. It is typically not used, but is handy if you want to transform a variable without creating a new data set. Standard symbols such as +, -, *, /, ** (for exponents) can be used (a full list is on p. 353 of the Mplus manual).

```
DEFINE: x1sq = x1**2;
```

Analysis: The analysis section gives information about the type of analysis, the estimation method, and the type of matrix that Mplus should use in the analysis. The specifications I use in the example below are the defaults—a general structural equation model, maximum likelihood estimation, and analysis using the covariance matrix are requested.

```
ANALYSIS:
     TYPE = GENERAL; ESTIMATOR = ML;
     MATRIX = COVARIANCE;
```

Model: The model section is where the user tells Mplus the variables and structure of the model to be tested. This section contains all statements that specify latent variables, causal paths, and correlations. The statements below illustrate the three basic statements in Mplus—BY, ON, and WITH.

```
MODEL:
     latent1 BY x1 x2 x3;
     y1 ON latent1 x4;
     latent1 WITH x4;
```

Output: This command requests certain information to be printed in the output file. By default, not too much is needed.

```
OUTPUT:   STANDARDIZED;
```

Data Set Preparation

Mplus reads only ASCII (text) data. This is inconvenient, but I'm going to give you some good tips and examples on how to prepare data sets for Mplus. There are several points that are helpful to keep in mind to avoid problems reading data out of your statistical package and into Mplus.

1. Remember the exact location of the data file, including the full path specification. I always keep folder names to a minimum, so that minor errors typing complex file paths don't trip me up on the FILE statement in Mplus.

2. Make sure that the list of the variables that you read out of your statistical packages matches exactly the list of variables you give Mplus in the DATA section. Always double check the order the variables were saved in with the order specified in Mplus.

3. If using free format for input, you need numerical values (asterisk or period are also acceptable) to designate missing data for all variables. This means that system missing values must be given a discrete missing value code.

4. It is good practice to open your data set to look at it to make sure it looks ok. You can open the file in any text editor, but I often just open it in Mplus.

SPSS

In SPSS, there are two ways to create the raw data files—through the menus and with syntax. My preferred method is with syntax, because a) I like to keep a record of the data files I created, 2) I can double check the variable list to verify the variables in my model are the ones I intended to use, and 3) I can copy the variable names into the Mplus program file to save time and avoid typos, incorrectly ordered variable lists, or omitted variables.

SPSS Using Menus.

1. file -> save as (specify location and filename and uncheck the "write variable names to spreadsheet" checkbox). Make sure that under "Save as type," you choose "Tab-delimited (*.dat)".

2. click the "variables" button and check the boxes next to the variables you wish to save out. Note that it is often convenient to first click the "drop all" button and then check the subset of variables that you desire, especially when working with a large data set.

3. Click "continue".

4. Click "save".

SPSS Using Syntax. The following syntax lines can be used to save the data as tab-delimited text. I use just a simple example with four variables here. The DESCRIPTIVES command issued afterwards is helpful for double checking that the N is the same as that used in Mplus, but it is optional. The MISSING=LISTWISE command is used to check the N in Mplus when listwise deletion is used (discussed later). If a DESCIPTIVES or other command is not used, an EXECUTE statement is needed following the SAVE command.

```
RECODE program TO b3p_conf (SYSMIS=-99).

MISSING VALUES program T0 b3p_conf (-99,-6 thru -1).

SAVE TRANSLATE OUTFILE='c:\jason\mplus\consult\ehs\temp.dat'
  /TYPE=TAB /MAP /REPLACE
  /KEEP=b1p_cesd b1v3pdet  b1v3pint  b1v3pneg .

DESCRIPTIVES VARS=b1p_cesd b1v3pdet  b1v3pint  b1v3pneg
   /MISSING=LISTWISE.
```

SAS Using Syntax. Use of the PUT statement on the DATA step in SAS will generate an ASCII file. The FILE statement is used to designate a location on the hard drive for the new file. The format statement at the end, (F10.6,'09'X), tells SAS to use a column width of 10 with 6 decimal places for all of the variables and to separate the variables with tabs. The latter is needed to avoid rounding. (I did not include any syntax here, but you may need to declare or recode missing values. The default period is acceptable in Mplus, but you must declare it as a missing value).

```
DATA one; SET data.ehs1;

DATA _NULL_; SET work.one;
   FILE 'c:\jason\mplus\consult\ehs\ehs1sas.dat';
   PUT (b1p_cesd b1v3pdet b1v3pint b1v3pneg) (F10.6,'09'X);
RUN;
```

Another option in SAS is to use the PROC EXPORT command, but it automatically lists variable names in the first line of the data file (which will cause problems in Mplus), so the file needs to be opened and edited to remove the names. This method also requires that a FORMAT statement be used to prevent rounding.

```
DATA one (keep=b1p_cesd b1v3pdet b1v3pint b1v3pneg); SET data.ehs1;

IF MISSING(b1p_cesd) THEN b1p_cesd=-99;
IF MISSING(b1v3pdet) THEN b1v3pdet=-99;
IF MISSING(b1v3pint) THEN b1v3pint=-99;
IF MISSING(b1v3pneg) THEN b1v3pneg=-99;

FORMAT b1p_cesd b1v3pdet b1v3pint b1v3pneg 10.6;

PROC EXPORT DATA=one OUTFILE='c:\jason\mplus\consult\ehs\temp2.dat'
DBMS=DLM REPLACE ;
run;
```

Example 1: Reading Data into Mplus

In Example 1 below, I read the data I created with the above examples into Mplus.

If you have user-defined missing values, you can identify those in Mplus with the MISSING statement in the VARIABLE section. The following are acceptable: MISSING = *; MISSING = . ; MISSING = BLANK; MISSING = varname(#); In the last example, "varname" is any variable name and # is the value in the data set that indicates missing where you can specify multiple discrete values or a range of values.

If using free format for input, as I illustrate below, you cannot use blanks to represent missing data in SPSS. An asterisk, period, or numerical value must be used.

The TYPE=BASIC command is not required, but generates some descriptive data useful for verifying that you have read the data correctly.

```
TITLE:  Example 1, Reading in raw data;

DATA:  FILE=ex1.dat;
       FORMAT=FREE;

VARIABLE:  NAMES = b1p_cesd b1v3pdet b1v3pint b1v3pneg;
             MISSING = b1p_cesd-b1v3pneg(-99,-6--1);

!  The following TYPE=BASIC command gives descriptive data
!   and is a good idea for checking to make sure the data
!   are read in correctly.

ANALYSIS:  TYPE=BASIC;
```

The path on the FILE statement can be abbreviated (e.g., FILE=ex1.dat) as long as the data file resides in the same folder as the input file. Otherwise, the full path is required (e.g., c:\jason\mplus\ehs\ex2.dat).

Overview of Structural Equation Modeling

- Structural equation models, sometimes called "covariance structural models," are a class of statistical techniques that combine elements of regression analysis and factor analysis.

- Structural equation modeling (SEM) is a general, flexible approach that encompasses or extends a number of common statistical models such as ANOVA, regression, hierarchical linear models, reliability estimation, multivariate analyses.

- The primary approach is one of model fitting: detailed models are specified and tested for fit with the obtained data.

- SEM employs a measurement orientation that posits "latent variables" representing constructs from which measurement error has been removed.

- Because relations among those latent variables are examined, more accurate estimates can be obtained

Path Analysis

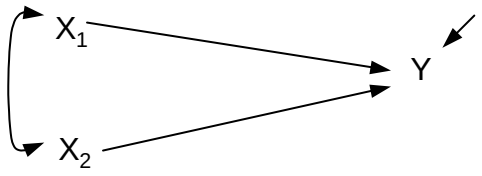
Path Diagrams

Beginning in 1918, Sewall Wright developed a system of decomposing sets of correlations into path coefficients in order to describe causal processes.

As it turns out, these path coefficients are simply regression slopes that can be derived from more complex models.

Path Diagrams

A set of conventions now exists for diagramming these models. For example, the model below represents a two-predictor regression model.

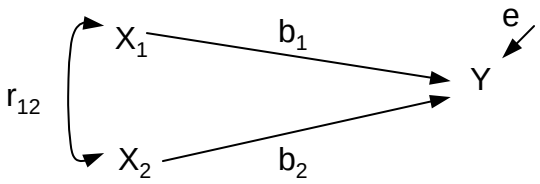


- Straight, unidirectional arrows represent hypothesized causal relations.
- Curved, double-headed arrows represent correlations among variables.
- The short arrow leading into variable Y represents the residual or error term (i.e., the variance not accounted for by X_1 and X_2). In path analysis and SEM, it is often referred to as a *disturbance* term.
-
- The causal flow is usually from left to right, sometimes said to be from "upstream" to "downstream"
- Variables on the left, not caused by any other variables are referred to as *exogenous* variables. Variables caused by other variables are referred to as *endogenous*.
- The regression weights associated with each directional arrow are usually referred to as *path coefficients*. The estimate for the relationship between variables, whether for correlations or a causal paths are generally called *parameters*

Deriving Path Estimates

Wright developed a set of rules for tracing through path diagrams that can be used to deriving path coefficients from correlations among a set of variables. Every route between the two variables is traced. The coefficients that make up each route are multiplied. If there are multiple routes that link the two variables, products for each route are added together.

Taking our original simple path diagram representing a two variable regression model, and assuming some values for the correlations between our three variables, we can derive the path coefficients using Wright's rules.



Assume that r_{12} = .50, r_{1Y} =.65, and r_{2Y} = .70.

The correlation between any two variables is a function of the possible tracing routes between those two variables. So, according to the rules $r_{1Y} = b_1 + r_{12}b_2$ and $r_{2Y} = b_2 + r_{12}b_1$.

We can then plug in the known values and solve for b_1 and b_2.

$$.65 = b_1 + .50b_2$$
$$.70 = b_2 + .50b_1$$

by rearranging, substituting, and solving for b_1 and then b_2, we get b_1 = .4 and b_2 = .5. These coefficients are actually the standardized regression coefficients, β_1 and β_2 because we started with a correlation matrix (i.e., standardized variables). Starting with a covariance matrix (the unstandardized version of a correlation matrix), we get unstandardized regression coefficients.

The disturbance term, *e*, is the amount of unaccounted for variance in *Y* and is equal to $\sqrt{1-R^2}$. We know from regression analysis that $R^2 = \beta_1^2 + \beta_2^2 + \beta_1\beta_2 r_{12}$. So, $R^2 = .4^2 + .5^2 + .5(.4)(.5) = .61$. The disturbance term then equals $e = \sqrt{1-.61} = .62$

What we have just done is decompose the correlation matrix into unique values for the coefficients that are implied by the model we specified.

Technical Note# 1 Wright's Rules of Tracing

Wright developed a method of estimating causal path coefficients by decomposing the correlations among a set of variables. He articulated a set of rules for examining a path diagram that would allow for this mathematical decomposition. The correlation of any two variables in a path diagram can be expressed as the sum of coefficients that connect the two variables.

1) No loops are allowed. In tracing from one variable to another, you cannot pass through the same variable twice.

2) No going forward and then backward. Once you have traveled along a path forward, you cannot travel backward across the path. However, going backward and then forward is possible.

3) Only one curved arrow is allowed in tracing from the first variable to the last variable.

Example 2: A Two-predictor Regression in Mplus

In Example 2 below, a regression analysis is specified in Mplus with parental negative regard (b1v3pneg) and depression (b1p_cesd) as predictors of conflict (b1p_conf). This introduces the Mplus model statements ON, for the dependent variable being regressed on an independent variable, and WITH, for correlations.

(Example 2)

```
TITLE:  Example 2, Two-predictor Regression;

DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
               FORMAT=FREE;

VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
    b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
   b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
  b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
    b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
           b3v3pneg b1p_conf b2p_conf b3p_conf;

    MISSING = program-b3p_conf(-99,-6--1);

    USEVARIABLES=b1p_cesd b1v3pneg b1p_conf;


ANALYSIS:
   TYPE = GENERAL; ESTIMATOR = ML; MATRIX = COVARIANCE;


MODEL:  b1p_conf ON b1p_cesd b1v3pneg;
        b1p_cesd WITH b1v3pneg;

OUTPUT:  STANDARDIZED;
```

Output from Example 2: Two-predictor Regression

```
Mplus VERSION 3.1
MUTHEN & MUTHEN
09/02/2004  11:07 AM

INPUT INSTRUCTIONS

  TITLE:  Example 2, Two-predictor Regression;

  DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
                FORMAT=FREE;

  VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
      b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
          b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
          b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
          b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
          b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
          b3v3pneg b1p_conf b2p_conf b3p_conf;

      MISSING = program-b3p_conf(-99,-6--1);

      USEVARIABLES=b1p_cesd b1v3pneg b1p_conf;


  ANALYSIS:
     TYPE = GENERAL; ESTIMATOR = ML; MATRIX = COVARIANCE;


  MODEL:  b1p_conf on b1p_cesd b1v3pneg;
          b1p_cesd with b1v3pneg;

  OUTPUT:  standardized;


INPUT READING TERMINATED NORMALLY

Example 2, Two-predictor Regression;

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        1593

Number of dependent variables                                    1
Number of independent variables                                  2
Number of continuous latent variables                            0

Observed dependent variables

  Continuous
   B1P_CONF

Observed independent variables
   B1P_CESD    B1V3PNEG


Estimator                                                       ML
Information matrix                                        EXPECTED
Maximum number of iterations                                  1000
Convergence criterion                                  0.500D-04
Maximum number of steepest descent iterations                   20

Input data file(s)
  c:\jason\mplus\ehs\ex2.dat

Input data format  FREE

THE MODEL ESTIMATION TERMINATED NORMALLY
```

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

Value	0.000
Degrees of Freedom	0
P-Value	0.0000

Chi-Square Test of Model Fit for the Baseline Model

Value	167.858
Degrees of Freedom	2
P-Value	0.0000

CFI/TLI

CFI	1.000
TLI	1.000

Loglikelihood

H0 Value	-8916.266
H1 Value	-8916.266

Information Criteria

Number of Free Parameters	6
Akaike (AIC)	17844.533
Bayesian (BIC)	17876.773
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	17857.712

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.000	
90 Percent C.I.	0.000	0.000
Probability RMSEA <= .05	0.000	

SRMR (Standardized Root Mean Square Residual)

Value	0.000

MODEL RESULTS

	Estimates	S.E.	Est./S.E.	Std	StdYX
B1P_CONF ON					
B1P_CESD	0.018	0.001	13.273	0.018	0.316
B1V3PNEG	0.000	0.017	0.022	0.000	0.001
B1P_CESD WITH					
B1V3PNEG	0.504	0.188	2.675	0.504	0.067
Variances					
B1P_CESD	93.256	3.304	28.222	93.256	1.000
B1V3PNEG	0.603	0.021	28.222	0.603	1.000
Residual Variances					
B1P_CONF	0.261	0.009	28.222	0.261	0.900

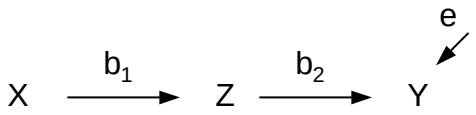
R-SQUARE

Observed Variable	R-Square
B1P_CONF	0.100

Mediation

Path analysis is especially useful for examining more complicated models that involve a causal chain of events. A variable that intervenes between a cause and an effect variable is called a *mediator* variable. A mediating relationship refers to a causal chain of events—that variable A causes variable B which, in turn, causes variable C. (Note that a *moderator* differs, because moderation refers to the statistical interaction between two predictors. That is, the effect of one predictor on the outcome depends on the value of the other predictor).

An *indirect effect* refers to a mediational effect; it is the effect of the predictor on the outcome that is due to changes in the mediator. For example, in the following diagram, the indirect effect of X on Y is the extent to which X causes Y through its effects on the mediator, Z.



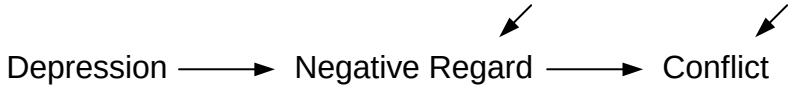
According to Wright's path tracing rules, the correlation between X and Y should equal the product of the two path coefficients, $r_{XY} = b_1 b_2$. If b_1 and b_2 are standardized coefficients, then $b_1 = r_{XZ}$ and $b2 = r_{ZY}$. Thus, r_{XY} should equal $r_{XZ} r_{ZY}$, if our path model is correct.

Knowing this fact, we could check our data to see if there is agreement between what our model implies and what our data indicate. As it turns out, this is the basis of tests of model fit in structural equation modeling. The model that we test implies certain correlations among variables given what we know about some of the other correlations. If the implied correlations are very close to the obtained correlations, then there is good fit.

In most applications, however, it is best that unstandardized information is analyzed, so covariances are used instead of correlations. A covariance can be thought of as an unstandardized correlation—the variances of the variables are not divided out. The chi-square test of model fit can be thought of as an index of the discrepancy between the implied covariance matrix and the obtained covariance matrix.

Example 3: Mediation

Using the same three variables as our previous example, we could test a mediational model, positing that depression leads to negative regard, which, in turn, leads to conflict.



In Mplus, this simply requires two regression statements. Negative regard is regressed on depression, and conflict is regressed on negative regard.

An optional section, MODEL INDIRECT, can be added to test the indirect effect. A coefficient is computed that represents the effect of depression on conflict as mediated through negative regard. The unstandardized coefficient for the indirect effect represents the change in conflict for each unit change in depression that is mediated by negative regard. The indirect coefficient is computed by multiplying the two direct path coefficients. The statement b1p_conf IND b1p_cesd refers to the indirect effect of the CES-D on conflict.

```
TITLE:  Example 3, Mediation Model;

DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
               FORMAT=FREE;

VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
    b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
   b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
  b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
    b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
            b3v3pneg b1p_conf b2p_conf b3p_conf;

    MISSING = program-b3p_conf(-99,-6--1);

    USEVARIABLES=b1p_cesd b1v3pneg b1p_conf;


ANALYSIS:
   TYPE = GENERAL; ESTIMATOR = ML; MATRIX = COVARIANCE;


MODEL:   b1v3pneg ON b1p_cesd ;
         b1p_conf ON b1v3pneg;

! The model indirect command is not required for this model
! but it produces a significance test of the indirect effect.

MODEL INDIRECT:  b1p_conf IND b1p_cesd;

OUTPUT: STANDARDIZED;
```

Output for Example 3: Mediation

(Note: I have omitted the model syntax to save paper)

```
INPUT READING TERMINATED NORMALLY

Example 3, Mediation Model;

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        1593

Number of dependent variables                                    2
Number of independent variables                                  1
Number of continuous latent variables                            0

Observed dependent variables

  Continuous
   B1V3PNEG    B1P_CONF

Observed independent variables
   B1P_CESD

Estimator                                                       ML
Information matrix                                        EXPECTED
Maximum number of iterations                                  1000
Convergence criterion                                  0.500D-04
Maximum number of steepest descent iterations                   20

Input data file(s)
  c:\jason\mplus\ehs\ex2.dat

Input data format  FREE

THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                             167.103
          Degrees of Freedom                      1
          P-Value                           0.0000



Chi-Square Test of Model Fit for the Baseline Model

          Value                             175.063
          Degrees of Freedom                      3
          P-Value                           0.0000
CFI/TLI

          CFI                                 0.035
          TLI                                -1.896

Loglikelihood
          H0 Value                       -8999.818
          H1 Value                       -8916.266

Information Criteria
          Number of Free Parameters               4
          Akaike (AIC)                   18007.636
          Bayesian (BIC)                 18029.129
          Sample-Size Adjusted BIC       18016.422
            (n* = (n + 2) / 24)

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

```
          Estimate                            0.323
          90 Percent C.I.                     0.283  0.365
          Probability RMSEA <= .05            0.000

SRMR (Standardized Root Mean Square Residual)
          Value                               0.129


MODEL RESULTS
                    Estimates     S.E.  Est./S.E.    Std     StdYX
 B1V3PNEG ON
    B1P_CESD            0.005    0.002      2.687    0.005    0.067

 B1P_CONF ON
    B1V3PNEG            0.015    0.017      0.869    0.015    0.022

 Residual Variances
    B1V3PNEG            0.600    0.021     28.222    0.600    0.995
    B1P_CONF            0.289    0.010     28.222    0.289    1.000

R-SQUARE
    Observed
    Variable  R-Square

    B1V3PNEG     0.005
    B1P_CONF     0.000
TOTAL, TOTAL INDIRECT, SPECIFIC INDIRECT, AND DIRECT EFFECTS

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

Effects from B1P_CESD to B1P_CONF

  Total               0.000    0.000      0.827    0.000    0.001
  Total indirect      0.000    0.000      0.827    0.000    0.001

  Specific indirect

    B1P_CONF
    B1V3PNEG
    B1P_CESD            0.000    0.000      0.827    0.000    0.001
```

Confirmatory Factor Analysis

Exploratory vs. Confirmatory Factor Analysis

Similarities

- Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are two statistical approaches used to examine the internal reliability of a measure.

- Both are used to investigate the theoretical constructs, or factors, that might be represented by a set of items.

- Either can assume the factors are uncorrelated, or *orthogonal*.

- Both are used to assess the quality of individual items.

- Both can be used for exploratory or confirmatory purposes.

Differences

- With EFA, researchers usually decide on the number of factors by examining output from a principal components analysis (i.e., eigenvalues are used). With CFA, the researchers must specify the number of factors a priori.

- CFA requires that a particular factor structure be specified, in which the researcher indicates which items load on which factor. EFA allows all items to load on all factors.

- CFA provides a fit of the hypothesized factor structure to the observed data.

- Researchers typically use maximum likelihood to estimate factor loadings, whereas maximum likelihood is only one of a variety of estimators used with EFA.

- CFA allows the researchers to specify correlated measurement errors, constrain loadings or factor correlations to be equal to one another, perform statistical comparisons of alternative models, test second-order factor models, and statistically compare the factor structure of two or more groups.

Latent Variables

The concept of latent variables is based on classical test theory, which assumes that any measure is a function of two variables: the true score and error variation. This assertion can be written as a formula:

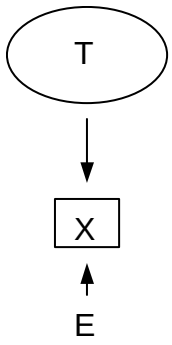
$$X = T + E$$

in which *X* represents the observed score on the measure, *T* is the person's true score, and *E* is error variation.

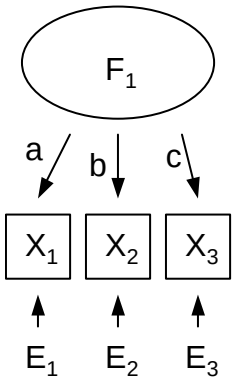
In the social sciences, we attempt measure many unobservable phenomena. The real variable or construct of interest is not precisely the one that is measured. A simple example is the measurement of an attitude, say about statistics. A response to a single items such as "Do you like statistics?" is a function of one's true attitude but also a function of other more transient factors such as the specific item wording, the respondents mood, or recent traumatic experiences with statistics. The true score, *T*, is the actual attitude, the observed score *X* is the expressed attitude on the question, and *E* is any factors that impact *X* other than *T*.

Notice that the classical test theory formula is also a regression formula. *X* is predicted by true score with some residual error remaining. In SEM, latent variables are thought to represent true scores. CFA models are visually represented in the following way:

Latent variables are represented by ellipses, and measured variables are represented by square boxes.



Deriving Factor Loadings



We can use Wright's tracing rules to derive factor loadings. The correlation between X_1 and X_2, r_{12}, should be equal to the product of *ab*, because we trace from X_1 to F_1 and back to X_2. Similarly, *bc* will equal the correlation between X_2 and X_3. To obtain the factor loadings for the above model, there are three equations:

$$r_{12} = ab$$
$$r_{23} = bc$$
$$r_{13} = ac$$

As long as we have values for r_{12}, r_{23}, and r_{13}, we can solve the equations for a, b, and c. Thus, there will be three equations and three unknowns. If we had just two variables loading on one factor, we would have two paths to estimate but only one correlation. That model is unsolvable.

If the number of unknowns is equal to the number of equations, the model is called *just identified*. If the number of unknowns is greater than the number of equations, the model is said to be *underidentified*, and there is no solution possible. An *overidentified* model is one in which there are fewer unknowns than equations. This is preferred.

Generally, the number of correlations among a set of variables can be described as:

$$\#correlations = \frac{v(v-1)}{2}$$

where *v* is the number of variables. One can determine if the model is identified by calculating whether there are more correlation elements than paths to be estimated. Thus, one formula for degrees of freedom for structural models is:

$$df = \frac{v(v-1)}{2} - p$$

where *v* is the number of measured variables in the model and *p* is the number of free parameters that need to be estimated (not including residual errors or disturbances).

Maximum Likelihood Estimation

In practice, correlations are not typically used to estimate factor loadings or path coefficients. The primary reason for this is that correlations use standardized variables and important information about the variances of the variables is lost (i.e., the variances of each of the variables is assumed to be 1). It can also be shown that for many different factor models, use of correlations can lead to erroneous results. Thus, covariances, the unstandardized version of correlations, are used.

The logic of deriving estimates of the loadings remains the same. There are a set of equations that describe the model (i.e., *structural equations*) and some known values about the relations among all of the variables (i.e., a matrix of covariances). For complicated models, the easier way to solve the set of structural equations is through a calculus-based method called *maximum likelihood* (ML). Maximum likelihood solves for the loadings by minimizing the discrepancy between the equations implied by the model and the obtained covariances. This discrepancy is mathematically described as:

$$S - \hat{\Sigma}(\theta)$$

Where *S* is the covariance matrix obtained from the data, and $\hat{\Sigma}(\theta)$ is matrix notation for a covariance matrix implied by the hypothesized model.

Certain values for the relations among the variables are implied by certain specified models. We can examine the fit of the hypothesized model to the data, by comparing the implied covariances to those obtained. The ML solution is obtained by minimizing the following (somewhat frightening) *fit function*:

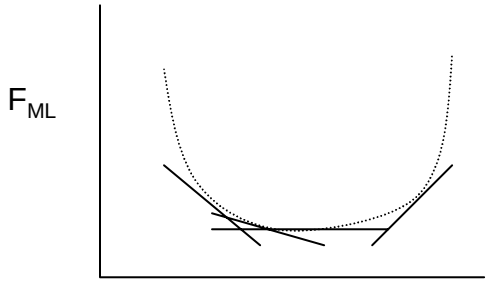
$$F_{ML} = \log\left|\hat{\Sigma}(\theta)\right| + tr\left(S\hat{\Sigma}^{-1}(\theta)\right) - \log|S| - (p+q)$$

log is the logarithm function, $\hat{\Sigma}(\theta)$ is the covariance matrix implied by the model, *S* is the observed covariance matrix, *tr* is the trace matrix algebra function, and *(p + q)* is equal to the number of coefficients that need to be estimated in the model. The superscript in the middle, $^{-1}$, is a matrix algebra function called the *inverse*.[2]

[2] Computer packages sometimes print an error message stating that the "inverse of sigma is not positive definite." This indicates a severe problem with the model because one or more of the implied variances from the variance/covariance matrix is negative.

Fit, Chi-square, and df

ML is an iterative process, so initial starting values (i.e., guesses) are generated by the computer, the discrepancy between the implied and the obtained covariance matrix is computed, then new guesses are entered, and so on, until the minimum possible discrepancy values is obtained. Each step is called an *iteration*. The idea is similar to the idea of ordinary least squares (OLS) in regression in which the squared errors or residuals are minimized to obtain the best fit of the regression line to the data and the regression coefficients.



To find the minimum value of the F_{ML} discrepancy (fit) function, derivatives from calculus are used to draw tangent lines that correspond a point on the curve. When the tangent line has a slope of zero, the minimized value of the function has been found. The computer stops and generates values for the fit of the overall model and the parameter values. The final value can be used in a chi-square test [$\chi^2 = (N-1)F_{ML}$]. If the fit is perfect, there will be no discrepancy between the implied and obtained covariances, and the chi-square will be zero. A chi-square nonsignificantly different from zero indicates a good fit. Significantly positive chi-squares indicate poor fit.

It should be noted that most textbooks give the following formula (or a variation using $p + q$ to distinguish paths between exogenous and endogenous variables from those between endogenous paths).

$$df = \frac{v(v+1)}{2} - p$$

This formula is used because the number of unique variance/covariance elements (including the diagonal) is $v(v+1)$. Using this method, however, means that one must count the number of variances in the model when determining the value of p. Both models lead to the same result.

Example 4: One-factor CFA

The syntax below specifies a one-factor model for the items of the CES-D. The model posits that a single latent variable, depression, is the common cause of the responses to each of the items on the measure.

To test this model, we need an additional Mplus statement, the BY statement. BY stands for "measured by." Under the MODEL section, we need a new variable name for our latent variable of depression, which is listed on the left (cesd1), and we need to specify which items are indicators of that latent variable (items b1pc04a through b1pc04t), which are listed on the right.

All CFAs need an arbitrary scaling constraint, and there are two ways to make that constraint. The first method is to choose an item and set the loading equal to 1.0, which I usually refer to as a "marker variable." Mplus chooses the first item in the list as a marker variable by constraining or "fixing" its loading to 1.0 (you will see this in the output), but any item can be used. Many prefer to use the item with the highest loading.

```
TITLE:  Example 4, 1-factor CFA;

DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
               FORMAT=FREE;

VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
    b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
   b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
  b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
    b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
           b3v3pneg b1p_conf b2p_conf b3p_conf;

    MISSING = program-b3p_conf(-99,-6--1);

    USEVARIABLES=b1pc04a-b1pc04t;

ANALYSIS:
   TYPE = GENERAL; ESTIMATOR = ML; MATRIX = COVARIANCE;

MODEL:   cesd1 BY b1pc04a-b1pc04t;

OUTPUT:  STANDARDIZED;
```

An alternative (but equally valid) approach to the scaling constraint is to free the first loading (override the default) and then set the variance of the latent variable equal to 1.0.

```
cesd1 BY b1pc04a*1 b1pc04b-b1pc04t;
cesd1@1;
```

Listing a latent or measured variable without any other statement is a reference to the variance or residual variance (i.e., error term) associated with that variable. The @ symbol sets the value to a specific number. Omitting the @ symbol or using a * symbol tells Mplus you want the program to freely estimate that value.

Output for Example 4: One-factor CFA

```
INPUT READING TERMINATED NORMALLY

Example 4, 1-factor CFA;

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        2250

Number of dependent variables                                   12
Number of independent variables                                  0
Number of continuous latent variables                            1

Observed dependent variables

  Continuous
   B1PC04A     B1PC04B     B1PC04C     B1PC04E     B1PC04F     B1PC04G
   B1PC04J     B1PC04K     B1PC04M     B1PC04N     B1PC04R     B1PC04T

Continuous latent variables
   CESD1


Estimator                                                       ML
Information matrix                                        EXPECTED
Maximum number of iterations                                  1000
Convergence criterion                                  0.500D-04
Maximum number of steepest descent iterations                   20

Input data file(s)
  c:\jason\mplus\ehs\ex2.dat

Input data format  FREE



THE MODEL ESTIMATION TERMINATED NORMALLY



TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                            520.952
          Degrees of Freedom                    54
          P-Value                           0.0000

Chi-Square Test of Model Fit for the Baseline Model

          Value                           7806.940
          Degrees of Freedom                    66
          P-Value                           0.0000

CFI/TLI

          CFI                               0.940
          TLI                               0.926

Loglikelihood

          H0 Value                     -32360.102
          H1 Value                     -32099.626

Information Criteria

          Number of Free Parameters             24
          Akaike (AIC)                   64768.204
```

```
          Bayesian (BIC)                 64905.453
          Sample-Size Adjusted BIC       64829.201
            (n* = (n + 2) / 24)

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.062
          90 Percent C.I.                    0.057  0.067
          Probability RMSEA <= .05           0.000

SRMR (Standardized Root Mean Square Residual)

          Value                              0.037


MODEL RESULTS

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

 CESD1    BY
    B1PC04A             1.000    0.000      0.000    0.505    0.594
    B1PC04B             0.788    0.044     17.954    0.398    0.437
    B1PC04C             1.168    0.045     25.886    0.590    0.700
    B1PC04E             1.005    0.049     20.674    0.508    0.518
    B1PC04F             1.410    0.051     27.673    0.712    0.775
    B1PC04G             0.740    0.054     13.825    0.374    0.325
    B1PC04J             0.834    0.038     21.809    0.421    0.554
    B1PC04K             1.031    0.051     20.079    0.521    0.499
    B1PC04M             0.813    0.043     18.733    0.410    0.459
    B1PC04N             1.234    0.049     25.072    0.623    0.668
    B1PC04R             1.301    0.048     27.291    0.657    0.758
    B1PC04T             1.015    0.047     21.765    0.512    0.553

 Variances
    CESD1               0.255    0.017     14.672    1.000    1.000

 Residual Variances
    B1PC04A             0.467    0.015     31.042    0.467    0.647
    B1PC04B             0.672    0.021     32.465    0.672    0.809
    B1PC04C             0.363    0.012     29.133    0.363    0.511
    B1PC04E             0.703    0.022     31.867    0.703    0.732
    B1PC04F             0.337    0.013     26.599    0.337    0.399
    B1PC04G             1.183    0.036     33.003    1.183    0.894
    B1PC04J             0.401    0.013     31.516    0.401    0.693
    B1PC04K             0.816    0.025     32.023    0.816    0.751
    B1PC04M             0.630    0.019     32.321    0.630    0.789
    B1PC04N             0.481    0.016     29.839    0.481    0.554
    B1PC04R             0.319    0.012     27.312    0.319    0.425
    B1PC04T             0.597    0.019     31.531    0.597    0.695


R-SQUARE

    Observed
    Variable  R-Square

    B1PC04A      0.353
    B1PC04B      0.191
    B1PC04C      0.489
    B1PC04E      0.268
    B1PC04F      0.601
    B1PC04G      0.106
    B1PC04J      0.307
    B1PC04K      0.249
    B1PC04M      0.211
    B1PC04N      0.446
    B1PC04R      0.575
    B1PC04T      0.305
```

Alternative Fit Indices

Although chi-square is nearly always reported, it has a number of serious problems as a measure of overall model fit.

- Sensitive to sample size. Larger samples increase power of chi-square leading the researcher to reject models that might be good.
- Sensitive to model complexity. Larger models will tend to be rejected.
- Sensitive to violations of multivariate normality assumption. Models with highly skewed or kurtotic variables will tend to be rejected.

In response to some of the problems with chi-square, statisticians have developed a plethora of alternative fit indices. The article by Tanaka (1993) is a good review of the basic fit indices, their rationales, and how to interpret them, and more information is available in Technical Note # 2.

Recent work by Hu and Bentler (1999) has been influential in narrowing down the choice of fit indices. Based on their simulation work examining appropriate cutoffs for fit indices, they make the following recommendations:

- Use the Comparative Fit Index (CFI; Bentler, 1990) or the Incremental Fit Index (IFI; Bollen, 1989) in conjunction with the standardized root mean square residual (SRMR; Bentler, 1995) or the root mean square error of approximation (RMSEA, Steiger & Lind, 1980).
- The following cutoffs for these indices are optimal for minimizing false rejection and acceptance: CFI or IFI—a good fit > .95, SRMR—good fit < .08, and RMSEA—good fit < .06.[3]

[3] Steiger previously recommended < .05 as the critical value for RMSEA and Mplus prints confidence intervals and a significance test of whether the sample estimate is less than .05 in the population.

Technical Note # 2 Some Clarifications and Recommendations on Fit Indices

Tanaka (1993), Maruyama (1998), and others distinguish between several types of fit indices: *absolute fit indices*, *relative fit indices*, *parsimony fit indices*, and those based on the *noncentrality* parameter.

Absolute Fit Indices (χ^2, GFI, AGFI, Hoelter's CN, AIC, BIC, ECVI)

Absolute fit indices do not use an alternative model as a base for comparison. They are simply derived from the fit of the obtained and implied covariance matrices and the ML minimization function. Chi-square (χ^2, sometimes referred to as *T*) is the original fit index for structural models because it is derived directly from the fit function [f_{ML}(N-1)].

Chi-square is not a very good fit index in practice under many situations because it is affected by the following factors (1) sample size: larger samples produce larger chi-squares that are more likely to be significant (Type I error). Small samples may be too likely to accept poor models (Type II error). Based on my experience, it is difficult to get a nonsignificant chi-square when samples sizes are much over 200 or so, even when other indices suggest a decent fitting model. (2) model size also has an increasing effect on chi-square values. Models with more variables and more complicated models tend to have larger chi-squares. (3) Chi-square is affected by the distribution of variables. Highly skewed and kurtotic variables increase chi-square values. This has to do with the multivariate normality assumption that we will discuss later in the class.

There are several other indices that fall into the category of absolute indices, including the Goodness-of-fit index (GFI, also known as gamma-hat or $\hat{\gamma}$), the adjusted goodness of fit index (AGFI), χ^2/df ratio, Hoelter's CN ("critical N"), Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Expected Cross-validation Index (ECVI), the root mean square residual (RMR), and the standardized root mean square residual (SRMR). Most of these indices, with the possible exception of the SRMR, have similar problems to those of the chi-square, because they are based on simple variations on chi-square. As one example, the AIC (as given by Tanaka, 1993) is $\chi^2 + 2(p)$, where *p* is the number of free parameters (the number counted in calculating df).

Relative Fit Indices (IFI, TLI, NFI)

Relative fit indices compare a chi-square for the model tested to one from a so-called *null model* (also called a "baseline" model or "independence" model). The null model is a model tested that specifies that all measured variables are uncorrelated (there are no latent variables). The null model should always have a very large chi-square (poor fit). Although other baseline models could be used, this is not often seen in practice.[4] There

[4] Mplus version 3.11 uses a slightly modified null model in which any correlations among exogenous variables that are estimated in the hypothesized model are also estimated in the null model. This adjusts the df downward for the null (baseline) model and has the effect of lowering the relative fit index values (IFI, TLI, NFI, as well as the CFI).

are several relative fit indices (which are not explicitly designed to be provide penalties for parsimonious models), including Bollen's Incremental Fit Index (IFI, also called BL89 or Δ_2), the Tucker-Lewis Index [TLI, Bentler-Bonett Nonnormed Fit Index (NFI or BBNFI), or ρ_2], and the Bentler-Bonett Normed Fit Index (NFI). Most of these fit indices are computed by using ratios of the model chi-square and the null model chi-square and dfs for the models. All of them have values that range between approximately 0 and 1.0. Some of these indices are "normed" so that their values cannot be below 0 or above 1 (e.g., NFI, CFI described below). Others are considered "nonnormed" because, on occasion, they may be larger than 1 or slightly below 0 (e.g., TLI, IFI). In the past, these indexes have generally been used with a conventional cutoff in which values larger than .90 are considered good fitting models.

Parsimonious Fit Indices (PGFI, PNFI, PNFI2, PCFI)

These fit indices are relative fit indices that are adjustments to most of the ones above. The adjustments are to penalize models that are less parsimonious, so that simpler theoretical processes are favored over more complex ones. The more complex the model, the lower the fit index. Parsimonious fit indices include PGFI (based on the GFI), PNFI (based on the NFI), PNFI2 (based on Bollen's IFI), PCFI (based on the CFI mentioned below). Mulaik et al. (1989) developed a number of these. Although many researchers believe that parsimony adjustments are important, there is some debate about whether or not they are appropriate. My own perspective is that researchers should evaluate model fit independent of parsimony considerations, but evaluate alternative theories favoring parsimony. With that approach, we would not penalize models for having more parameters, but if simpler alternative models fit equally well, we might want to favor the simpler model.

Noncentrality-based Indices (RMSEA, CFI, RNI, CI)

The concept of the *noncentrality parameter* is a somewhat difficult one. The rationale for the noncentrality parameter is that our usual chi-square fit is based on a test of the null hypothesis is true (χ^2=0). This gives a distribution of the "central" chis-square. Because our we are hoping to reject the null hypothesis, it can be argued that we should be testing to reject the alternative hypothesis (H_a). Therefore, we should be conducting tests taking into account the noncentral chi-square distribution created under the case when H_a is true and thus the noncentral chi-square representing a model that is actually incorrect in the population. The estimate of what is the best possible fit for this incorrect model is based on the df for the model being tested. So, a model with a df of 2 would have a perfect fit if the chi-square equaled 2 under this rationale (rather than 0 as before). Thus, the noncentrality parameter is calculated by subtracting the df of the model from the chi-square ($\chi^2 - df$). Usually this value is adjusted for sample size and referred to as the rescaled noncentrality parameter:

$$d = \frac{\chi^2 - df}{N - 1}$$

In models with few exogenous variables, this will probably have a minor effect, but the incremental fit values may be substantially lower in models with more exogenous variables.

A population version is usually referred to as δ and is computed by dividing by N rather than N-1. Noncentrality-based indices include the Root Mean Square Error of Approximation (RMSEA)—not to be confused with RMR or SRMR, Bentler's Comparative Fit Index (CFI), McDonald and Marsh's Relative Noncentrality Index (RNI), and McDonald's Centrality Index (CI). Because the noncentrality parameter is simply a function of chi-square, df, and N, several of the formulas for the relative fit indices described above can be algebraically manipulated to include the noncentrality parameter. For example the TLI can also be presented as:

$$TLI = \frac{\left(d_0 / df_0\right) - \left(d_{model} / df_{model}\right)}{d_0 / df_0}$$

Where d_{model} and df_{model} are the noncentrality parameter and the degrees of freedom for the model tested and d_0 and df_0 are the noncentrality parameter for the null model. A recent article by Raykov (2000) shows that noncentrality parameter sample estimates are biased and that this problem may affect fit indices computed based on noncentrality.

Sample Size Independence

Many of the relative fit indices (and the noncentrality fit indices) are affected by sample size, so that larger samples are seen as better fitting (i.e., have a higher fit index value). Bollen (1990) made a very useful distinction between fit indices that can be shown to explicitly include N in their calculation and those that are dependent on sample size empirically. That is, even though a fit index may not include N in the formula, or even attempt to adjust for it, does not mean that the fit index will really turn out to be independent of sample size. He also showed that the TLI and IFI are relatively unaffected by sample size (see also Anderson & Gerbing, 1993; Hu & Bentler, 1995; Marsh, Balla, & McDonald, 1988). This is the basis for why I tend to favor these two indices.

$$TLI = \frac{\chi^2_{null} / df_{null} - \chi^2_{model} / df_{model}}{\chi^2_{null} / df_{null} - 1}$$

$$IFI = \frac{\chi^2_{null} - \chi^2_{model}}{\chi^2_{null} - df_{model}}$$

If you are interested in adjusting for parsimony, you might consider the Mulaik et al.'s PNFI2 which is a parsimony adjusted version of the IFI. One can make a similar argument about parsimony adjustment. There may be an important distinction between fit indices that are explicitly adjusting for parsimony and ones that are empirically affected by model complexity. The TLI is a example of an index that adjusts for parsimony, even though that was not its original intent.

Recommendations

Every researcher and every statistician seems to have a favorite index or set of indices. You should be prepared for reviewers to suggest the addition of one or two of their favorite indices, but it would not be fair to yourself or others to pick the index that is most optimistic about the fit of your model. In recent years, there has been concern that the recommended cutoff values for relative fit indices of .90 are too low and that higher

values, such as .95 should be used. Hu and Bentler (1999) empirically examine various cutoffs for many of these measures, and there data suggest that to minimize Type I and Type II errors under various conditions, one should use a combination of one of the above relative fit indexes and the SRMR (good models < .08). These values should not be written in stone, but I believe this is useful work and hope it will helpful for establishing a more concrete basis for conventional cutoff values in the future. Based on the IFI's and TLI's independence of sample size and the data from Hu and Bentler, I expect to report the IFI and/or the TLI in combination with the SRMR in my work.

Example 5: Two-factor CFA

The CES-D has been shown to be a multifactor scale, with the full 20-item comprised of three or four factors (i.e., negative affect, positive affect, somatic symptoms, interpersonal symptoms). The 12-item scale used in the EHS has items from two factors—the negative affect and the somatic symptom factors. So, a two-factor CFA may show a better fit with the data. In Mplus, here is how a two-factor scale would be specified:

```
TITLE:  Example 5, 2-factor CFA;

DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
               FORMAT=FREE;

VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
    b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
   b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
  b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
    b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
           b3v3pneg b1p_conf b2p_conf b3p_conf;

    MISSING = program-b3p_conf(-99,-6--1);

    USEVARIABLES=b1pc04a-b1pc04t;


ANALYSIS:
   TYPE = GENERAL; ESTIMATOR = ML; MATRIX = COVARIANCE;


MODEL:   cesd1som BY b1pc04a b1pc04b b1pc04e b1pc04g b1pc04k
                  b1pc04m b1pc04t;
         cesd1neg BY b1pc04c b1pc04f b1pc04j b1pc04n b1pc04r;
         cesd1som WITH cesd1neg;

OUTPUT:  STANDARDIZED;
```

Output for Example 5: Two-factor CFA

```
INPUT READING TERMINATED NORMALLY

Example 5, 2-factor CFA;

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        2250

Number of dependent variables                                   12
Number of independent variables                                  0
Number of continuous latent variables                            2

Observed dependent variables

  Continuous
   B1PC04A     B1PC04B     B1PC04C     B1PC04E     B1PC04F     B1PC04G
   B1PC04J     B1PC04K     B1PC04M     B1PC04N     B1PC04R     B1PC04T

Continuous latent variables
   CESD1SOM    CESD1NEG


Estimator                                                       ML
Information matrix                                        EXPECTED
Maximum number of iterations                                  1000
Convergence criterion                                    0.500D-04
Maximum number of steepest descent iterations                   20

Input data file(s)
  c:\jason\mplus\ehs\ex2.dat

Input data format  FREE

THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                            315.134
          Degrees of Freedom                53
          P-Value                           0.0000



Chi-Square Test of Model Fit for the Baseline Model

          Value                           7806.940
          Degrees of Freedom                66
          P-Value                           0.0000

CFI/TLI

          CFI                               0.966
          TLI                               0.958

Loglikelihood

          H0 Value                     -32257.193
          H1 Value                     -32099.626

Information Criteria

          Number of Free Parameters            25
          Akaike (AIC)                  64564.386
          Bayesian (BIC)                64707.353
          Sample-Size Adjusted BIC      64627.924
            (n* = (n + 2) / 24)
```

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.047	
90 Percent C.I.	0.042	0.052
Probability RMSEA <= .05	0.841	

SRMR (Standardized Root Mean Square Residual)

Value	0.029

MODEL RESULTS

	Estimates	S.E.	Est./S.E.	Std	StdYX
CESD1SOM BY					
B1PC04A	1.000	0.000	0.000	0.528	0.621
B1PC04B	0.828	0.044	18.903	0.437	0.480
B1PC04E	1.048	0.049	21.580	0.553	0.565
B1PC04G	0.759	0.053	14.265	0.401	0.348
B1PC04K	1.100	0.052	21.344	0.581	0.557
B1PC04M	0.793	0.043	18.537	0.419	0.469
B1PC04T	1.068	0.047	22.842	0.564	0.608
CESD1NEG BY					
B1PC04C	1.000	0.000	0.000	0.589	0.699
B1PC04F	1.247	0.037	33.496	0.734	0.799
B1PC04J	0.720	0.030	24.128	0.424	0.557
B1PC04N	1.073	0.037	28.990	0.632	0.678
B1PC04R	1.146	0.035	32.792	0.675	0.779
CESD1SOM WITH					
CESD1NEG	0.265	0.014	19.579	0.852	0.852
Variances					
CESD1SOM	0.279	0.019	14.833	1.000	1.000
CESD1NEG	0.347	0.019	17.925	1.000	1.000
Residual Variances					
B1PC04A	0.443	0.015	28.672	0.443	0.614
B1PC04B	0.639	0.020	31.259	0.639	0.770
B1PC04C	0.364	0.013	28.693	0.364	0.512
B1PC04E	0.654	0.022	29.944	0.654	0.681
B1PC04F	0.305	0.012	24.463	0.305	0.361
B1PC04G	1.162	0.036	32.495	1.162	0.879
B1PC04J	0.399	0.013	31.272	0.399	0.689
B1PC04K	0.750	0.025	30.090	0.750	0.690
B1PC04M	0.623	0.020	31.394	0.623	0.780
B1PC04N	0.470	0.016	29.227	0.470	0.541
B1PC04R	0.295	0.012	25.635	0.295	0.394
B1PC04T	0.542	0.019	29.007	0.542	0.630

R-SQUARE

Observed Variable	R-Square
B1PC04A	0.386
B1PC04B	0.230
B1PC04C	0.488
B1PC04E	0.319
B1PC04F	0.639
B1PC04G	0.121
B1PC04J	0.311
B1PC04K	0.310
B1PC04M	0.220
B1PC04N	0.459
B1PC04R	0.606
B1PC04T	0.370

Nested Models & Chi-square Difference Tests

- It is often recommended that researchers compare the fit of their model to alternative models.

- A chi-square difference test can be conducted using chi-square values and degrees of freedom from any two *nested models*.

- Nested models are models that use the same variables but specify at least one different parameter (e.g., comparing a one-factor to a two-factor model).

- For a model to be nested, all of the same measured variables and the same cases must be used. There are some cases where a model is not nested even though both of these conditions are met (see Rigdon, 1995)

- The chi-square test is simply the difference between the original model and the nested model, using the difference in degrees of freedom as the degrees of freedom for the test.

Example
In Example 4, the one-factor CFA of the CES-D scale, the chi-square value was 520.952 with 54 degrees of freedom. One can test whether the two-factor model fits significantly better. The chi-square for the two-factor model was 315.134 with 53 df. Notice that the only difference between the two models is that one more parameter is being estimated, namely the correlation between the two factors.

$$\begin{aligned}\chi^2_{diff} &= \chi^2 - \chi^2_{nested} \\ &= 520.952 - 315.134 \\ &= 205.818\end{aligned}$$

$$\begin{aligned}df_{diff} &= df - df_{nested} \\ &= 54 - 53 \\ &= 1\end{aligned}$$

The critical value for chi-square difference value with 1 df is 3.84, so the two-factor model fits significantly better than the one-factor model.

Model Modification and Modification Indices

Modification indices, which can be requested in Mplus and from most computer packages, are one degree of freedom chi-square tests of the addition of a new parameter or the deletion of a parameter (in EQS these tests are called Lagrange multiplier and Wald tests). Each modification index represents the change in the overall chi-square for the fit of the model if that particular parameter is changed. Thus, a significant chi-square value (greater than 3.84) will significantly improve the fit of the model. Some packages also will print the expected change in a parameter, representing what the unstandardized or standardized value of the added or deleted path would be.

A few comments on modification indices:

- Model modifications are subject Type I errors, so most researchers do not recommend making changes in a model that are not theoretically sensible.

- Many modifications to a model are considered exploratory and can lead to development of incorrect models.

- Modification indices can be an important source of information about whether a given model can be improved beyond a certain point. Fit index cutoffs are useful, but comparisons to alternative models and evaluation of theoretically sensible modifications to a model are also essential to evaluating models.

- Because chi-square tests (including chi-square difference tests) are sensitive to sample size, large samples may produce many significant modification indices. It is generally impractical and unreasonable change or even consider changing all significant MIs. MI values should be evaluated in relationship to the magnitude of the change in the overall model chi-square (I often approximate the percentage change in the overall chi-square).

- Modifications will not provide information about major changes that are needed (e.g., whether a one vs. two-factor model is appropriate)

Example 6: Modification Indices

In Mplus, the addition of one statement under the OUTPUT command, MODINDICES, will produce modification indices in the output. The default minimum modification index value is 10, but this can be overridden. Because a 1-df chi-square is significant at 3.84, I often use this as the minimum value. I modified the OUTPUT command from Example 5, in the following way:

```
OUTPUT:  STANDARDIZED MODINDICES(3.84);
```

The output looks exactly the same as that from Example 5, but the following information is appended.

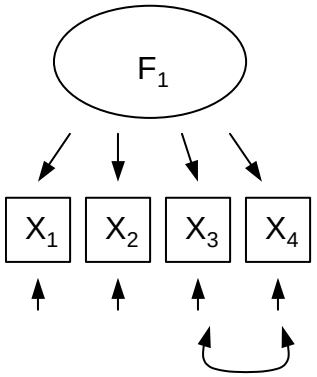
```
MODEL MODIFICATION INDICES

Minimum M.I. value for printing the modification index     3.840

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

CESD1SOM BY B1PC04C        34.798     0.521      0.275        0.326
CESD1SOM BY B1PC04F        17.453    -0.393     -0.208       -0.226
CESD1SOM BY B1PC04J         4.354     0.178      0.094        0.124
CESD1SOM BY B1PC04R         6.801    -0.231     -0.122       -0.141
CESD1NEG BY B1PC04A        24.649     0.461      0.272        0.320
CESD1NEG BY B1PC04B         4.748    -0.213     -0.125       -0.138
CESD1NEG BY B1PC04K        14.616    -0.429     -0.253       -0.242
CESD1NEG BY B1PC04M        21.008     0.440      0.259        0.290
CESD1NEG BY B1PC04T         5.374    -0.234     -0.138       -0.149

WITH Statements

B1PC04C  WITH B1PC04A      21.793     0.046      0.046        0.064
B1PC04C  WITH B1PC04B      21.189     0.052      0.052        0.068
B1PC04F  WITH B1PC04B      14.789    -0.043     -0.043       -0.051
B1PC04F  WITH B1PC04C      10.832     0.033      0.033        0.042
B1PC04F  WITH B1PC04E       7.997     0.033      0.033        0.036
B1PC04J  WITH B1PC04C      10.819    -0.030     -0.030       -0.047
B1PC04K  WITH B1PC04A      11.386    -0.049     -0.049       -0.055
B1PC04K  WITH B1PC04B      28.628     0.087      0.087        0.091
B1PC04K  WITH B1PC04C       4.291    -0.026     -0.026       -0.029
B1PC04K  WITH B1PC04F       8.227    -0.035     -0.035       -0.037
B1PC04M  WITH B1PC04C      11.978     0.038      0.038        0.051
B1PC04M  WITH B1PC04E       4.731    -0.032     -0.032       -0.037
B1PC04N  WITH B1PC04C      20.177    -0.047     -0.047       -0.060
B1PC04N  WITH B1PC04M      47.280     0.086      0.086        0.103
B1PC04R  WITH B1PC04C       9.071    -0.028     -0.028       -0.039
B1PC04R  WITH B1PC04E       8.768    -0.033     -0.033       -0.039
B1PC04R  WITH B1PC04M       6.479    -0.027     -0.027       -0.035
B1PC04R  WITH B1PC04N       9.606     0.032      0.032        0.040
B1PC04T  WITH B1PC04A      16.486    -0.051     -0.051       -0.065
B1PC04T  WITH B1PC04E      13.826     0.055      0.055        0.060
B1PC04T  WITH B1PC04K      32.979     0.091      0.091        0.094
B1PC04T  WITH B1PC04M       8.783    -0.041     -0.041       -0.050
```

The first column gives the change in chi-square for each modification to the model. The first set of results, under “BY Statements” are for adding loadings that were assumed to be zero in the model (i.e., cross-loadings onto the other factor). The second set of MIs are for correlated measurement errors that could be added that would significantly improve the fit. E.P.C. stands for “expected parameter change” and provides the value of the parameter if added. Unstandardized values are given under the E.P.C. column and standardized values are found under the StdXY E.P.C. column.

Correlated Errors

Typically, researchers begin testing a confirmatory factor model by assuming that measurement errors are independent of one another. This assumption implies that the variance of a particular item not caused by the factor has a source that is unique to that particular variable. This assumption is not always valid, so researchers may incorporate correlated errors in a factor model.

Inclusion of the correlation will decrease the loadings for the items involved if the correlation is positive. The following figure graphically illustrates a correlation between errors for items 3 and 4. The basis for including a correlation between measurement errors can be data driven or theoretically driven.



- Researchers frequently incorporate one or two correlated errors because modification indices suggest an important improvement in fit of the model. In such instances, correlated errors are often due to similar item wording or content (e.g., “I feel blue whenever my spouse is around” and “I feel sad whenever my spouse is around”).

- Correlated errors are often used in longitudinal designs in which the same item is asked twice. In this case, parallel items across factors are allowed to correlate. When they are not included under these circumstances, predictive paths across time may be inflated.

- Correlated errors may be used to account for methods effects (e.g., telephone interviews and face to face interviews).

In Example 6 above, the largest modification index was for b1pc04n (felt lonely) WITH b1pc04m (talked less) and would improve the chi-square value by 47.280, approximately 15% improvement in the overall chi-square value of the model. To make this change, we would simply add the statement b1pc04n WITH b1pc04m to the MODEL command.

Full Structural Equation Models

Overview

- In practice, most structural models are more complex and include latent variables and predictive paths.

- Any combination of measured variables and latent variables can be used. The portion of the model involving the latent variables is often referred to as the *measurement* model. The predictive path model portions are referred to as the *causal or structural* portion of the model.

- A general two-step approach is often recommended in which the measurement portion is tested and improved until and adequate fit is achieved. Then the full structural model is tested.

- The complexity of the model tested is entirely up to the researcher barring limitations of the data. Instead of one mediator variable, the researcher can include four mediators if theory dictates.

- A major advantage to full structural equation models is the ability to incorporate latent variables in the model. Using latent variables instead of measured variables for predictive relationships is important because measurement error attenuates relationships.

    - With simple bivariate relationships, associations between two measured variables (e.g., a composite index), can be substantially weaker than the association between to latent variables, depending on the reliability of the index.

    - With more complex relationships, in which covariates are involved, the attenuation problem can lead to overestimation of some predictive paths.

    - In the figures below, assume X1 and X2 are composite indexes made up of several items, and Eta1 and Eta2 are latent variables constructed with the same items. If X1 is not perfectly reliable, path b2 will be overestimated. b1* and b2* will be more accurately estimated, because latent measurement error is removed from the model.

Fig 1a

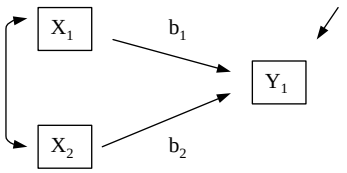
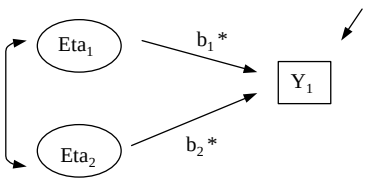
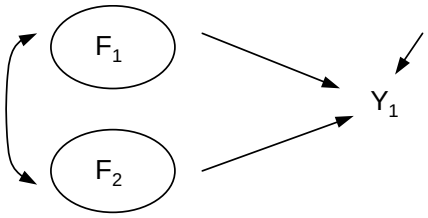


Fig 1b



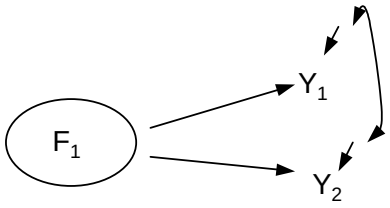
Predictor Intercorrelation & Correlated Disturbances

It is fairly common practice to allow exogenous variables to correlate with one another. If a correlation between two predictor variables is not included, the researcher assumes that the correlation between those variables is zero.



The above figure illustrates a correlation between the two exogenous variables, F_1 and F_2. If this correlation is greater than zero (positive), but the researcher does not specify it should be estimated (i.e., it is set to zero), the predictive paths from F_1 and F_2 to Y_1 will simply reflect the zero-order correlation between the predictors and the outcome.

For endogenous variables, one cannot estimate a correlation between variables—only between their disturbances. Correlations between disturbances are fairly common in practice. Such correlations represent a common source of error variation affecting both dependent variables. Fit can be substantially affected if there is residual correlation between the two endogenous variables, but that correlation is not estimated.



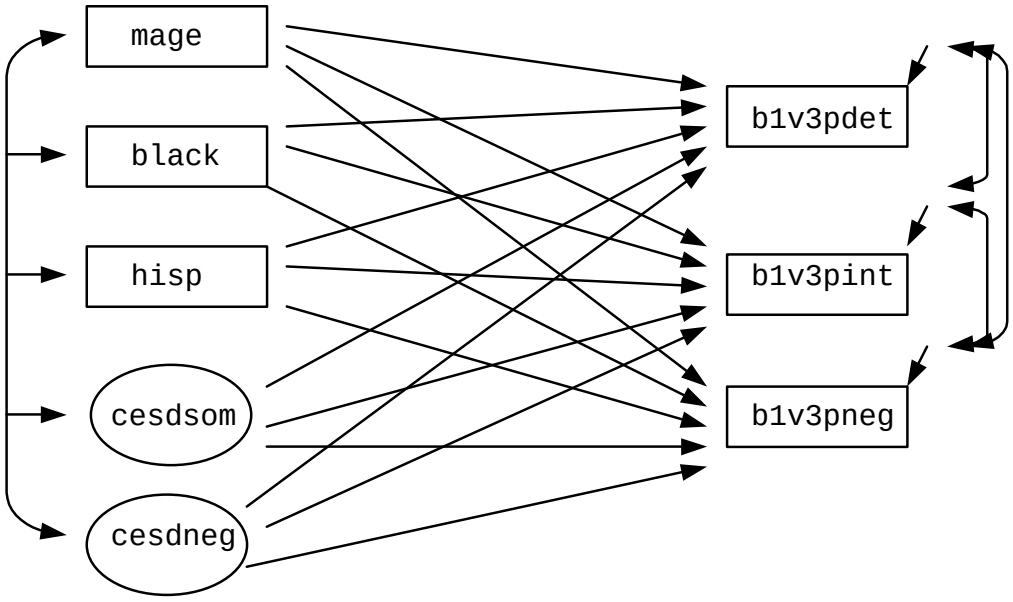
In Mplus, any two variables, measurement errors, or disturbances can be correlated by using a WITH statement. By default, exogenous variables are correlated, measurement errors are uncorrelated, and endogenous disturbances are correlated. I have noticed that Mplus often does not report these correlations in the output if they are the default (even though Mplus is estimating them). I recommend adding these statements in the model so you know they are there when you look back at the model at a future point, and so that you can see their estimated values in the output.

Full SEM Example

The following is an example of a full structural model with the two CES-D factors as predictors of parental detachment (b1v3pdet), intrusiveness (b1v3pint), and negative regard (b1v3pneg) at 14 months. I have added mother's age (mage) and race as covariates (race).

Because race is a nominal variable with four categories (1=White, 2=Black, 3=Hispanic, 4=Other), dummy variables are needed to represent several categories. In the example below, I illustrate the use of the DEFINE command in Mplus to create two dummy variables called "black" and "hisp" (the other category is combined with the White category).

The following figure illustrates the structural portion of the model:



Example 7: Full Structural Equation Model

```
INPUT INSTRUCTIONS

  TITLE:  Example 7, full SEM;

  DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
                FORMAT=FREE;

  VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
      b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
          b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
          b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
          b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
          b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
          b3v3pneg b1p_conf b2p_conf b3p_conf;

      MISSING = program-b3p_conf(-99,-6--1);

      USEVARIABLES=b1pc04a-b1pc04t b1v3pdet b1v3pint b1v3pneg
               mage black hisp;

  DEFINE:  IF (race EQ 1 OR race GE 3) THEN black = 0;
           IF (race EQ 2) THEN black = 1;
           IF (race EQ 1 OR race EQ 2 OR race EQ 4) THEN hisp = 0;
           IF (race EQ 3) THEN hisp = 1;

  ANALYSIS:
          TYPE = GENERAL; ESTIMATOR = ML; MATRIX = COVARIANCE;

  MODEL:  cesd1som BY b1pc04a b1pc04b b1pc04e b1pc04g b1pc04k
                      b1pc04m b1pc04t;
          cesd1neg BY b1pc04c b1pc04f b1pc04j b1pc04n b1pc04r;
          b1v3pdet b1v3pint b1v3pneg ON cesd1som cesd1neg
                    mage black hisp;
          cesd1som with cesd1neg mage black hisp;
          cesd1neg with mage black hisp;
          mage with black hisp;
          black with hisp;
          b1v3pdet with b1v3pint b1v3pneg;
          b1v3pint with b1v3pneg;

  OUTPUT:  STANDARDIZED;


INPUT READING TERMINATED NORMALLY



Example 7, full SEM;

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        1832

Number of dependent variables                                   15
Number of independent variables                                  3
Number of continuous latent variables                            2

Observed dependent variables

  Continuous
   B1PC04A     B1PC04B     B1PC04C     B1PC04E     B1PC04F     B1PC04G
   B1PC04J     B1PC04K     B1PC04M     B1PC04N     B1PC04R     B1PC04T
   B1V3PDET    B1V3PINT    B1V3PNEG

Observed independent variables
   MAGE        BLACK       HISP
```

```
Continuous latent variables
   CESD1SOM    CESD1NEG


Estimator                                                       ML
Information matrix                                        EXPECTED
Maximum number of iterations                                  1000
Convergence criterion                                  0.500D-04
Maximum number of steepest descent iterations                   20

Input data file(s)
  c:\jason\mplus\ehs\ex2.dat

Input data format  FREE



THE MODEL ESTIMATION TERMINATED NORMALLY



TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                            551.587
          Degrees of Freedom                   113
          P-Value                           0.0000

Chi-Square Test of Model Fit for the Baseline Model

          Value                           7580.111
          Degrees of Freedom                   150
          P-Value                           0.0000

CFI/TLI

          CFI                                0.941
          TLI                                0.922

Loglikelihood

          H0 Value                      -41257.104
          H1 Value                      -40981.310

Information Criteria

          Number of Free Parameters             58
          Akaike (AIC)                   82630.208
          Bayesian (BIC)                 82949.971
          Sample-Size Adjusted BIC       82765.708
            (n* = (n + 2) / 24)

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.046
          90 Percent C.I.                    0.042  0.050
          Probability RMSEA <= .05           0.954

SRMR (Standardized Root Mean Square Residual)

          Value                              0.034



MODEL RESULTS

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

 CESD1SOM BY
```

B1PC04A	1.000	0.000	0.000	0.522	0.614
B1PC04B	0.827	0.049	16.997	0.432	0.477
B1PC04E	1.069	0.054	19.733	0.558	0.573
B1PC04G	0.743	0.059	12.633	0.388	0.340
B1PC04K	1.088	0.057	19.151	0.568	0.552
B1PC04M	0.765	0.047	16.177	0.399	0.450
B1PC04T	1.086	0.052	20.937	0.567	0.620
CESD1NEG BY					
B1PC04C	1.000	0.000	0.000	0.588	0.700
B1PC04F	1.247	0.041	30.457	0.733	0.801
B1PC04J	0.702	0.033	21.370	0.413	0.545
B1PC04N	1.070	0.040	26.511	0.629	0.684
B1PC04R	1.143	0.038	29.936	0.672	0.784
B1V3PDET ON					
CESD1SOM	0.077	0.156	0.494	0.040	0.040
CESD1NEG	0.002	0.131	0.015	0.001	0.001
B1V3PINT ON					
CESD1SOM	0.141	0.189	0.747	0.074	0.060
CESD1NEG	0.031	0.159	0.196	0.018	0.015
B1V3PNEG ON					
CESD1SOM	0.083	0.122	0.676	0.043	0.054
CESD1NEG	0.008	0.103	0.075	0.005	0.006
B1V3PDET ON					
MAGE	-0.014	0.004	-3.514	-0.014	-0.083
BLACK	0.359	0.056	6.361	0.359	0.170
HISP	0.147	0.070	2.084	0.147	0.062
B1V3PINT ON					
MAGE	-0.014	0.005	-2.875	-0.014	-0.066
BLACK	0.728	0.068	10.688	0.728	0.278
HISP	0.533	0.085	6.272	0.533	0.181
B1V3PNEG ON					
MAGE	-0.010	0.003	-3.017	-0.010	-0.069
BLACK	0.474	0.044	10.707	0.474	0.279
HISP	0.125	0.055	2.274	0.125	0.066
CESD1SOM WITH					
CESD1NEG	0.262	0.015	17.630	0.853	0.853
MAGE	-0.193	0.081	-2.367	-0.369	-0.065
BLACK	-0.009	0.007	-1.278	-0.016	-0.035
HISP	-0.035	0.006	-5.778	-0.067	-0.161
CESD1NEG WITH					
MAGE	-0.115	0.085	-1.353	-0.196	-0.034
BLACK	-0.002	0.007	-0.351	-0.004	-0.009
HISP	-0.001	0.006	-0.184	-0.002	-0.005
MAGE WITH					
BLACK	-0.482	0.064	-7.538	-0.482	-0.179
HISP	0.214	0.056	3.810	0.214	0.089
BLACK WITH					
HISP	-0.076	0.005	-15.356	-0.076	-0.384
B1V3PDET WITH					
B1V3PINT	0.140	0.027	5.111	0.140	0.113
B1V3PNEG	0.193	0.018	10.629	0.193	0.242
B1V3PINT WITH					
B1V3PNEG	0.365	0.023	15.916	0.365	0.369
Variances					
MAGE	32.592	1.077	30.265	32.592	1.000
BLACK	0.223	0.007	30.265	0.223	1.000
HISP	0.176	0.006	30.265	0.176	1.000

CESD1SOM	0.273	0.021	13.282	1.000	1.000
CESD1NEG	0.346	0.021	16.245	1.000	1.000

Residual Variances

B1PC04A	0.449	0.017	26.251	0.449	0.622
B1PC04B	0.635	0.022	28.355	0.635	0.773
B1PC04C	0.360	0.014	25.954	0.360	0.510
B1PC04E	0.636	0.024	27.047	0.636	0.671
B1PC04F	0.300	0.014	22.156	0.300	0.358
B1PC04G	1.152	0.039	29.423	1.152	0.884
B1PC04J	0.404	0.014	28.392	0.404	0.703
B1PC04K	0.737	0.027	27.397	0.737	0.696
B1PC04M	0.630	0.022	28.623	0.630	0.798
B1PC04N	0.449	0.017	26.317	0.449	0.531
B1PC04R	0.283	0.012	23.022	0.283	0.385
B1PC04T	0.515	0.020	26.127	0.515	0.616
B1V3PDET	0.961	0.032	30.247	0.961	0.964
B1V3PINT	1.400	0.046	30.223	1.400	0.919
B1V3PNEG	0.590	0.019	30.232	0.590	0.919

R-SQUARE

Observed Variable	R-Square
B1PC04A	0.378
B1PC04B	0.227
B1PC04C	0.490
B1PC04E	0.329
B1PC04F	0.642
B1PC04G	0.116
B1PC04J	0.297
B1PC04K	0.304
B1PC04M	0.202
B1PC04N	0.469
B1PC04R	0.615
B1PC04T	0.384
B1V3PDET	0.036
B1V3PINT	0.081
B1V3PNEG	0.081

Some Practical Considerations

Here, I would like to cover several miscellaneous topics that arise when using SEM analyses in practice.

- There is no minimum sample size for SEM, but models based on an N less than 100 may have convergence problems (i.e., a ML solution is not found). Bentler and Chou (1988) suggest a convention of having at least 10 cases for every parameter. Tanaka (1987) suggests only a 5:1 ratio is needed.

- Nonconvergence occurs when a ML solution is not reached. Iterations continue until the maximum allowable is reached. Causes can include identification problems (i.e., the model is underidentified), specification errors, complex models, or latent variables with items that correlate poorly with one another.

- A warning message may indicate a "nonpositive definite sigma matrix." This means that the program was not able to take the inverse of the implied covariance matrix, because the model and data imply negative variances. The most common causes are model specification errors or identification problems. The error "the PSI matrix is nonpositive definite" is also a common error message and this results from a residual variance that is estimated to be negative.

- *Heywood cases* are nonsensical results that occur in the output. Usually they involve negative variances (usually measurement error variances) or standardized coefficients over 1.0. Often the computer packages do not print a warning, so always examine your output carefully!!

- Underidentified models are a common problem. Latent variables require at least three indicators (i.e., measured variables). Models with negative degrees of freedom are *theoretically underidentified*, but even some models that are theoretically identified, are *empirically underidentified*. Empirically underidentified models usually result from a model that, overall has sufficient df, but insufficient information is available for a portion of the model (e.g., two-indicator latent variables, bi-directional paths).

- A good practice is to test portions of your model first, then build up to a more complex model. Starting out with a model that is too complex can result in errors in syntax or specifications that cause estimation or fit problems that are difficult to locate.

Nonnormality and Alternative Estimators

Multivariate Normality Assumption

Most of the statistical assumptions for SEM are the same assumptions for regression analysis (e.g., independent observations, normally distributed errors, identically distributed errors or homoscedasticity) However, SEM with ML also assumes that the variables are *multivariate normal* in the population. Multivariate normality assumes *univariate normality*—that each variable is normally distributed. However, univariate normality is not sufficient for multivariate normality. It is possible to have univariate normality and multivariate nonnormality.

When the multivariate normality assumption is violated, chi-square values will tend to be overestimated (indicating poorer fit), and (in general) standard errors used in significance tests for parameters are underestimated. In other words, fit will tend to be poorer and significance of paths will be overestimated (Type I errors). Nonnormality does not affect the parameter estimates themselves—only the standard errors are biased.

An article by West, Finch, and Curran (1995)presents a nice introduction to problems, their detection, and solutions for multivariate nonnormality, but I will make a few brief points here.

Detection

The first step is to carefully examine univariate distributions and skew and kurtosis. West, Finch, & Curran (1995) recommend concern if skewness > 2 and kurtosis > 7. Kurtosis is usually a greater concern than skewness. If the univariate distributions are nonnormal, then the multivariate distribution will be nonnormal. Keep in mind that one can have multivariate nonnormality (i.e., the joint distributions of all the variables is a nonnormal joint distribution) even when all the individual variables are normally distributed (although this is relatively infrequent in practice). Therefore, one should also examine multivariate kurtosis and skewness.

Tests of multivariate normality are only available in EQS and Lisrel, but it is difficult to interpret them. No one has really provided good cutoff recommendations. Mardia's multivariate skewness and kurtosis tests are distributed normally (z-test) in very large samples, so can be evaluated against a t or z-distribution, but they tend to be sensitive to sample size. Other than a significance tests, there are no suggested cutoffs to identify when substantial problems exist. Lawrence DeCarlo (1997) has developed macros for SPSS and SAS to calculate a variety of multivariate nonnormality indices (available at http://www.columbia.edu/~ld208/).

In Mplus, when the Satorra-Bentler robust statistics (ESTIMATOR= MLM statement) are requested, a "scaling correction factor" is printed in the output. This correction factor can be taken as an index of the degree to which the chi-square value is inflated by multivariate kurtosis. A value of 1.10 represents 10% inflation and a value of 2.0

represents 100% inflation. There is no accepted cutoff on what value of the scaling correction factor is problematic, but I tend to have minimal concern if it is 1.05 or less.

Remedies

If there are non-normality concerns, one can use a correction to provide better estimates of the chi-square value and the standard errors. Satorra and Bentler (1994) developed a “rescaled” chi-square and robust standard error estimates that are corrected by a multivariate kurtosis weight matrix. This method seems to perform well in simulation studies (e.g., Hu, Bentler, & Kano, 1992; Curran, West, & Finch, 1996). Sample sizes of 250 or greater may be needed to avoid over correction of chi-square and standard errors.

Bootstrapping is an increasingly popular approach to correcting standard errors, but it seems that more work is needed to understand how well it performs under various conditions (e.g., specific bootstrap approach, sample sizes needed). Simulation work (e.g., Hancock and Nevitt, 1999).

I will only illustrate the Satorra-Bentler correction method here. Implementation is simple—one just needs to specify ESTIMATOR=MLM under the ANALYSIS section. As an example, I analyze the two-factor CFA from Example 5.

```
TITLE:  Example 8, 2-factor CFA with Satorra-Bentler corrections;

DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
               FORMAT=FREE;

VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
    b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
   b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
  b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
    b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
           b3v3pneg b1p_conf b2p_conf b3p_conf;

    MISSING = program-b3p_conf(-99,-6--1);

    USEVARIABLES=b1pc04a-b1pc04t;


ANALYSIS:
   TYPE = GENERAL; ESTIMATOR = MLM; MATRIX = COVARIANCE;


MODEL:   cesd1som BY b1pc04a b1pc04b b1pc04e b1pc04g b1pc04k
                  b1pc04m b1pc04t;
         cesd1neg BY b1pc04c b1pc04f b1pc04j b1pc04n b1pc04r;
         cesd1som WITH cesd1neg;

OUTPUT:  STANDARDIZED;
```

Example 8: Two-factor CFA with Rescaled Chi-square and Robust Standard Errors

```
THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                            241.193*
          Degrees of Freedom                    53
          P-Value                           0.0000
          Scaling Correction Factor          1.307
            for MLM

*   The chi-square value for MLM, MLMV, MLR, WLSM and WLSMV cannot be used for
    chi-square difference tests.  MLM, MLR and WLSM chi-square difference
    testing is described in the Mplus Technical Appendices at www.statmodel.com.
    See chi-square difference testing in the index of the Mplus User's Guide.

Chi-Square Test of Model Fit for the Baseline Model

          Value                           5998.889
          Degrees of Freedom                    66
          P-Value                           0.0000

CFI/TLI

          CFI                               0.968
          TLI                               0.960

Number of Free Parameters                       37

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                          0.040

SRMR (Standardized Root Mean Square Residual)

          Value                             0.027

WRMR (Weighted Root Mean Square Residual)

          Value                             1.425


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

 CESD1SOM BY
    B1PC04A            1.000    0.000      0.000    0.528    0.621
    B1PC04B            0.828    0.047     17.498    0.437    0.480
    B1PC04E            1.048    0.051     20.627    0.553    0.565
    B1PC04G            0.759    0.050     15.124    0.401    0.348
    B1PC04K            1.100    0.055     19.997    0.581    0.557
    B1PC04M            0.793    0.048     16.573    0.419    0.469
    B1PC04T            1.068    0.050     21.191    0.564    0.608

 CESD1NEG BY
    B1PC04C            1.000    0.000      0.000    0.589    0.699
    B1PC04F            1.247    0.043     29.293    0.734    0.799
    B1PC04J            0.720    0.043     16.743    0.424    0.557
    B1PC04N            1.073    0.045     23.873    0.632    0.678
    B1PC04R            1.146    0.042     27.316    0.675    0.779

 CESD1SOM WITH
    CESD1NEG           0.265    0.019     14.092    0.852    0.852
```

```
 Intercepts
    B1PC04A            1.658    0.018     92.544    1.658    1.951
    B1PC04B            1.683    0.019     87.596    1.683    1.847
    B1PC04C            1.532    0.018     86.250    1.532    1.818
    B1PC04E            1.907    0.021     92.294    1.907    1.946
    B1PC04F            1.652    0.019     85.337    1.652    1.799
    B1PC04G            2.280    0.024     94.044    2.280    1.983
    B1PC04J            1.417    0.016     88.387    1.417    1.863
    B1PC04K            1.963    0.022     89.311    1.963    1.883
    B1PC04M            1.609    0.019     85.423    1.609    1.801
    B1PC04N            1.653    0.020     84.121    1.653    1.773
    B1PC04R            1.671    0.018     91.500    1.671    1.929
    B1PC04T            1.744    0.020     89.181    1.744    1.880

 Variances
    CESD1SOM           0.279    0.022     12.495    1.000    1.000
    CESD1NEG           0.347    0.026     13.572    1.000    1.000

 Residual Variances
    B1PC04A            0.443    0.019     23.139    0.443    0.614
    B1PC04B            0.639    0.024     27.011    0.639    0.770
    B1PC04C            0.364    0.017     20.947    0.364    0.512
    B1PC04E            0.654    0.023     27.905    0.654    0.681
    B1PC04F            0.305    0.017     17.826    0.305    0.361
    B1PC04G            1.162    0.028     41.685    1.162    0.879
    B1PC04J            0.399    0.019     21.364    0.399    0.689
    B1PC04K            0.750    0.027     28.055    0.750    0.690
    B1PC04M            0.623    0.024     25.702    0.623    0.780
    B1PC04N            0.470    0.022     21.554    0.470    0.541
    B1PC04R            0.295    0.017     17.870    0.295    0.394
    B1PC04T            0.542    0.022     24.962    0.542    0.630


R-SQUARE

    Observed
    Variable  R-Square

    B1PC04A      0.386
    B1PC04B      0.230
    B1PC04C      0.488
    B1PC04E      0.319
    B1PC04F      0.639
    B1PC04G      0.121
    B1PC04J      0.311
    B1PC04K      0.310
    B1PC04M      0.220
    B1PC04N      0.459
    B1PC04R      0.606
    B1PC04T      0.370
```

Note that you cannot conduct chi-square difference tests with the rescaled chi-square in the usual way. The difference test needs to be weighted by the rescaling factor. With your example files, I have included an Excel spreadsheet that does this calculation.

Categorical Measured Variables

It is important to distinguish between categorical variables and continuous variables. Categorical variables are those with two values (i.e., binary, dichotomous) or those with a few ordered categories (say 3 to 5). Examples might include gender, dead vs. alive, audited vs. not audited, or variables with few response options like “never,” “sometimes,” or “always.” Continuous variables are variables measured on a ratio or interval scale, such as temperature, height, or income in dollars. Ordinal variables with many categories, such as 7-point Likert-type scales of agreement, are usually treated as “continuous.”

ML estimation is generally not appropriate for binary dependent variables or dependent variables with few ordered categories, and special estimation techniques are needed. A dependent variable in this context is any variable by predicted other variables in the model and includes any indicator for a latent variable. A categorical independent variable not predicted by any variable in the data set (i.e., an exogenous variable) does not require any special treatment and can be modeled using traditional ML estimation.

There are two common ways of estimating models when one or more dependent variables are binary or ordinal. The first method, which is available in some other statistical packages, such as Lisrel, is an analysis of polychoric correlation matrices (see Technical Note # 3) using a weighted least squares (WLS) estimator. This method has computational limitations when there are many variables in the model. The second method, an analysis only available in Mplus, is Muthen’s categorical variable model (CVM) estimation. CVM estimation uses a process similar to the polychoric correlation matrix approach but has computational and statistical advantages (it performs well with small sample sizes and larger models). In Mplus, the CVM approach is invoked by using the CATEGORICAL option on the VARIABLE command (make sure you use the default estimation method of WLSMV by omitting the TYPE statement under the ANALYSIS paragraph).

When the binary variables are not latent variable indicators, one can obtain logistic regression estimates by using the CATEGORICAL statement together with TYPE=LOGISTIC, ESTIMATOR=ML, or ESTIMATOR=MLR on the ANALYSIS command. If the variable listed on the CATEGORICAL statement has more than two categories, Mplus assumes it is an ordinal variable and generates an ordinal logistic regression. Probit regression estimates for binary outcomes are obtained when the CATEGORICAL statement is used and neither TYPE=LOGISTIC nor ESTIMATOR=ML or MLR are specified.

When binary variables are used as latent variable indicators, the CVM approach should be used by specifying those variables on the CATEGORICAL statement and using the default estimation procedure (i.e., omit the TYPE statement).

Alternative Estimation Approaches

Maximum likelihood (ML) is by far the most common estimation used for structural equation modeling (and, indeed, is the default in all packages), but there are a number of alternatives to ML available in Mplus.

I will not review all the estimation procedures here or discuss their details (see Technical Note # 3), but here is a guide that I think will be helpful when fitting Mplus options in with the literature. This is a simplification and embellishment of information on p. 366 of the Mplus users guide (Chapter 15).

Estimator	Purpose	Comment	Mplus Specification
ML	Continuous normal variables	Most widely used estimator in SEM.	ESTIMATOR=ML; (default for GENERAL models without any categorical variables)
MLM	Non-normal continuous variables	Mean-adjusted maximum likelihood. Produces Satorra-Bentler scaled chi-square and robust standard errors (Satorra & Bentler, 1988; 1994)	ESTIMATOR=MLM;
MLMV	Non-normal continuous variables, but less commonly used	Maximum likelihood with mean and variance-adjusted chi-square and robust standard errors. A (less preferable) alternative to MLM.	ESTIMATOR=MLMV;
MLR	Non-normal continuous variables with missing data	Maximum likelihood robust. Yuan-Bentler (2000) robust estimator for missing data (sometimes referred to the sandwich estimator)	ESTIMATOR=MLR; (default if TYPE=MISSING or if TYPE=LOGISTIC)
MLF	Not commonly used	Generates approximate standard errors (using first order derivatives) and traditional chi-square	ESTIMATOR=MLF;
MUML	Possible estimator for multilevel models	Muthen's limited information estimator	ESTIMATOR=MUML;
WLS	Not commonly used	Also known as ADF or AGLS elsewhere. Can be used for categorical data, but requires many cases (e.g., approx 5,000) and simple models	ESTIMATOR=WLS;
WLSM	Not commonly used	Weighted least squares with mean-adjusted chi-square and robust standard errors. Diagonal weight matrix is used for parameter estimation. Sometimes referred to as a "diagonally weighted least squares" estimator.	ESTIMATOR=WLSM;
WLSMV	One or more categorical dependent variables (binary or ordinal).	Weighted least squares with mean and variance-adjusted chi-square and robust standard errors. Provides Muthen's CVM estimation. Sometimes referred to as a "diagonally weighted least squares" estimator.[5]	ESTIMATOR=WLSMV; (default if CATEGORICAL option used on the VARIABLE command)
GLS	Rarely used in practice	General estimator, of which ML is a special case	ESTIMATOR=GLS;
ULS	Rarely used in SEM models	Unweighted least squares. Simple least squares estimator. Could be used to generate starting values if a model does not converge or has other estimation difficulties. Used in Mplus' EFA.	ESTIMATOR=ULS; (default for TYPE=EFA)

[5] Note that models that use categorical outcomes estimated by MLSMV require that data are MCAR, not just MAR.

Technical Note #3 : Alternative Estimation Methods

ML

Remember that the usual approach to estimating fit and coefficients in SEM is the maximum likelihood (ML) approach. ML uses derivatives to minimize the following fit function:

$$F_{ML} = \log\left|\Sigma(\theta)\right| + tr\left(S\Sigma^{-1}(\theta)\right) - \log|S| - (p+q)$$

The ML estimator assumes that the variables in the model are multivariate normal (i.e., the joint distribution of the variables is distributed normally).

GLS

Generalized least squares is an alternative fitting function. The GLS fit function also minimizes the discrepancy between S and Σ, but uses a weight matrix for the residuals, designated W.

$$F_{GLS} = \left(\frac{1}{2}\right) tr\left(\left\{\left[S - \Sigma(\theta)W^{-1}\right]\right\}^2\right)$$

Notice that this is a much simpler function (e.g., no logs), and it is clear that the discrepancy between the obtained covariance matrix and the covariance matrix implied by the model (S-Σ) is minimized after weighting it by W. Although any W can be chosen for the weight matrix, most commonly, the inverse of the covariance matrix, S, is used in SEM packages. F_{GLS} is asymptotically equivalent to F_{ML}, meaning that as sample sizes increase, they are approximately equal. F_{GLS} is based on the same assumptions as F_{ML} and would be used under the same conditions. It is thought to perform less well, however, in small samples, so F_{ML} is usually chosen instead of F_{GLS}. The simplicity of the function, however, means that other weight matrices could be used in an attempt to correct for violations of distributional assumptions.

ADF

The asymptotic distribution free function was developed by Browne (1984). It is described as arbitrary generalized least squares (AGLS) by Bentler in the EQS package and weighted least squares (WLS) by Joreskog and Sorbom in Lisrel. The main advantage of the ADF estimator is that it does not require multivariate normality. The ADF estimator is based on the F_{GLS}, except a different W is chosen. It can be written in a general form that encompasses GLS, ML, and ULS (not discussed here) where the difference depends on the choice of W:

$$F_{ADF} = F_{AGLS} = F_{WLS} = (s-\sigma)' W^{-1} (s-\sigma)$$

W used in F_{ADF} is based on a covariance of all of the elements of the covariance matrix, S. That is, a covariance matrix is constructed that estimates the covariances between each s_{ij} element of S, and is therefore a ½[v(v+1)] by ½[v(v+1)] matrix. The reason for this is that these "covariances of covariances" are the related to kurtosis estimates (so called "fourth-order moments"). So, the GLS fit function is weighted by variances and kurtosis in attempt to correct for violations of the normality assumption. Another way of saying this is that when the data are normal, the ADF estimator reduces to GLS because there is no kurtosis. The large weight matrix causes serious practical difficulties when there is a large number of variables in the model (e.g., more that 20 or so), and computer packages (e.g., EQS) do not allow estimation unless the number of cases is equal or greater than number of elements in the weight matrix (i.e., ½[v(v+1)] times ½[v(v+1)] divided by 2). Simulations studies suggest that chi-square values are severely overestimated with small samples and that sample sizes of about 5000 are necessary for good estimates. A recent study by Olsson, Foss, Troye, and Howell (2000) suggests that ADF estimation performs poorly when the model is misspecified. Combined with the limitation of variables, this is usually seen as an unattractive approach when nonnormality exists.

Muthen's CVM

Muthen (1993) suggested a categorical variable model (CVM) for use when models measured variables are categorical (either dichotomous or ordered categorical). Models with categorical variables are always considered to be in violation of the normality assumption and, thus, the usual F_{ML} estimator is not recommended. The CVM approach uses the general ADF function (which Muthen and Lisrel call WLS), but does not have a practical limit on the number of variables nor require such large samples, because it avoids inversion of the large weight matrix (using something called "Taylor expansion"). The idea behind the method is that categorical variables have an underlying continuous latent variable, called y*. y* is estimated by *polychoric* correlations which correct for loss of information in Pearson correlations due to categorization of a continuous variable (See MacCallum, Zhang, Preacher, & Rucker, 2002). *Tetrachoric* correlations are a special case of polychoric correlations involving only binary variables, and *polyserial* correlations are those involving the correlation between a binary and a continuous variable. The polychoric correlations are then used to estimate the model using the F_{WLS} estimator. Mplus has special features that implement the CVM approach. A similar approach is available in Lisrel by creating a polychoric correlation matrix in Prelis (the Lisrel preprocessor) and then analyzing the new matrix in Lisrel with WLS (not ML). The CVM approach with Mplus is a simpler process and is able to avoid inversion of the large W matrix.

Missing Data

Missing Data and Missing Data Estimation

Listwise Deletion

Until recently, listwise deletion has been the most common way of dealing with missing data in SEM. That is, complete data was required on all variables in the analysis—any cases with missing data on one or more of the variables was eliminated from the analysis. In the last few years, however, researchers have begun to use data estimation techniques when there are missing data among the variables in a structural model. And simulation data convincingly shows that when there are a lot of missing data, listwise deletion will have biased parameters and standard errors.

MAR and MCAR

A distinction of the type of missing data was made by Rubin (1976), who classified missing data as missing at random (MAR), missing completely at random (MCAR), or neither.[6] Both MAR and MCAR require that the variable with missing data be unrelated to whether or not a person has missing data on that variable. For example, if those with lower incomes are more likely to have missing data on the income variable, the data cannot be MAR or MCAR. The difference between MAR and MCAR is whether or not other variables in the data set are associated with whether or not someone has missing data on a particular variable. For example, are older people more likely to refuse to respond to the income variable? The term MAR is confusing because data are not really missing at random, because missingness does depend on some of the variables in the data set.

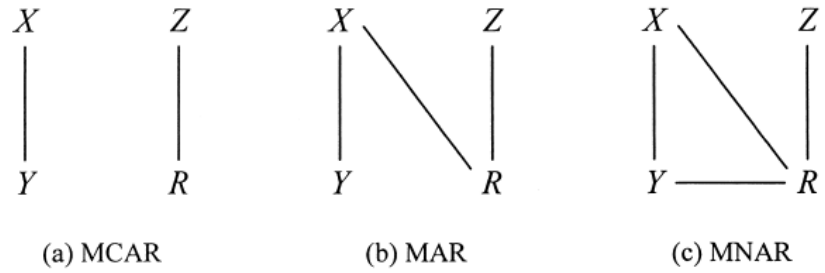


Figure 2. Graphical representations of (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR) in a univariate missing-data pattern. *X* represents variables that are completely observed, *Y* represents a variable that is partly missing, *Z* represents the component of the causes of missingness unrelated to *X* and *Y*, and *R* represents the missingness.

From Schafer, J. L. & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

FIML

Probably the best missing data estimation approach is full information maximum likelihood (FIML), which has been shown to produce unbiased parameter estimates and standard errors under MAR and MCAR. FIML, sometimes called "raw maximum likelihood" or just "ML," is currently available in Amos, Mplus, and Mx. It requires that data be at least MAR (i.e., either MAR or MCAR are ok). FIML works by estimating a

[6] Muthen uses the term “non-ignorable missing data” to describe anything not MCAR or MAR.

likelihood function for each individual based on the variables that are present so that all the available data are used. For example, there may be some variables with data for all 389 cases but some variables may have data for only 320 of the cases. The fitting function for FIML is computed by summing all the individual fit functions, and, thus, it is able to use all 389 cases. Rather than the traditional approach to calculating chi-square, FIML estimates two models, the H_0 model and the H_1 model. The H_0 model is the "unrestricted" model, meaning that all variables are correlated. The H_1 model is the specified model. The difference between the two log-likelihoods is used to derive the chi-square. This approach allows one to use all the available information in the variables.

Determining If Missing Data are at Least MAR

Practically speaking, it is quite difficult to determine if your data are at least MAR. With a single variable that has missing data, it is not too difficult to determine if any of the other variables in the data set predict whether there are missing data on a particular variable. In practice, however, data will be missing on a number of variables, and so determining if other variables are related may be considerably complex. But the real importance is determining if missingness is associated with values of the variables that are missing data. Determining whether data are at least MAR may be quite difficult to do. In a recent discussion of missing data estimation, Schafer and Graham (2002) state: "When missingness is beyond the researcher's control, its distribution is unknown and MAR is only an assumption. In general, there is no way to test whether MAR holds in a data set, except by obtaining follow-up data from nonrespondents or by imposing an unverifiable model." (p. 152). There may be some ways to try to explore the issue, however. With attrition over time, it may be possible to test whether missingness is associated with the value of the variable by examining whether the variable at Time 1 (i.e., with complete data) is associated with the missingness for that variable at Time 2. If data are missing on individual items from a scale, an approximate approach might be to attempt to show that missingness on particular items is unrelated to scale scores for that measure. In other circumstances, one may have to provide a theoretical argument that missingness is not associated with the variable or rely on information in the literature. There are many writings on missing data estimation, but few on how to go about determining if data are at least MAR.

Multigroup SEM Approach

Another approach to missing data analysis uses a multigroup structural model approach, suggested by Muthen, Kaplan, and Hollis (1987). The same model is estimated in different groups. The groups are based on different patterns of missing data—one group for each pattern. A few hand calculations must be done. This is a fairly impractical approach if there are many patterns of missing data, but might be especially useful if data are missing by design.

Pairwise Deletion

Pairwise deletion is sometimes used to estimate models when there are missing data. With pairwise deletion, a covariance (or correlation) matrix is computed where each element is based on the full number of cases with complete data for each pair of

variables. This approach may lead to nonpositive definite matrices and to standardized values over 1. There are other potential problems with the approach and I do not recommend it.

Other Imputation Methods

There are several other estimation approaches in which the data are imputed. That is, a full data set is created based on the imputation method that fills in data based on information from existing data. Some examples are: mean imputation (the average scores is filled in), regression-based methods (a regression is used to predict a score), resemblance-based "hot-deck imputation" (which imputes new values from similar cases), and Expectation Maximization (EM; which is a maximum likelihood-based approach). The regression method and the EM approach build in some error (so that the imputed values are not perfectly correlated with the existing data). Of the two, the EM approach seems to perform the best. The EM approach requires that data are at least MAR.

Comments

Particularly when there is a large amount of missing data, researchers are better off using a FIML approach to estimation. Given that it is fairly easy to implement in the packages where it is available, there is no reason *not* to do it. In the conclusions of their paper, Schafer and Graham suggest that under many circumstances there may be advantages to missing data estimation relative to listwise deletion even when data are not MAR. What is a large amount of missing data? The percentage of missing data is sometimes discussed based on the percentage missing for a certain variable. It makes more sense to me to examine the percentage of cases missing if listwise deletion were to be used. With this method, data sets (i.e., the set of variables in the model) in which more than roughly 20% of the cases are excluded by listwise deletion seem to lead to substantial bias in estimates (e.g., Arbuckle, 1996). With fewer than this much missing data, it may not be a major difference whether listwise deletion or FIML is used.

Example 9: Missing Data Estimation

By default, Mplus uses listwise deletion whenever there are missing data present. FIML estimation can be obtained, however, simply by including a TYPE=MISSING H1 statement on the analysis command in Mplus. The H1 statement is needed to obtain an overall model chi-square.[7] Below is syntax for a re-estimation of the full SEM model used in Example 7. In that example, there were only 1832 cases when listwise deletion was used. The new analysis uses 3001 cases, suggesting about a 40% loss of data without the missing data estimation.

```
TITLE:  Example 9, full SEM with missing data estimation;

DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
               FORMAT=FREE;

VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
    b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
   b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
  b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
    b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
           b3v3pneg b1p_conf b2p_conf b3p_conf;

    MISSING = program-b3p_conf(-99,-6--1);

    USEVARIABLES=b1pc04a-b1pc04t b1v3pdet b1v3pint b1v3pneg
             mage black hisp;

DEFINE:  IF (race EQ 1 OR race GE 3) THEN black = 0;
         IF (race EQ 2) THEN black = 1;
         IF (race EQ 1 OR race EQ 2 OR race EQ 4) THEN hisp = 0;
         IF (race EQ 3) THEN hisp = 1;

ANALYSIS:
        TYPE = MISSING H1; ESTIMATOR = ML; MATRIX = COVARIANCE;

MODEL:  cesd1som BY b1pc04a b1pc04b b1pc04e b1pc04g b1pc04k
                  b1pc04m b1pc04t;
        cesd1neg BY b1pc04c b1pc04f b1pc04j b1pc04n b1pc04r;
        b1v3pdet b1v3pint b1v3pneg ON cesd1som cesd1neg
                  mage black hisp;
        cesd1som with cesd1neg mage black hisp;
        cesd1neg with mage black hisp;
        mage with black hisp;
        black with hisp;
        b1v3pdet with b1v3pint b1v3pneg;
        b1v3pint with b1v3pneg;

OUTPUT:  STANDARDIZED;
```

[7] Without this statement, Mplus only provides a likelihood value, testing what is sometimes referred to as the H0 model. The likelihood value from that model can be used for comparisons with other models, but does not give an overall chi-square for the model. With a lot of missing data, computation can be intensive and convergence is sometimes not achieved. In those instances, omitting the H1 statement might be useful to get model results (if not only for the purpose of identifying model problems etc.).

Example 9 Output: Missing Data Estimation

```
INPUT READING TERMINATED NORMALLY

Example 9, full SEM with missing data estimation;

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        3001

Number of dependent variables                                   15
Number of independent variables                                  3
Number of continuous latent variables                            2

Observed dependent variables

  Continuous
   B1PC04A     B1PC04B     B1PC04C     B1PC04E     B1PC04F     B1PC04G
   B1PC04J     B1PC04K     B1PC04M     B1PC04N     B1PC04R     B1PC04T
   B1V3PDET    B1V3PINT    B1V3PNEG

Observed independent variables
   MAGE        BLACK       HISP

Continuous latent variables
   CESD1SOM    CESD1NEG

Estimator                                                       ML
Information matrix                                        OBSERVED
Maximum number of iterations                                  1000
Convergence criterion                                    0.500D-04
Maximum number of steepest descent iterations                   20
Maximum number of iterations for H1                           2000
Convergence criterion for H1                             0.100D-03

Input data file(s)
  c:\jason\mplus\ehs\ex2.dat

Input data format  FREE

SUMMARY OF DATA

     Number of patterns          37

COVARIANCE COVERAGE OF DATA

Minimum covariance coverage value   0.100

    PROPORTION OF DATA PRESENT


           Covariance Coverage
              B1PC04A       B1PC04B       B1PC04C       B1PC04E       B1PC04F
              ________      ________      ________      ________      ________
 B1PC04A        0.767
 B1PC04B        0.766         0.767
 B1PC04C        0.766         0.765         0.766
 B1PC04E        0.766         0.766         0.765         0.766
 B1PC04F        0.766         0.766         0.765         0.766         0.766
 B1PC04G        0.765         0.764         0.764         0.764         0.764
 B1PC04J        0.764         0.764         0.764         0.764         0.764
 B1PC04K        0.765         0.765         0.764         0.765         0.765
 B1PC04M        0.760         0.760         0.759         0.760         0.760
 B1PC04N        0.765         0.765         0.764         0.765         0.765
 B1PC04R        0.764         0.764         0.763         0.764         0.764
 B1PC04T        0.765         0.765         0.764         0.765         0.765
 B1V3PDET       0.635         0.634         0.634         0.634         0.634
 B1V3PINT       0.635         0.634         0.634         0.634         0.634
 B1V3PNEG       0.635         0.634         0.634         0.634         0.634
 MAGE           0.765         0.765         0.765         0.765         0.765
```

BLACK	0.754	0.754	0.754	0.753	0.753
HISP	0.754	0.754	0.754	0.753	0.753

Covariance Coverage

	B1PC04G	B1PC04J	B1PC04K	B1PC04M	B1PC04N
B1PC04G	0.765				
B1PC04J	0.763	0.765			
B1PC04K	0.764	0.764	0.766		
B1PC04M	0.759	0.759	0.760	0.761	
B1PC04N	0.764	0.764	0.765	0.760	0.766
B1PC04R	0.762	0.763	0.764	0.758	0.764
B1PC04T	0.764	0.764	0.765	0.760	0.765
B1V3PDET	0.633	0.634	0.634	0.631	0.634
B1V3PINT	0.633	0.634	0.634	0.631	0.634
B1V3PNEG	0.633	0.634	0.634	0.631	0.634
MAGE	0.764	0.763	0.764	0.759	0.764
BLACK	0.752	0.752	0.753	0.748	0.753
HISP	0.752	0.752	0.753	0.748	0.753

Covariance Coverage

	B1PC04R	B1PC04T	B1V3PDET	B1V3PINT	B1V3PNEG
B1PC04R	0.764				
B1PC04T	0.764	0.766			
B1V3PDET	0.633	0.634	0.652		
B1V3PINT	0.633	0.634	0.652	0.652	
B1V3PNEG	0.633	0.634	0.652	0.652	0.652
MAGE	0.763	0.764	0.651	0.651	0.651
BLACK	0.751	0.753	0.641	0.641	0.641
HISP	0.751	0.753	0.641	0.641	0.641

Covariance Coverage

	MAGE	BLACK	HISP
MAGE	0.998		
BLACK	0.978	0.980	
HISP	0.978	0.980	0.980

THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

Value	642.380
Degrees of Freedom	113
P-Value	0.0000

Chi-Square Test of Model Fit for the Baseline Model

Value	9296.714
Degrees of Freedom	150
P-Value	0.0000

CFI/TLI

CFI	0.942
TLI	0.923

Loglikelihood

H0 Value	-53479.484
H1 Value	-53158.294

Information Criteria

Number of Free Parameters	76
Akaike (AIC)	107110.969
Bayesian (BIC)	107567.478
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	107325.997

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.040

```
          90 Percent C.I.                    0.037  0.043
          Probability RMSEA <= .05           1.000

SRMR (Standardized Root Mean Square Residual)
          Value                              0.032

MODEL RESULTS

                    Estimates     S.E.  Est./S.E.    Std     StdYX
 CESD1SOM BY
    B1PC04A             1.000    0.000      0.000    0.522    0.611
    B1PC04B             0.835    0.044     18.956    0.436    0.479
    B1PC04E             1.067    0.049     21.850    0.557    0.568
    B1PC04G             0.774    0.053     14.488    0.404    0.353
    B1PC04K             1.117    0.053     21.182    0.583    0.560
    B1PC04M             0.775    0.043     18.151    0.405    0.452
    B1PC04T             1.104    0.048     22.837    0.576    0.623

 CESD1NEG BY
    B1PC04C             1.000    0.000      0.000    0.593    0.701
    B1PC04F             1.244    0.036     34.389    0.737    0.802
    B1PC04J             0.716    0.029     24.296    0.424    0.559
    B1PC04N             1.057    0.036     28.976    0.627    0.675
    B1PC04R             1.141    0.035     33.059    0.676    0.780

 B1V3PDET ON
    CESD1SOM            0.117    0.151      0.778    0.061    0.061
    CESD1NEG           -0.032    0.126     -0.251   -0.019   -0.019

 B1V3PINT ON
    CESD1SOM            0.115    0.183      0.631    0.060    0.049
    CESD1NEG            0.063    0.152      0.414    0.037    0.030

 B1V3PNEG ON
    CESD1SOM            0.095    0.116      0.820    0.050    0.063
    CESD1NEG            0.000    0.097     -0.004    0.000    0.000

 B1V3PDET ON
    MAGE               -0.016    0.004     -3.907   -0.016   -0.088
    BLACK               0.369    0.056      6.622    0.369    0.175
    HISP                0.155    0.069      2.239    0.155    0.066

 B1V3PINT ON
    MAGE               -0.015    0.005     -3.108   -0.015   -0.069
    BLACK               0.714    0.066     10.762    0.714    0.275
    HISP                0.505    0.083      6.119    0.505    0.174

 B1V3PNEG ON
    MAGE               -0.010    0.003     -3.256   -0.010   -0.072
    BLACK               0.466    0.042     11.020    0.466    0.280
    HISP                0.124    0.053      2.364    0.124    0.067

 CESD1SOM WITH
    CESD1NEG            0.262    0.014     19.188    0.847    0.847
    MAGE               -0.235    0.071     -3.299   -0.450   -0.080
    BLACK              -0.004    0.006     -0.740   -0.009   -0.018
    HISP               -0.039    0.006     -6.946   -0.074   -0.175

 CESD1NEG WITH
    MAGE               -0.078    0.075     -1.038   -0.132   -0.023
    BLACK              -0.001    0.006     -0.207   -0.002   -0.005
    HISP               -0.003    0.006     -0.436   -0.004   -0.010

 MAGE     WITH
    BLACK              -0.496    0.051     -9.808   -0.496   -0.185
    HISP                0.302    0.045      6.741    0.302    0.126

 BLACK    WITH
    HISP               -0.082    0.004    -20.349   -0.082   -0.405

 B1V3PDET WITH
```

```
    B1V3PINT           0.129    0.027      4.854    0.129    0.104
    B1V3PNEG           0.191    0.017     10.952    0.191    0.241

 B1V3PINT WITH
    B1V3PNEG           0.358    0.022     16.361    0.358    0.367

 Means
    MAGE              22.657    0.103    219.704   22.657    4.014
    BLACK              0.347    0.009     39.559    0.347    0.729
    HISP               0.236    0.008     30.162    0.236    0.556

 Intercepts
    B1PC04A            1.660    0.018     93.358    1.660    1.942
    B1PC04B            1.683    0.019     88.761    1.683    1.848
    B1PC04C            1.537    0.018     87.239    1.537    1.819
    B1PC04E            1.907    0.020     93.464    1.907    1.945
    B1PC04F            1.657    0.019     86.489    1.657    1.803
    B1PC04G            2.278    0.024     95.261    2.278    1.986
    B1PC04J            1.417    0.016     89.414    1.417    1.865
    B1PC04K            1.961    0.022     90.449    1.961    1.883
    B1PC04M            1.611    0.019     86.024    1.611    1.797
    B1PC04N            1.651    0.019     85.272    1.651    1.778
    B1PC04R            1.673    0.018     92.482    1.673    1.929
    B1PC04T            1.739    0.019     90.327    1.739    1.880
    B1V3PDET           1.820    0.103     17.671    1.820    1.811
    B1V3PINT           2.476    0.124     20.029    2.476    2.007
    B1V3PNEG           1.501    0.079     18.979    1.501    1.898

 Variances
    MAGE              31.856    0.823     38.698   31.856    1.000
    BLACK              0.227    0.006     38.347    0.227    1.000
    HISP               0.180    0.005     38.336    0.180    1.000
    CESD1SOM           0.273    0.019     14.739    1.000    1.000
    CESD1NEG           0.352    0.019     18.211    1.000    1.000

 Residual Variances
    B1PC04A            0.458    0.016     29.416    0.458    0.627
    B1PC04B            0.639    0.020     31.668    0.639    0.771
    B1PC04C            0.363    0.013     28.891    0.363    0.508
    B1PC04E            0.650    0.021     30.333    0.650    0.677
    B1PC04F            0.301    0.012     24.561    0.301    0.356
    B1PC04G            1.152    0.035     32.798    1.152    0.876
    B1PC04J            0.397    0.013     31.584    0.397    0.688
    B1PC04K            0.744    0.025     30.274    0.744    0.686
    B1PC04M            0.639    0.020     31.818    0.639    0.796
    B1PC04N            0.469    0.016     29.611    0.469    0.544
    B1PC04R            0.294    0.011     25.852    0.294    0.391
    B1PC04T            0.523    0.018     28.866    0.523    0.612
    B1V3PDET           0.971    0.031     31.204    0.971    0.961
    B1V3PINT           1.402    0.045     31.210    1.402    0.921
    B1V3PNEG           0.574    0.018     31.215    0.574    0.918

R-SQUARE
    Observed
    Variable  R-Square

    B1PC04A      0.373
    B1PC04B      0.229
    B1PC04C      0.492
    B1PC04E      0.323
    B1PC04F      0.644
    B1PC04G      0.124
    B1PC04J      0.312
    B1PC04K      0.314
    B1PC04M      0.204
    B1PC04N      0.456
    B1PC04R      0.609
    B1PC04T      0.388
    B1V3PDET     0.039
    B1V3PINT     0.079
    B1V3PNEG     0.082
```

Longitudinal Models

Longitudinal Cross-lagged Models

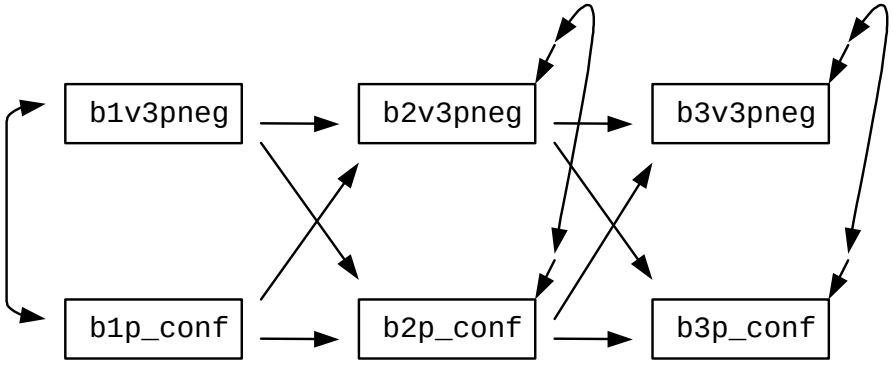
With longitudinal data of two waves or more, a useful strategy to investigate causal directionality between two variables is the *cross-lagged panel* model.

- The model examines two variables, each predicting the other across time.
- Paths from one variable to another are called *cross-lagged paths*. Cross-lagged paths can be interpreted as prediction of the change in the dependent variable, because the initial value of the dependent variable is controlled.
- A path from the same variable to itself over time is called a *stability path*.
- Generally, correlations among exogenous variables and among endogenous disturbances are estimated.
- With latent variables, correlated measurement errors over time must be used to avoid overestimation of longitudinal causal paths.
- Because initial levels of each variable are controlled, this is a powerful design for investigating causal relationships with passive observational data.

The most basic cross-lagged panel model uses only measured variables at each time point.

- For two waves, this model is just identified (df=0), so no information about fit is available.
- Equality constraints can be added, forcing the same paths to be equal at different time points (e.g., y1 → y2 equals y2 → y3).

In Example 10 below, I examine whether negative regard leads to conflict or conflict leads to negative regard. The following picture illustrates the model:



Example 10: Cross-lagged Panel Model with Measured Variables

Mplus input for a cross-lagged is shown below. Because the variables used in this model are non-normal and have missing data, I illustrate the MLR estimator which is preferable in this case. The WITH statement is optional, because the synchronous correlations are estimated by default.

```
TITLE:  Example 10, Cross-lagged Panel Model with Measured Variables;

DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
               FORMAT=FREE;

VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
    b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
    b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
   b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
    b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
            b3v3pneg b1p_conf b2p_conf b3p_conf;

    MISSING = program-b3p_conf(-99,-6--1);

    USEVARIABLES=b1v3pneg b2v3pneg b3v3pneg
                 b1p_conf b2p_conf b3p_conf;

ANALYSIS:
   TYPE = MISSING H1; ESTIMATOR = MLR; MATRIX = COVARIANCE;


MODEL:  ! Stability paths;
        b2v3pneg ON b1v3pneg;
        b3v3pneg ON b2v3pneg;
        b2p_conf ON b1p_conf;
        b3p_conf ON b2p_conf;
        !Cross-lagged paths;
        b2p_conf ON b1v3pneg;
        b3p_conf ON b2v3pneg;
        b2v3pneg ON b1p_conf;
        b3v3pneg ON b2p_conf;
        ! Synchronous Correlations;
        b1v3pneg WITH b1p_conf;
        b2v3pneg WITH b2p_conf;
        b3v3pneg WITH b3p_conf;


OUTPUT:  STANDARDIZED;
```

Output for Example 10: Cross-lagged panel model with measured variables

```
*** WARNING
  Data set contains cases with missing on all variables.
  These cases were not included in the analysis.
  Number of cases with missing on all variables:  392
   1 WARNING(S) FOUND IN THE INPUT INSTRUCTIONS


Example 10, Cross-lagged Panel Model with Measured Variables;

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        2609

Number of dependent variables                                    4
Number of independent variables                                  2
Number of continuous latent variables                            0

Observed dependent variables

  Continuous
   B2V3PNEG    B3V3PNEG    B2P_CONF    B3P_CONF

Observed independent variables
   B1V3PNEG    B1P_CONF

Estimator                                                      MLR
Information matrix                                        OBSERVED
Maximum number of iterations                                 1000
Convergence criterion                                   0.500D-04
Maximum number of steepest descent iterations                  20
Maximum number of iterations for H1                          2000
Convergence criterion for H1                            0.100D-03

Input data file(s)
  c:\jason\mplus\ehs\ex2.dat

Input data format  FREE


SUMMARY OF DATA

     Number of patterns          63


COVARIANCE COVERAGE OF DATA

Minimum covariance coverage value   0.100


     PROPORTION OF DATA PRESENT


           Covariance Coverage
              B2V3PNEG      B3V3PNEG      B2P_CONF      B3P_CONF      B1V3PNEG
              ________      ________      ________      ________      ________
 B2V3PNEG       0.688
 B3V3PNEG       0.522         0.635
 B2P_CONF       0.590         0.496         0.711
 B3P_CONF       0.524         0.549         0.560         0.701
 B1V3PNEG       0.570         0.512         0.553         0.536         0.750
 B1P_CONF       0.538         0.484         0.576         0.550         0.618


           Covariance Coverage
              B1P_CONF
              ________
 B1P_CONF       0.744
```

THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

Value	53.558*
Degrees of Freedom	4
P-Value	0.0000
Scaling Correction Factor for MLR	1.433

* The chi-square value for MLM, MLMV, MLR, WLSM and WLSMV cannot be used for chi-square difference tests. MLM, MLR and WLSM chi-square difference testing is described in the Mplus Technical Appendices at www.statmodel.com. See chi-square difference testing in the index of the Mplus User's Guide.

Chi-Square Test of Model Fit for the Baseline Model

Value	586.470
Degrees of Freedom	14
P-Value	0.0000

CFI/TLI

CFI	0.913
TLI	0.697

Loglikelihood

H0 Value	-10168.283
H1 Value	-10129.918

Information Criteria

Number of Free Parameters	23
Akaike (AIC)	20382.566
Bayesian (BIC)	20517.501
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	20444.423

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.069

SRMR (Standardized Root Mean Square Residual)

Value	0.038

MODEL RESULTS

	Estimates	S.E.	Est./S.E.	Std	StdYX
B2V3PNEG ON					
B1V3PNEG	0.269	0.040	6.747	0.269	0.256
B1P_CONF	0.139	0.050	2.792	0.139	0.091
B3V3PNEG ON					
B2V3PNEG	0.270	0.036	7.503	0.270	0.366
B2P_CONF	0.037	0.031	1.209	0.037	0.033
B2P_CONF ON					
B1P_CONF	0.356	0.029	12.445	0.356	0.356
B1V3PNEG	0.042	0.019	2.237	0.042	0.061
B3P_CONF ON					
B2P_CONF	0.418	0.029	14.385	0.418	0.424
B2V3PNEG	-0.008	0.018	-0.430	-0.008	-0.012

```
 B1V3PNEG WITH
    B1P_CONF           0.009    0.011      0.856    0.009    0.022

 B2V3PNEG WITH
    B2P_CONF           0.014    0.011      1.233    0.014    0.030

 B3V3PNEG WITH
    B3P_CONF          -0.002    0.008     -0.234   -0.002   -0.006

 Means
    B1V3PNEG           1.451    0.017     82.920    1.451    1.839
    B1P_CONF           1.722    0.012    141.131    1.722    3.174

 Intercepts
    B2V3PNEG           0.804    0.100      8.067    0.804    0.969
    B3V3PNEG           0.831    0.070     11.835    0.831    1.360
    B2P_CONF           1.034    0.055     18.966    1.034    1.903
    B3P_CONF           0.969    0.054     18.073    0.969    1.812

 Variances
    B1V3PNEG           0.623    0.047     13.318    0.623    1.000
    B1P_CONF           0.294    0.012     24.822    0.294    1.000

 Residual Variances
    B2V3PNEG           0.636    0.049     13.035    0.636    0.925
    B3V3PNEG           0.323    0.030     10.898    0.323    0.863
    B2P_CONF           0.256    0.011     22.765    0.256    0.869
    B3P_CONF           0.235    0.011     21.710    0.235    0.821


R-SQUARE

    Observed
    Variable  R-Square

    B2V3PNEG     0.075
    B3V3PNEG     0.137
    B2P_CONF     0.131
    B3P_CONF     0.179
```

Example 11 Cross-lagged Panel Model with Latent Variables

In this example, I examine a two-wave cross-lagged model using one latent variable. I use negative regard, detachment, and intrusiveness as indicators for a latent factor. Naturally, two latent variables or more than two waves can be used for this model as well.

In the example below, I illustrate the use of equality constraints. An important first step, which is not shown here, involves chi-square difference tests to see if these constraints on the loadings are appropriate (i.e., *longitudinal invariance*).

It is also critical that correlated measurement errors be estimated for same items over time.

```
TITLE:  Example 11, Cross-lagged Panel Model with Measured Variables;

DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
              FORMAT=FREE;

VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
    b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
           b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
          b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
            b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
       b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
                   b3v3pneg b1p_conf b2p_conf b3p_conf;

    MISSING = program-b3p_conf(-99,-6--1);

    USEVARIABLES=b1v3pdet b1v3pint b1v3pneg
       b3v3pdet b3v3pint b3v3pneg b1p_conf b3p_conf;

ANALYSIS:
   TYPE = MISSING H1; ESTIMATOR = MLR; MATRIX = COVARIANCE;


MODEL:  ! Measurement model for 3-bag measure;
        bag1 BY b1v3pdet*1 (1);
        bag1 BY b1v3pint*1 (2);
        bag1 BY b1v3pneg@1 (3);
        bag3 BY b3v3pdet*1 (1);
        bag3 BY b3v3pint*1 (2);
        bag3 BY b3v3pneg@1 (3);

        ! Stability paths;
        bag3 ON bag1;
        b3p_conf ON b1p_conf;

        !Cross-lagged paths;
        bag3 ON b1p_conf;
        b3p_conf on bag1;

        ! Synchronous Correlations;
        bag1 WITH b1p_conf;
        bag3 WITH b3p_conf;

        ! Correlated Measurement Errors Over Time;
         b1v3pdet b1v3pint b1v3pneg PWITH b3v3pdet b3v3pint b3v3pneg;


OUTPUT:  STANDARDIZED;
```

Output for Example 11: Cross-lagged Panel Model with Latent Variables

```
*** WARNING
  Data set contains cases with missing on all variables.
  These cases were not included in the analysis.
  Number of cases with missing on all variables:  463
   1 WARNING(S) FOUND IN THE INPUT INSTRUCTIONS


Example 11, Cross-lagged Panel Model with Measured Variables;

SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        2538

Number of dependent variables                                    7
Number of independent variables                                  1
Number of continuous latent variables                            2

Observed dependent variables

  Continuous
   B1V3PDET    B1V3PINT    B1V3PNEG    B3V3PDET    B3V3PINT    B3V3PNEG
   B3P_CONF

Observed independent variables
   B1P_CONF

Continuous latent variables
   BAG1        BAG3

Estimator                                                      MLR
Information matrix                                        OBSERVED
Maximum number of iterations                                  1000
Convergence criterion                                    0.500D-04
Maximum number of steepest descent iterations                   20
Maximum number of iterations for H1                           2000
Convergence criterion for H1                             0.100D-03

Input data file(s)
  c:\jason\mplus\ehs\ex2.dat

Input data format  FREE


SUMMARY OF DATA

     Number of patterns          16


COVARIANCE COVERAGE OF DATA

Minimum covariance coverage value   0.100


     PROPORTION OF DATA PRESENT


           Covariance Coverage
              B1V3PDET      B1V3PINT      B1V3PNEG      B3V3PDET      B3V3PINT
              ________      ________      ________      ________      ________
 B1V3PDET       0.771
 B1V3PINT       0.771         0.771
 B1V3PNEG       0.771         0.771         0.771
 B3V3PDET       0.527         0.527         0.527         0.654
 B3V3PINT       0.527         0.527         0.527         0.654         0.654
 B3V3PNEG       0.526         0.526         0.526         0.653         0.653
 B3P_CONF       0.551         0.551         0.551         0.565         0.565
 B1P_CONF       0.635         0.635         0.635         0.498         0.498
```

Covariance Coverage

	B3V3PNEG	B3P_CONF	B1P_CONF
B3V3PNEG	0.653		
B3P_CONF	0.565	0.721	
B1P_CONF	0.498	0.565	0.765

THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

Value	41.498*
Degrees of Freedom	15
P-Value	0.0003
Scaling Correction Factor for MLR	1.246

* The chi-square value for MLM, MLMV, MLR, WLSM and WLSMV cannot be used for chi-square difference tests. MLM, MLR and WLSM chi-square difference testing is described in the Mplus Technical Appendices at www.statmodel.com. See chi-square difference testing in the index of the Mplus User's Guide.

Chi-Square Test of Model Fit for the Baseline Model

Value	1109.655
Degrees of Freedom	28
P-Value	0.0000

CFI/TLI

CFI	0.976
TLI	0.954

Loglikelihood

H0 Value	-15568.382
H1 Value	-15542.538

Information Criteria

Number of Free Parameters	29
Akaike (AIC)	31194.763
Bayesian (BIC)	31364.098
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	31271.957

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.026

SRMR (Standardized Root Mean Square Residual)

Value	0.028

MODEL RESULTS

		Estimates	S.E.	Est./S.E.	Std	StdYX
BAG1	BY					
B1V3PDET		0.339	0.049	6.962	0.246	0.249
B1V3PINT		0.858	0.109	7.833	0.623	0.499
B1V3PNEG		1.000	0.000	0.000	0.727	0.921
BAG3	BY					
B3V3PDET		0.339	0.049	6.962	0.168	0.276
B3V3PINT		0.858	0.109	7.833	0.424	0.550
B3V3PNEG		1.000	0.000	0.000	0.495	0.810

BAG3 ON					
BAG1	0.264	0.038	6.887	0.388	0.388
BAG3 ON					
B1P_CONF	0.057	0.032	1.766	0.116	0.063
B3P_CONF ON					
BAG1	0.011	0.020	0.559	0.008	0.015
B3P_CONF ON					
B1P_CONF	0.269	0.029	9.342	0.269	0.273
BAG1 WITH					
B1P_CONF	0.013	0.011	1.121	0.018	0.033
BAG3 WITH					
B3P_CONF	-0.005	0.008	-0.623	-0.010	-0.019
B1V3PDET WITH					
B3V3PDET	0.168	0.024	6.895	0.168	0.281
B1V3PINT WITH					
B3V3PINT	0.135	0.025	5.514	0.135	0.141
B1V3PNEG WITH					
B3V3PNEG	-0.019	0.024	-0.769	-0.019	-0.039
Means					
B1P_CONF	1.722	0.012	140.379	1.722	3.172
Intercepts					
B1V3PDET	1.618	0.022	72.094	1.618	1.640
B1V3PINT	2.488	0.028	89.714	2.488	1.994
B1V3PNEG	1.453	0.018	82.460	1.453	1.843
B3V3PDET	1.207	0.023	52.345	1.207	1.989
B3V3PINT	1.504	0.049	30.805	1.504	1.948
B3V3PNEG	1.183	0.056	21.164	1.183	1.937
B3P_CONF	1.207	0.049	24.479	1.207	2.259
Variances					
B1P_CONF	0.295	0.012	24.764	0.295	1.000
BAG1	0.528	0.079	6.664	1.000	1.000
Residual Variances					
B1V3PDET	0.913	0.052	17.486	0.913	0.938
B1V3PINT	1.169	0.063	18.681	1.169	0.751
B1V3PNEG	0.094	0.066	1.425	0.094	0.151
B3V3PDET	0.341	0.034	10.162	0.341	0.924
B3V3PINT	0.416	0.030	13.652	0.416	0.698
B3V3PNEG	0.128	0.034	3.827	0.128	0.344
B3P_CONF	0.264	0.012	21.407	0.264	0.925
BAG3	0.207	0.038	5.464	0.844	0.844

R-SQUARE

Observed

Variable	R-Square
B1V3PDET	0.062
B1V3PINT	0.249
B1V3PNEG	0.849
B3V3PDET	0.076
B3V3PINT	0.302
B3V3PNEG	0.656
B3P_CONF	0.075

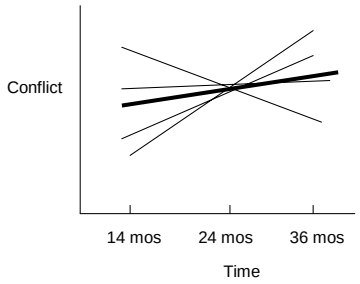
Latent

Variable	R-Square
BAG3	0.156

Latent Growth Curve Models

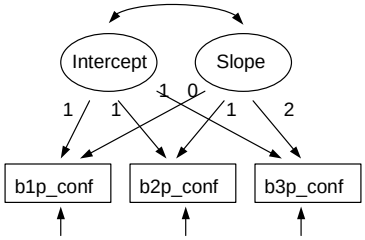
Longitudinal models that trace the growth or decline of individuals over time can also be tested with SEM. The approach is an extension of growth curve models tested using a hierarchical linear modeling (e.g., Raudenbush & Bryk, 2002), and possesses several advantages. More flexible error structures can be specified and more complex models can be tested. The model can also be extended to incorporate latent variables at each time point, leading to more accurate estimates of growth.

In general, growth curve analysis models individual trajectories of change. For linear models, each individual has a predicted intercept and slope. Researchers are not only interested in the average intercept and slope for the sample but also the extent to which intercepts and slopes vary across individuals. The following hypothetical graph of individual slopes for the change in conflict over time is illustrative.



The heavy line represents an average change over time for the sample. The thinner lines represent predicted change for 4 individuals from the data set. Notice that there is considerable variability in the level of conflict across participants at 14 months and there is considerable variability in how conflict changes over the 22 months. Growth curve models provide estimates of the variability of baseline scores and variability of change over time in addition to average baseline values and average change over time.

The figure below represents the basic specification for a latent growth curve model of conflict with three time points.



As illustrated, measurement is an indicator for both the intercept and the slope latent variable.

- Special instructions are given to obtain means for the latent variables (called “meanstructures”), which provide information about average baseline values and average slopes.

- The loadings are constrained to particular values, and the loadings shown above represent the most common loadings used (although others are possible).

- By setting the loadings for the slope factor to 0, 1, 2, the mean for the slope factor represents average linear growth.

- Because zero is chose as the first loading, the mean of the intercept factor has a special meaning—it is the average value at the first time point.

- The correlation between the intercept and slope provides information whether the initial value is associated with the rate of change (e.g., Are those with higher conflict at 14 mos more likely to decline in conflict?).

Example 12: Latent Growth Curve Model

Specifying these models in Mplus requires few additional syntax elements. The only new specification involves meanstructures. First, TYPE = MEANSTRUCTURE is needed on the analysis command (here I also use missing data estimation). Second, to refer to means or intercepts (i.e., the term used for the mean if the variable is predicted by another variable), square brackets, [ ] ,are used. In general, means for the latent variables are freely estimated (no @ sign is use) while intercepts for the indicators are set to zero.

```
TITLE:  Example 12, Latent Growth Curve Analysis;

DATA:  FILE=c:\jason\mplus\ehs\ex2.dat;
               FORMAT=FREE;

VARIABLE:  NAMES = program mrisk3 site c_maler mage race b3p35a
    b3p35b b3p35c b3p35d b3p35e b2p35a b2p35b b2p35c
   b2p35d b2p35e b1pc04a b1pc04b b1pc04c b1pc04e b1pc04f
  b1pc04g b1pc04j b1pc04k b1pc04m b1pc04n b1pc04r b1pc04t
    b1p69a b1p69b b1p69c b1p69d b1p69e b1p_cesd b1v3pdet
b1v3pint b1v3pneg b2v3pdet b2v3pint b2v3pneg b3v3pdet b3v3pint
           b3v3pneg b1p_conf b2p_conf b3p_conf;

    MISSING = program-b3p_conf(-99,-6--1);

    USEVARIABLES=b1p_conf b2p_conf b3p_conf;

ANALYSIS:
   TYPE = MEANSTRUCTURE MISSING H1; ESTIMATOR = ML; MATRIX =
COVARIANCE;

MODEL: intrcept BY b1p_conf@1 b2p_conf@1 b3p_conf@1;
       slope BY b1p_conf@0 b2p_conf@1 b3p_conf@2;
       intrcept WITH slope;
       [intrcept slope];
       [b1p_conf@0 b2p_conf@0 b3p_conf@0];

OUTPUT:  STANDARDIZED;
```

There is a shortcut specification on the model statement that produces the exact same results:

```
MODEL: intrcept slope |  b1p_conf@0 b2p_conf@1 b3p_conf@2;
```

Output for Example 12: Latent Growth Curve Model

```
SUMMARY OF ANALYSIS

Number of groups                                                 1
Number of observations                                        2457

Number of dependent variables                                    3
Number of independent variables                                  0
Number of continuous latent variables                            2

Observed dependent variables

  Continuous
   B1P_CONF    B2P_CONF    B3P_CONF

Continuous latent variables
   INTRCEPT    SLOPE

Estimator                                                       ML
Information matrix                                        OBSERVED
Maximum number of iterations                                 1000
Convergence criterion                                   0.500D-04
Maximum number of steepest descent iterations                  20
Maximum number of iterations for H1                          2000
Convergence criterion for H1                            0.100D-03

Input data file(s)
  c:\jason\mplus\ehs\ex2.dat

Input data format  FREE


SUMMARY OF DATA

     Number of patterns           7

COVARIANCE COVERAGE OF DATA

Minimum covariance coverage value   0.100


     PROPORTION OF DATA PRESENT


           Covariance Coverage
              B1P_CONF      B2P_CONF      B3P_CONF
              ________      ________      ________
 B1P_CONF       0.790
 B2P_CONF       0.611         0.755
 B3P_CONF       0.584         0.595         0.744


THE MODEL ESTIMATION TERMINATED NORMALLY


TESTS OF MODEL FIT

Chi-Square Test of Model Fit

          Value                              0.704
          Degrees of Freedom                     1
          P-Value                           0.4015

Chi-Square Test of Model Fit for the Baseline Model

          Value                            505.591
          Degrees of Freedom                     3
          P-Value                           0.0000
```

```
CFI/TLI
          CFI                                1.000
          TLI                                1.002

Loglikelihood
          H0 Value                       -4255.880
          H1 Value                       -4255.528

Information Criteria

          Number of Free Parameters              8
          Akaike (AIC)                    8527.761
          Bayesian (BIC)                  8574.214
          Sample-Size Adjusted BIC        8548.796
            (n* = (n + 2) / 24)

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                           0.000
          90 Percent C.I.                    0.000  0.050
          Probability RMSEA <= .05           0.950

SRMR (Standardized Root Mean Square Residual)

          Value                              0.005


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

 INTRCEPT BY
    B1P_CONF           1.000    0.000      0.000    0.357    0.658
    B2P_CONF           1.000    0.000      0.000    0.357    0.658
    B3P_CONF           1.000    0.000      0.000    0.357    0.667

 SLOPE    BY
    B1P_CONF           0.000    0.000      0.000    0.000    0.000
    B2P_CONF           1.000    0.000      0.000    0.184    0.340
    B3P_CONF           2.000    0.000      0.000    0.368    0.689

 INTRCEPT WITH
    SLOPE             -0.025    0.009     -2.866   -0.374   -0.374

 Means
    INTRCEPT           1.726    0.012    149.716    4.837    4.837
    SLOPE             -0.026    0.008     -3.430   -0.143   -0.143

 Intercepts
    B1P_CONF           0.000    0.000      0.000    0.000    0.000
    B2P_CONF           0.000    0.000      0.000    0.000    0.000
    B3P_CONF           0.000    0.000      0.000    0.000    0.000

 Variances
    INTRCEPT           0.127    0.014      8.847    1.000    1.000
    SLOPE              0.034    0.007      4.582    1.000    1.000

 Residual Variances
    B1P_CONF           0.167    0.015     11.226    0.167    0.567
    B2P_CONF           0.182    0.008     23.012    0.182    0.619
    B3P_CONF           0.121    0.014      8.500    0.121    0.424


R-SQUARE

    Observed
    Variable  R-Square

    B1P_CONF     0.433
    B2P_CONF     0.381
    B3P_CONF     0.576
```

Other Latent Growth Curve Analyses

There are many other growth curve applications that I will not have time to cover in detail. However, below I list just a few ideas that you should be able to do, given the other modeling knowledge you have gained here.

- Intercepts and slopes as outcomes. Any number of predictors and covariates can be used to explain variation in intercepts and slopes. For example, perhaps the mother's age is a significant predictor of initial values of conflict or change in conflict over time.

- Intercepts and slopes as predictors. An advantage of latent growth curve models over HLM (multilevel regression) is that intercepts and slope variables can be used as predictor variables. Perhaps increases in conflict lead to educational problems for the children later.

- Multi-group analyses. Growth curves can be compared across groups if there is a naturally categorical grouping variable (e.g., race). Perhaps there are differnces among White, Black, and Hispanic families in terms of the initial values or growth in conflict over time.

- Time-varying covariates. Predictors of the outcome at each measurement point can also be incorporated, and this may provide interesting theoretical findings or more accurate estimates of growth over time. For example, mother's depression level at each time point might be used as a predictor of conflict at each time point, thus providing estimates of conflict initial values and growth that have been adjusted for depression level.

- The new PLOT feature on the OUTPUT command can be used to view individual growth curves (See Chapter 17, pp.464-472, of the Mplus 3 User's Guide)

Technical Note #4: Some Recommended Readings on Longitudinal Analysis

General

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis for field settings.* Boston: Houghton Mifflin.

Dwyer, J.H. (1983). *Statistical models for the social and behavioral sciences*. New York: Oxford University Press.

Chapter 1. Kenny, D.A. (1979). *Correlation and causation*. New York: Wiley.

Menard, S. (2002). *Longitudinal research* (2nd Edition). Thousand, Oaks: Sage (QASS #76).

Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research," plus supplemental questions. In J. M. Gottman, (Ed.), *The analysis of change* (pp. 3-66) Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Taris, T.W. (2000). *A primer in longitudinal data analysis*. London: Sage.

Regression

Chapter 15: Longitudinal regression models. Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd Edition). Wahwah, NJ: Erlbaum.

Campbell, D.T., & Kenny, D.A. (1999). *A primer on regression artifacts.* New York: The Guilford Press.

Cross-lagged Panel Models

Finkel, S.E. (1995). *Causal analysis with panel data*. Thousand Oaks, CA: Sage. (QASS #105).

Gollob, H.F., & Reichardt, C.S. (1991). *Interpreting and estimating effects assuming time lags really matter.* In L. Collins & J. Horn (Eds.), Best methods for the analysis of change (pp. 243-259).

Kessler, R.C., Greenberg, D.F. (1981). *Linear panel analysis: Models quantitative change.* New York: Academic Press.

Latent Growth Curve Analysis

Duncan, T.E., Duncan, S.C., Stycker, L.A., Fuzhong, L., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Erlbaum.

Singer, J.D., & Willett, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

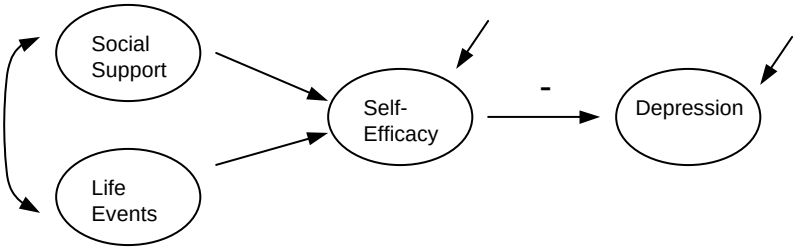
Willet, J.B., & Sayer, A.G. (1994). Using covariance structure analysis to detect correlates and predictors of change. *Psychological Bulletin, 116*, 363-381.

Other Topics in SEM

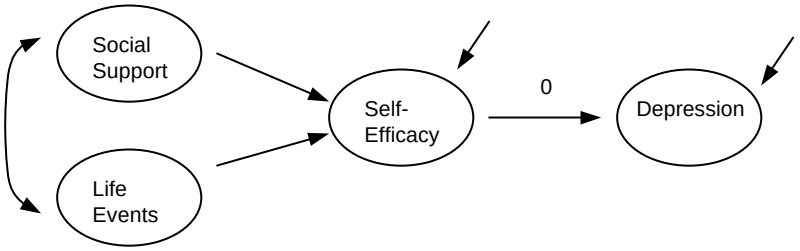
Multigroup Analysis

There are two approaches to testing moderator (i.e., interaction) hypotheses in SEM. The most common approach is to compare models in two groups. This can be done in what is called a *multigroup analysis* or sometimes called a *stacked model*. It is possible to test a model in two or more groups simultaneously and make statistical comparisons across groups.

Females



Males



- One or more paths are constrained to be equal across groups, and the fit of the constrained model is compared to the fit of the unconstrained model.
- The entire model can also be compared.
- Measurement models can be tested for *invariance* across groups. This can be complicated and tricky business (e.g., Cheung & Rensvold, 1999; Millsap, 1995), but it is important to establish that measurement characteristics are the same across groups before making inferences about predictive paths.
- Different models with the same variables can be tested in the two groups.
- Multigroup analysis is not available for mixture models (although see the KNOWNCLASS option), EFA, or logistic regression.
- Chapter 13, pp. 296-307 of the Mplus 3 User's guide provides specification details for multigroup models.

Other advanced capabilities in Mplus

Latent variable interactions. Mplus 3 includes a special preprogrammed approach to latent variable interaction using the Klein & Moosbruger (2000) full-information maximum likelihood approach. This method seems to work well for approximate N > 300. See p. 61-62 of the Mplus 3 User's guide for an example.

Latent class analysis. Mplus has special features for confirmatory factor analysis with categorical latent variables, known as "latent class analysis". Continuous, binary, or ordered categorical indicators can be used to define the latent class variable. Akin to cluster analysis, the main purpose is to identify subgroups of individuals defined by the set of indicator variables. See Chapter 7 of the Mplus 3 User's guide for more details.

Mixture modeling. The term "mixture modeling" refers to structural equation models that use latent class variables. Mplus allows integration of latent class variables in virtually any type of model (e.g., multi-group models, growth models). See Chapter 7 of the Mplus 3 User's guide for more details.

Poisson and zero-inflated Poisson variables. Count variables that are highly skewed require special estimators (using the Poisson distribution). can be analyzed by –using count variables with many zero frequencies (e.g. drug use). The COUNT statement is used under the VARIABLE command to designate such variables. These variables can be incorporated into most other types of Mplus models. See pp. 25-26, 336, 340 of the Mplus 3 User's guide.

Discrete time survival analysis. Muthen and Masyn (in press) have shown how to use latent class variables to test discrete-time survival models—a longitudinal analysis that has not been available in a structural equation framework before. See pp. 179-181 of the Mplus 3 User's guide.

Multilevel regression and multilevel structural models. Mplus is also capable of testing mulitilevel regression (HLM) models designed for hierarchically structured data and growth models that are usually analyzed with packages such as HLM, MLWIN, and SAS Proc Mixed. Mplus has special features for extending hierarchically structured models to multilevel confirmatory factor models, path models, and structural models. See Chapter 9 of the Mplus 3 User's guide.

Complex sampling. Mplus is the only SEM package that has incorporated methodology for adjusting parameters and standard errors for cluster or stratified survey sampling designs. Weight variables with or without design specifications can be used. See Chapter 9 of the Mplus 3 User's guide.

Web Resources

MPlus

Free lecture movies at UCLA statistics site:

More Mplus oriented:

http://www.ats.ucla.edu/stat/seminars/

More statistical in nature:

http://www.ats.ucla.edu/stat/seminars/ed231e/

Technical appendices for Mplus user's guide provides details on estimators, missing data, complex sampling designs, and other topics.

http://www.statmodel.com/mplus/techappen.pdf

Many examples of simple and advanced analyses with Mplus:

http://www.statmodel.com/mplus/examples/

Tutorial from the University of Texas research consulting site:

http://www.utexas.edu/its/rc/tutorials/stat/mplus/

SEM in General

SEMNET Discussion List and Archive
http://www.gsu.edu/~mkteer/semnet.html

Dave Kenny's site (great didactic information on SEM and other statistical topics)
http://users.rcn.com/dakenny/causalm.htm

Ed Rigdon's site (many SEM links)
http://www.gsu.edu/~mkteer/index.html

Patrick Curran's website (growth curve models):
http://www.unc.edu/~curran/

Jason Newsom's SEM references page:
http://www.ioa.pdx.edu/newsom/semrefs.htm

EHS Example Data Set 1

EHSID Identification number
PROGRAM Program Group? (0=Comparison, 1=Program)
MRISK3 Risk group: 1=0-2, 2=3, 3=4-5 risks
SITE Site code
C_MALER 1= Focus Child is Male: For Regr
MAGE Age of Mother at Rand Asn (years) (trk)
RACE White, Black, Hisp, Other
B3P35A conf1-fights
B3P35B conf2-lose tempers
B3P35C conf3-get angry
B3P35D conf4-criticize
B3P35E conf5-hit
B2P35A conf1-fights
B2P35B conf2-lose tempers
B2P35C conf3-get angry
B2P35D conf4-criticize
B2P35E conf5-hit
B1PC04A cesd1-bothered
B1PC04B cesd2- eating
B1PC04C cesd3- blues
B1PC04E cesd4- mind on things
B1PC04F cesd5- depressed
B1PC04G cesd6- effort
B1PC04J cesd7- fearful
B1PC04K cesd8- restless
B1PC04M cesd9- talked less
B1PC04N cesd10- felt lonely
B1PC04R cesd11- felt sad
B1PC04T cesd12-not get going
B1P69A conf1-fights
B1P69B conf2-lose tempers (R)
B1P69C conf3-get angry
B1P69D conf4-criticize
B1P69E conf5-hit
B1P_CESD 14m CES-Depression total scale
B1V3PDET 14m Parent Detachment 3-bag
B1V3PINT 14m Parent Intrusiveness3-bag
B1V3PNEG 14m Parent Negative Regard 3-bag
B2V3PDET 24m Parent Detachment 3-bag
B2V3PINT 24m Parent Intrusiveness 3-bag
B2V3PNEG 24m Parent Negative Regard 3-bag
B3V3PDET 36m Parent Detachment 3-bag
B3V3PINT 36m Parent Intrusiveness3-bag
B3V3PNEG 36m Parent Negative Regard 3-bag
B1P_CONF 14m FES CONFLICT
B2P_CONF 24m FES CONFLICT
B3P_CONF 36m FES CONFLICT

References

Arbuckle, J.L. (1996) Full information estimation in the presence of incomplete data. In G.A. Marcoulides and R.E. Schumacker [Eds.] Advanced structural equation modeling: Issues and Techniques. Mahwah, NJ: Lawrence Erlbaum Associates.

Bentler (1990), Comparative Fit Indices in Structural Models, Psychological Bulletin, 107, 238-246.

Bentler, P. M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software.

Bentler, P.M., & Chou, C.-P. (1988). Practical issues in structural modeling. In J.S. Long (Ed.), Common problems/proper solutions (pp. 161-192). Beverly Hills, CA: Sage.
Bollen, 1990, Overall fit in covariance structure models: Two types of sample size effects. Psychological Bulletin, 107, 256-259.

Bollen, K.A. (1989). Structural equations with latent variables. New York: Wiley.

Browne, M.W. (1984). Asymptotic distribution free methods in analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 37, 62-83.

Cheung, G.W. & Rensvold, R.B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. Journal of Management, 25, 1-27.

Curran, P.J., Harford, T., & Muthen, B.O. (1996). The relation between heavy alcohol use and bar patronage: A latent growth model. Journal of Studies on Alcohol, 57, 410-418.

DeCarlo, L. T. (1997), On the meaning and use of kurtosis. Psychological Methods, 2, 292-307.

Gerbing, D.W., & Anderson, J.C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K.A. Bollen, & J.S. Long (eds.), Testing structural equation models. Newbury Park, CA: Sage.

Hancock, G. R. & Nevitt, J. (1999). Bootstrapping and identification of exogenous latent variables within structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(4), 394-399.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1-55.
Hu, L., Bentler, P.M., & Kano, Y. (1988). Can test statistics in covariance structure analysis be trusted? Psychological Bulletin, 112, 351-362.

Hu, L.-T., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), Structural Equation Modeling. Concepts, Issues, and Applications (pp. 76-99). London: Sage.

Hu, L.-T., & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), Structural Equation Modeling. Concepts, Issues, and Applications (pp. 76-99). London: Sage.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65, 457-474.*

MacCallum, R.X., Zhang, S., Preacher, K.J., & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19-40.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness of fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Maruyama (1998).Basics of Structural Equation Modeling. Thousand Oaks: Sage.

Millsap, R.E. (1995). Measurement invariance, predictive invariance, and the duality paradox. Multivariate Behavioral Research, 30, 577-605

Mulaik, S.A., James, L.R., Van Alstine, J., Bennett, N., Lind, S., & Stillwell, C.D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105, 430-445.

Muthén, B. & Masyn, K. (in press). Discrete-time survival mixture analysis. Journal of Educational and Behavioral Statistics.

Muthen, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K.A. Bollen & J.S. Long (Eds.), Testing structural equation models (pp. 205-234). Newbury Park, CA: Sage.

Muthen, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. Psychometrika, 51,431-462.

Olsson, U.H., Foss, T., Troye, S. V., & Roy D. Howell (2000). The Performance of ML, GLS and WLS Estimation in Structural Equation Modeling Under Conditions of Misspecification and Nonnormality. Structural Equation Modeling, 7 (4), 557-595.

Raudenbush, S.W., & Bryk, A.S., (2002) *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage

Raykov, T. (1997). Growth curve analysis of ability means and variances in measures of fluid intelligence of older adults. Structural Equation Modeling,4(4), 283-319.

Raykov, T. (2000). On the large-sample bias, variance, and mean squared error of the conventional noncentrality parameter estimator of covariance structure models. Structural Equation Modeling, 7, 431-441.

Rigdon, E. E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. Multivariate Behavioral Research, 30, 359-383.

Rubin, D. (1987). Multiple imputation for nonresponse in surveys. Wiley.

Satorra, A., & Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. 1988 Proceedings of the Business and Economic Statistics Section of the American Statistical Association, 308-313.

Satorra, A., & Bentler, P.M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye and C.C. Clogg (eds.), Latent Variable Analysis: Applications to Developmental Research (pp. 399-419). Newbury Park: Sage.
Steiger, J.H. (1989). EZPATH: A supplementary module for SYSTATand SYGRAPH. Evanston, IL: SYSTAT.

Steiger, J.H., & Lind, J.C. (1980). Statistically-based tests for the number of factors. Paper presented at the Annual Spring Meeting of the Psychometric Society. Iowa City, Iowa.

Tanaka, J.S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. Child Development, 58, 134-146.

Tanaka, J.S. (1993). Multifaceted conceptions of fit in structural equation models. In K.A. Bollen, & J.S. Long (eds.), Testing structural equation models. Newbury Park, CA: Sage.

West, S. G., Finch, J.F, & Curran, P.J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R.H. Hoyle (Ed), *Structural equation modeling: Concepts, issues, and applications.* (pp. 56-75). Thousand Oaks, CA: Sage Publications.).

Yuan, K.H. & Bentler, P.M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In Sobel, M.E. & Becker, M.P. (eds.), Sociological Methodology 2000 (pp. 165-200). Washington, D.C.: ASA.