An Introduction to SEM in Mplus

Ben Kite

Saturday Seminar Series
Quantitative Training Program
Center for Research Methods and Data Analysis
Goals

• Provide an introduction to *Mplus*
• Demonstrate *Mplus* with a variety of examples
• Provide *Mplus* tips and suggestions
• Answer questions regarding statistical modeling with *Mplus*
• Provide additional sources for information
What is **Mplus**?

- Created by Linda K. Muthén and Bengt O. Muthén
- Now in 7\textsuperscript{th} version!
  - Version 7 September 2012
  - Version 6 April 2010
  - Version 5 November 2007
  - Version 4 February 2006
  - Version 3 March 2004
  - Version 2 February 2001
  - Version 1 November 1998
What is Mplus?

• Statistical modeling program
  – Wide choice of models, estimators, algorithms
  – Cross-sectional & longitudinal
  – Single & multi-level models
  – Observed & latent variables
  – Continuous, categorical, binary, censored, count, or combinations of these
  – Missing data
  – Monte Carlo simulation
  – Latent Classes
Some Advantages

• Compact syntax (few lines of code)
• Very flexible & powerful for complex modeling
• Good user’s guide with lots of examples
• Quick & extensive customer support for registered users
• Website has a discussion board open to the public http://www.statmodel.com/
Some Disadvantages

• Expensive $$$
  – Student Pricing $195-350
  – University Pricing $595-895

• Syntax based
  – No pull-down menus

• Has many defaults that can be tricky for new users/modelers to learn
Basic *Mplus*

- Open the program and use its syntax editor
- Write your syntax, save it
- Click  
- *Mplus* will open your output file (results) for you!
Importing Data Files

- Must be a text file
  - We suggest using a .csv or .dat (tab delimited) file
- No importer for Excel, Stata, SPSS, etc
- Can use
  - Raw data
  - Correlation or covariance matrix
  - A few other types (incl. means, imputation)
Data Files

• Numeric only (No variable names)
• Maximum of 500 variables
• MUST have a code for missing values.
  – E.g., -999
Main Commands

- TITLE
- DATA (required)
- VARIABLE(required)
- ANALYSIS
- MODEL
- OUTPUT

They are often written in all uppercase to make it easy to read, but it is not required.
Other Commands

- DEFINE
- SAVEDATA
- PLOT

They are often written in all uppercase to make it easy to read, but it is not required
Basic Syntax Conventions

• ! Can be used to make comments
• End statements with a semicolon;
• * is used to free a parameter estimate
• *.5 indicates .5 is your start value
• @ fixes parameter to a specified value
• (1) or (a) means equate all parameters that have the same number/letter in parentheses
• Dash can be used, as in var1-var5
  – Same as saying var1 var2 var3 var4 var5
More Syntax Conventions

- Not case sensitive
  - i.e. not picky about upper/lower case letters
- Commands can be in any order
- Commands start on a new line and are followed by a colon (e.g. DATA: )
- Only Data & Variable commands are required for all analyses
- No more than one option per line
- No more than 90 characters wide
TITLE:

• Not required
• Provide a title for the analysis (for your own use, not part of the program)

e.g.
TITLE: My Analysis of Happiness;
DATA:

• Information about the data set & where it is found
  FILE is “C:/mydocuments/happiness.dat”;
  or
  FILE is “data/happiness.dat”;
  or
  FILE is happiness.dat;

• For Summary Statistics
  – TYPE IS
VARIABLE:

- Information about the variables
  - Order they appear in set
  - No more than 8 characters in variable names

VARIABLE:

NAMES ARE id age sex marital happines;
USEVARIABLES ARE age sex marital happines;
CATEGORICAL ARE sex marital; !(not always needed)
MISSING ARE (-999);
ANALYSIS:

• Technical details of analysis (not always needed)

e.g.
TYPE = MIXTURE;
or
TYPE IS TWOLEVEL RANDOM;
BOOTSTRAP = 500;
MODEL:

- Describes the model to be estimated

  e.g. in a factor analysis
  f1 BY var1 var2 var3;
  f2 BY var4-var6;
  f1 WITH f2;
  f2 ON X;
MODEL:

• Describes the model to be estimated

  e.g. in a factor analysis
  f1 BY var1 var2 var3;
  f2 BY var4-var6;
  f1 WITH f2;
  f2 ON X;
MODEL: BY, WITH, & ON

• BY – indicator (e.g. factor 1 is indicated BY var1-var3)
• WITH - correlate
• ON - regress DV on IV (order matters)
  – Y on X
• To prevent estimation of a pathway
  – Var1 WITH Var2 @0;
OUTPUT:

- Request output that is not part of the default
- Common Requests
  - MODINDICES: Reports modification indices for the model.
  - Use (#) to have indices only $\geq$ # reported (e.g. modindices(10))
• Common Requests
e.g.
standardized std tech1;
!To get standardized estimates & technical output to help with trouble shooting;
Example Models in Mplus

- Exploratory Factor Analysis (EFA)
- Confirmatory Factor Analysis (CFA)
- Multiple Group CFA
- Multiple Linear Regression
- Structural Equation Modeling
- Latent Growth Curve Modeling
Data

• Items collected from job seekers using the Excelsior Spring Job Corps center

• Variables
  – id: Subject identification number
  – wjcalc: Subject's score on the WJ calculation subtest. (Numeric)
  – wjspl: Subject's score on the WJ spelling subtest. (Numeric)
  – wratcalc: Subject's score on the WRAT calculation subtest. (Numeric)
  – wratspl: Subject's score on the WRAT spelling subtest. (Numeric)
  – waiscalc: Subject's score on the WAIS arithmetic calculations subtest. (Numeric)
  – waisspl: Subject's score on the WAIS spelling subtest. (Numeric)
  – edlevel: What is the highest level of education completed by the subject? (Ordinal)
  – newschl: Did the subject ever change high schools? (Binary: 1=Yes, 0=No)
  – suspend: Has the subject ever been suspended from high school? (Binary: 1=Yes, 0=No)
  – expelled: Has the subject ever been expelled from high school? (Binary: 1=Yes, 0=No)
  – haveld: Has the subject been diagnosed with a learning disorder? (Binary: 1=Yes, 0=No)
  – female: Gender (Binary: 1=Female, 0=Male)
  – age: Age in years (Numeric)

• Missing data are coded as “99999”
Example 1 – Exploratory Factor Analysis

• The Excelsior Springs data. How many latent variables underlie subtest scores (6 items)?
• *Mplus* can quickly run many EFAs and report the results.
• Unlike SPSS/PASW it does not default to principle components analysis.
• We can run EFAs for one and two possible latent factors.
EFA – One Factor

Factor 1

\[ \lambda_{11}, \lambda_{21}, \lambda_{31}, \lambda_{41}, \lambda_{51}, \lambda_{61} \]

\[ \theta_{11}, \theta_{22}, \theta_{33}, \theta_{44}, \theta_{55}, \theta_{66} \]

- wratcalc
- wjcalc
- waiscalc
- wratspl
- wjspl
- waisspl
EFA – Two Factors

Factor 1

Factor 2

\[ \Phi_{21} \]

\[ \lambda_{11} \]
\[ \lambda_{21} \]
\[ \lambda_{31} \]
\[ \lambda_{41} \]
\[ \lambda_{51} \]
\[ \lambda_{61} \]
\[ \lambda_{12} \]
\[ \lambda_{22} \]
\[ \lambda_{32} \]
\[ \lambda_{42} \]
\[ \lambda_{52} \]
\[ \lambda_{62} \]

wratcalc
wijcalc
waiscalc
wratspl
wjspl
waisspl

\[ \theta_{11} \]
\[ \theta_{22} \]
\[ \theta_{33} \]
\[ \theta_{44} \]
\[ \theta_{55} \]
\[ \theta_{66} \]
TITLE: 
Example 1 - Exploratory Factor Analysis 
DATA: 
FILE IS "../../data/job_placement.csv"; 
VARIABLE:
NAMES ARE 
  id wjcalc wjspl wratspl wratcalc waiscalc waisspl 
edlevel newschl suspend expelled haveld female age; 
USEVARIABLES ARE 
  wjcalc wjspl wratspl wratcalc waiscalc waisspl; 
MISSING ARE all (99999); 
ANALYSIS: TYPE = EFA 1 2; 
!Default estimator is ML; 
!Default rotation is Oblique Geomin;
EFA- Beyond Defaults

ANALYSIS: TYPE = EFA 1 2;
!Default estimator is ML;
!Default rotation is Oblique Geomin;

• ESTIMATOR = ML;
  – MLM;
  – MLMV
  – MLR;
  – MLF;
  – MULML;
  – WLS;
  – WLSM;
  – WLSMV;
  – ULS;
  – ULSMV;

ROTATION = GEOMIN; GEOMIN
GEOMIN (OBLIQUE);
(ORTHOGONAL);
QUARTIMIN; OBLIQUE
CF-VARIMAX
CF-QUARTIMAX;
CF-EQUAMAX;
CF-PARSIMAX;
CF-FACPARSIM;
CRAWFER; OBLIQUE 1/p
CRAWFER (OBLIQUE or ORTHOGONAL value);
OBLIMIN; OBLIQUE 0
OBLIMIN (OBLIQUE or ORTHOGONAL value);
VARIMAX;
PROMAX;
EFA Results
Example 2 - Confirmatory Factor Analysis

• CFA where only certain parameters are estimated.

• We believe that Math is indicated by:
  • wratcalc, wjcalc, waiscalc

• We believe that Spelling is indicated by:
  • wratspl, wjspl, waisspl
CFA- Greek Notation

The CFA model asserts these $= 0$
CFA - Mplus Notation

Math @1.0;

WITH

BY

Spell @1.0;

BY

wratcalc
wjcalc
waiscalc
wratspl
wjspl
waisspl

Math

Spell

Math with Spell
CFA Mplus Syntax

TITLE:
Example 2 - Confirmatory Factor Analysis

DATA:
FILE IS "../data/job_placement.csv";

VARIABLE:
NAMES ARE
id wjcalc wjspl wratspl wratcalc waisscalc waisspl edlevel newschl suspend expelled haveld female age;
USEVARIABLES ARE
wratspl wjspl waisspl wratcalc wjcalc waiscalc;
MISSING ARE all(99999);

MODEL:
MATH BY wratcalc* wjcalc waiscalc;
SPELL BY wratspl* wjspl waisspl;
MATH@1 SPELL@1; Set scale using fixed factor ID;

OUTPUT:
TECH1;
STDYX;

NOTE: Mplus defaults to marker variable identification.
* Frees parameter
CFA Results
Example 3 - Multiple Group CFA

- Compare males and female on Math and Spelling
- Determine if the measurement parameters \((\lambda_j, \tau_j)\) are equal for both groups
- Three stages (models)
  - Configural invariance (same model)
  - Metric invariance (equal factor loadings)
  - Scalar invariance (equal factor loadings and item intercepts)
Configural Invariance Model
TITLE:
Example 3 - Multiple Group Confirmatory Factor Analysis (Configural Invariance)

DATA:
FILE IS "../../data/job_placement.csv";

VARIABLE:
NAMES ARE
id wjcalc wjspl wratspl wratcalc waiscalc waisspl
edlevel newschl suspend expelled haveld female age;
USEVARIABLES ARE
wratspl wjspl waisspl wratcalc wjcalc waiscalc;
MISSING ARE all(99999);
GROUPING IS female (0=male 1 =female);
Configural Model - Syntax

MODEL:
MATH BY wratcalc* wjcalc waiscalc;
SPELL BY wratspl* wjspl waisspl;
[wratspl wjspl waisspl wratcalc wjcalc waiscalc];
MATH@1 SPELL@1;
[MATH@0 SPELL@0];
MODEL female:
MATH BY wratcalc* wjcalc waiscalc;
SPELL BY wratspl* wjspl waisspl;
[wratspl wjspl waisspl wratcalc wjcalc waiscalc];

OUTPUT:
TECH1;
STDXY;
Configural Model Results
Metric Invariance Model

Math

Spell

\[ \Phi_{21g} \]

\[ \alpha \]

\[ \lambda_{11} \]
\[ \lambda_{21} \]
\[ \lambda_{31} \]
\[ \lambda_{42} \]
\[ \lambda_{52} \]
\[ \lambda_{62} \]

\[ \tau_{1g} \]
\[ \tau_{2g} \]
\[ \tau_{3g} \]
\[ \tau_{4g} \]
\[ \tau_{5g} \]
\[ \tau_{6g} \]
Metric Model - Syntax

**TITLE:**
Example 3 - Multiple Group Confirmatory Factor Analysis (Weak Invariance)

**DATA:**
FILE IS "../data/job_placement.csv";

**VARIABLE:**
NAMES ARE
id wjcalc wjspl wratspl wratcalc waiscalc waisspl edlevel newschl suspend expelled haveld female age;
USEVARIABLES ARE
wratspl wjspl waisspl wratcalc wjcalc waiscalc;
MISSING ARE all(99999);
GROUPING IS female (0=male 1 =female);
MODEL:
MATH BY wratcalc* wjcalc waiscalc;
SPELL BY wratspl* wjspl waisspl;
[wratspl wjspl waisspl wratcalc wjcalc waiscalc];
MATH@1 SPELL@1;
[MATH@0 SPELL@0];

MODEL female:
[wratspl wjspl waisspl wratcalc wjcalc waiscalc];
MATH SPELL;

OUTPUT:
TECH1;
STDXY;
Metric Model Results
Scalar Invariance Model

1, $\Phi_{112}$

Math

$\lambda_{11.}$

$\lambda_{21.}$

$\lambda_{31.}$

$\Phi_{21g}$

$0, \alpha_{12}$

$0, \alpha_{22}$

Spell

$\alpha$

1, $\Phi_{222}$

$\lambda_{42.}$

$\lambda_{52.}$

$\lambda_{62.}$

$\tau_1.$

$\tau_2.$

$\tau_3.$

$\tau_4.$

$\tau_5.$

$\tau_6.$

$\tau$
Scalar Model - Syntax

**TITLE:**
Example 3 - Multiple Group Confirmatory Factor Analysis (Strong Invariance)

**DATA:**
FILE IS "../../data/job_placement.csv";

**VARIABLE:**
NAMES ARE
id wjcalc wjspl wratspl wratcalc waiscalc waisspl
edlevel newschl suspend expelled haveld female age;
USEVARIABLES ARE
wratspl wjspl waisspl wratcalc wjcalc waiscalc;
MISSING ARE all(99999);
GROUPING IS female (0=male 1 =female);
Scalar Model - Syntax

**MODEL:**
MATH BY wratcalc* wjcalc waiscalc;
SPELL BY wratspl* wjspl waisspl;
[wratspl wjspl waisspl wratcalc wjcalc waiscalc];
MATH@1 SPELL@1;
[MATH@0 SPELL@0];

MODEL female:
MATH SPELL;
[MATH SPELL];

**OUTPUT:**
TECH1;
STDXY;
Scalar Model Results
Example 3 – Multiple Group CFA

- Model comparisons can be conducted to test for metric invariance
  - Fit for configural model vs. metric model
- Or scalar model
  - Fit for metric model vs. scalar model
- This can be done with Chi-square difference testing
  - $\chi^2_{\text{metric}} - \chi^2_{\text{config}} = \chi^2_{\text{diff}}$
  - $df_{\text{metric}} - df_{\text{config}} = df_{\text{diff}}$
- *Mplus* 7.1 added a new command that conducts model comparisons for you
MULTIPLE MODELS - SYNTAX

TITLE:
Example 3 - Multiple Group CFA Multiple Models

DATA:
(same as before)

VARIABLE:
(same as before)

ANALYSIS:
Model = configural metric scalar;

MODEL:
MATH BY wratcalc wjcalc waiscalc;
SPELL BY wratspl wjspl waisspl;

OUTPUT:
(same as before)
Multiple Models Result
Example 4 - Multiple Linear Regression

- Regress WJ spelling scores (wjspl) on predictor variables
- Same as ordinary regression, but it is easily done in Mplus
Multiple Linear Regression Model

- edlevel: $\beta_1$
- newschl: $\beta_2$
- suspend: $\beta_3$
- expelled: $\beta_4$
- haveld: $\beta_5$
- female: $\beta_6$
- age: $\beta_7$

$\theta_{11}$
$\tau_1$
1

KU THE UNIVERSITY OF KANSAS
KU COLLEGE OF LIBERAL ARTS & SCIENCES
Center for Research Methods & Data Analysis
TITLE:
Example 4 - Multiple Linear Regression

DATA:
FILE IS "../../data/job_placement.csv";

VARIABLE:
NAMES ARE
id wjcalc wjspl wratspl wratcalc waisscalc waisspl
edlevel newschl suspend expelled haveld female age;

USEVARIABLES ARE
wjspl edlevel newschl suspend expelled haveld female age;

MISSING ARE all(99999);

MODEL:
wjspl ON edlevel newschl suspend expelled haveld female age;
MLR Results
Example 5 - Structural Equation Modeling

- Estimate coefficients for predictors of the Math and Spelling latent variables
- This can be done by combining CFA and MLR
- This is an example of a LISREL model (measurement and structural component)
Structural Equation Modeling

edlevel
newschl
suspend
expelled
haveld
female
age

Math

Spell

\( \Psi_{11} \)
\( \Psi_{21} \)
\( \Psi_{22} \)

\( \lambda_{21} \)
\( \lambda_{31} \)
\( \lambda_{52} \)
\( \lambda_{62} \)

1

1

wratcalc
wjcaltc
waisscalc
wratspl
wjspl
waisspl

edlevel
newschl
suspend
expelled
haveld
female
age
SEM - Syntax

TITLE:
Example 5 - Structural Equation Model

DATA:
FILE IS "../..//data/job_placement.csv";

VARIABLE:
NAMES ARE
id wjcalc wjspl wratspl wratcalc waiscalc waisspl
edlevel newschl suspend expelled haveld female age;

USEVARIABLES ARE
wratspl wjspl waisspl wratcalc wjcalc waiscalc
edlevel newschl suspend expelled haveld female age;
MISSING ARE all(99999);
SEM - Syntax

MODEL:
MATH BY wratcalc wjcalc waiscalc;
SPELL BY wratspl wjspl waisspl;
MATH ON edlevel newschl suspend expelled haveld female age;
SPELL ON edlevel newschl suspend expelled haveld female age;
MATH WITH SPELL;

OUTPUT:
SAMPSTAT;
TECH1;
STDYX;
SEM Results
Example 5b - Structural Equation Modeling

- Seven items that are indicators of three latent variables about insomnia
  - Impact of insomnia
  - Severity of insomnia
  - Satisfaction with sleep
- Four items that are indicators of Quality of Life
- Estimate regression coefficients for insomnia variables predicting Quality of life
TITLE: Example 5B
DATA:
FILE IS "../data/insomnia.dat";
VARIABLE:
names are insom1-insom7 cesd1-cesd20 phy psy soc env;
usevariables are insom1-insom7 phy psy soc env;
missing all (999);
MODEL:
  Impact by insom5 insom6 insom7;
  Severity by insom1 insom2 insom3;
  Satisf by insom4 insom1 insom7;
  Qol by phy psy soc env;
  Qol on Impact Severity Satisf;
SEM 2 Results
Example 6 - Latent Growth Curve Modeling

• We recorded participants anxiety level at four different time points.
• Do anxiety levels change over time?
• Initial anxiety (intercept) and change in anxiety (slope) can be modeled as latent variables.
Latent Growth Curve Modeling

\[ \Phi_{11} \]

\[ \Phi_{21} \]

\[ \Phi_{22} \]

\[ \alpha_1 \]

\[ \alpha_2 \]

\[ \alpha \]

\[ \tau \]

Intercept \hspace{2cm} Slope

anxiety1 \hspace{1cm} anxiety2 \hspace{1cm} anxiety3 \hspace{1cm} anxiety4
TITLE:
Example 6 - Linear Latent Growth Curve Model with 4 Time Points
DATA:
FILE IS "../data/anxiety.dat";
VARIABLE:
NAMES ARE anxiety1 anxiety2 anxiety3 anxiety4;
MODEL:
INTCEPT BY anxiety1@1 anxiety2@1 anxiety3@1 anxiety4@1;
LINEAR BY anxiety1@0 anxiety2@1 anxiety3@2 anxiety4@3;
anxiety1-anxiety4;
[anxiety1-anxiety4@0];
INTCEPT LINEAR;
[INTCEPT LINEAR];
INTCEPT WITH LINEAR;
LGCM Results
LGCM Results

Intercept

Slope

\(\alpha\)

\(\tau\)

\begin{align*}
\text{anxiety1} & \quad 0.067 \\
\text{anxiety2} & \quad 0.048 \\
\text{anxiety3} & \quad 0.048 \\
\text{anxiety4} & \quad 0.040 \\
\end{align*}

\begin{align*}
\text{Intercept} & \quad 0.151 \\
\text{Slope} & \quad 0.007 \\
\end{align*}

\begin{align*}
\text{Intercept} & \quad -0.011 \\
\text{Slope} & \quad -0.062 \\
\end{align*}

\begin{align*}
\text{anxiety1} & \quad 1 \\
\text{anxiety2} & \quad 1 \\
\text{anxiety3} & \quad 1 \\
\text{anxiety4} & \quad 1 \\
\end{align*}
Troubleshooting (1)

- Check if model is specified correctly
- Draw a diagram of your commands
- Read your summary of analysis (output)
  - Number of groups
  - Number of observations
  - Number of DVs, IVs, continuous latent vars
  - Names of DVs, latent vars
Troubleshooting (2)

• Check data file is reading in correctly
  – Eyeball your data file for funny patterns
  – Check your missing codes
  – Warning Message
e.g. 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: 34

Make sure this number is correct based on your data file
Troubleshooting (3)

COMMON Warning Message:
The residual covariance matrix (Theta) is not positive definite. This could indicate a negative variance/residual variance for an observed variable, a correlation greater or equal to 1 between two observed variables or a linear dependency among more than 2 observed variables.

PROBLEM INVOLVING VARIABLE VAR8
• Check the technical output (e.g. tech1)
Troubleshooting (4)

• Exceeds iterations & did not converge
  – Try start values for estimated parameters (can start with what the output gives you)
    • e.g., y ON x *.1;

• Increase the number of iterations
  – ANALYSIS: ITERATIONS = 10,000;
    (default is 1,000)

• Try a smaller piece of the model and work your way up to more complex
Troubleshooting (5)

• Model is unidentified/under identified
  – Estimate fewer parameters
  – Try fixing one or more parameters to 0 or 1 or to be equal to something else in the model
    e.g. in a longitudinal model, fix residuals to be equal across all time points. Saves you some parameter estimates
Additional Help/Resources

- Mplus Website
  - [http://statmodel.com/](http://statmodel.com/)
  - Contains downloadable PDF of User’s guide and example code with data.
  - Has support forum for users