Using Mplus To Do Multistep Mixture Modeling: Latent Class Analysis

Bengt Muthén
Professor Emeritus, UCLA
Mplus: https://www.statmodel.com
bmuthen@statmodel.com

Tihomir Asparouhov
Linda Muthén
Mplus

Mplus Web Talks: No. 8
July 2025

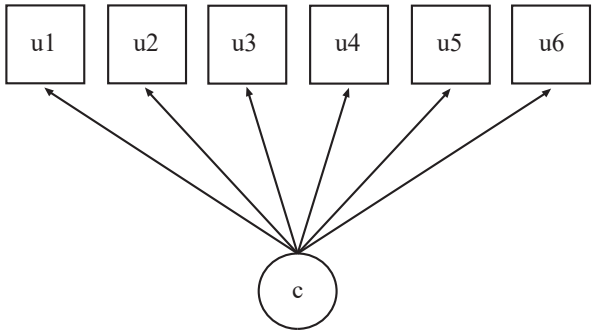
We thank Thuy Nguyen and Noah Hastings for expert assistance.

Outline

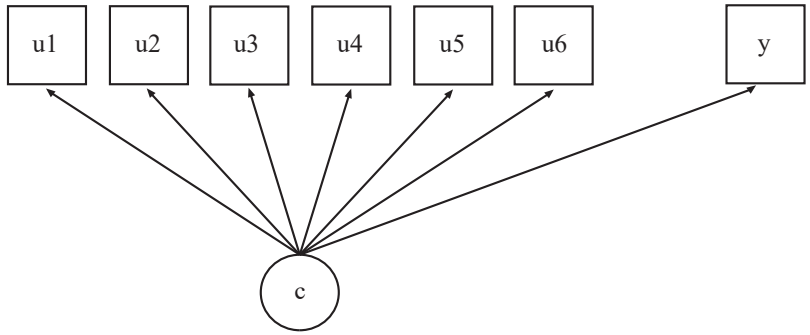
Outline, Continued

- Real-data example using antisocial behavior measures: 93-101
- Further topics: 103-115
  - Class-varying regression: 103-106
  - Last step considerations: 108-109
  - CPROBABILITIES with multiple imputation: 111-115
- Statistical methodology, appendices, references: 117-139
  - Posterior probabilities and MLC for 3-step and BCH: 117-122
  - Missing data on a binary covariate: 124-125
  - Why class-varying regression is an interaction: 127-128
  - Overview of Monte Carlo studies for 3-step, BCH, and 2-step: 130-132
  - Appendix: Generating missing data on covariates: 134-137

Background: Measurement Model and Full Model



(a) LCA measurement model



(b) Full model with distal outcome

- Measurement model
  - Determines the latent classes
  - Example: LCA
- Full model
  - Adds other variables to the measurement model
  - Example: Distal outcome predicted by the latent class variable

Multistep Rationale

- There is a need to separate the estimation of the measurement model from the full model so the latent classes don't change
  - Distal outcomes in the full model may be measured later in time than the variables of the measurement model
  - Covariates in the full model may affect the latent class enumeration
- The separation of the estimation is done by multistep approaches
  - First step: The latent classes are determined only by the measurement model
  - Last step: Distal outcomes and/or covariates of the full model are added
  - Theory is shown in the Statistical Methodology section

Multistep Approaches: 3-Step, BCH, 2-Step

- The 3-step and BCH approaches use the posterior probabilities from the estimated measurement model to determine each individual's most likely class membership and take its measurement error into account
  - 3-step: Last step uses fixed logits from measurement model to take measurement error into account in the Most Likely Class (MLC) classification
  - BCH: Last step uses weights from measurement model to take measurement error into account in the MLC classification
- 2-step: Last step uses fixed measurement model parameters to avoid measurement error

Posterior Probabilities and Most Likely Class (MLC)

- A posterior probability is the probability of an individual being in each class based on the estimated model and the individual's data
- Posterior probabilities for each individual and class are saved along with MLC using the CPROBABILITIES option of the SAVEDATA command
- Example: 6 individuals, 3 latent classes:

	CPROB1	CPROB2	CPROB3	MLC
i = 1	**0.5**	0.3	0.2	1
i = 2	0.1	0.1	**0.8**	3
i = 3	0.1	**0.6**	0.3	2
i = 4	0.3	**0.4**	0.3	2
i = 5	0.3	**0.7**	0.0	2
i = 6	**0.5**	0.2	0.3	1
P(C):	0.30	0.38	0.32	

- For each individual (row), CPROB1 + CPROB2 + CPROB3 = 1
- For each column, the average CPROB gives the class probability

Measurement Error in MLC

- From the table of posterior probabilities for all individuals it is possible to compute a summary table of classification quality:

Classification Probabilities for the Most Likely Latent Class Membership (Column) by Latent Class (Row)

	MLC=1	MLC=2	MLC=3
C=1:	**0.56**	0.38	0.06
C=2:	0.22	**0.75**	0.04
C=3:	0.26	0.31	**0.42**

- MLC is a perfect measure of C if all diagonal elements are 1 and all off-diagonal elements 0
- Non-zero off-diagonal elements correspond to MLC measurement error
- The 3-step and BCH multistep approaches take this measurement error into account in their last step analysis
- A full description is given in the Statistical Methodology section

Automatic and Manual Multistep Approaches

- Mplus offers automatic multistep approaches in certain setttings which simplify the analysis, as well as manual approaches which are more generally applicable

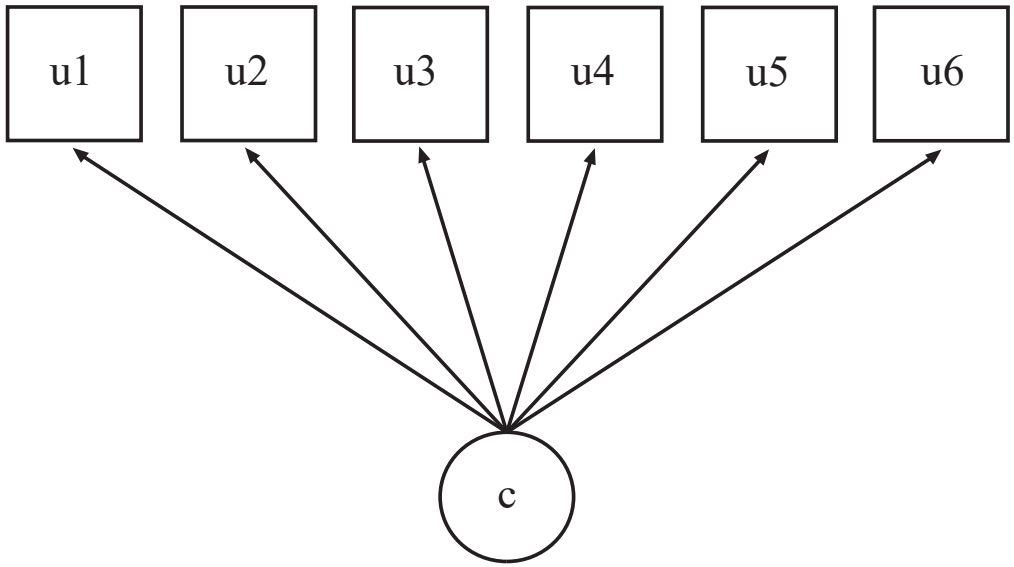
	Automatic	Manual
Distals	3-step BCH	3-step BCH 2-step
Covariates	3-step (R3STEP)	3-step BCH 2-step

- Following is a discussion of these two possibilities, beginning with their common starting point of the measurement model

Outline

- Background
- **Measurement model analysis**
- Automatic multistep approaches
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

LCA Measurement Model Analysis



- Latent class indicators: u1-u6
- Latent class variable: c
- Class enumeration determined in the analysis of the measurement model
- For a discussion of LCA, see our Short Course Videos, Topic 5

Input for Measurement Model

TITLE:	Measurement model, entropy = 0.681
DATA:	FILE = distal.dat;
VARIABLE:	NAMES = u1-u6 x1 y2 y3 x2 y1 x3 class; **USEVARIABLES = u1-u6;** MISSING = ALL(999); CATEGORICAL = u1-u6; CLASSES = c(3);
ANALYSIS:	TYPE = MIXTURE; STARTS = 40 10; PROC = 8;
OUTPUT:	TECH1 TECH10;

No MODEL command needed

The u thresholds/probabilities vary across all classes by default

Results for LCA Measurement Model

RANDOM STARTS RESULTS RANKED FROM THE BEST TO THE WORST LOGLIKELIHOOD VALUES

Final stage loglikelihood values at local maxima, seeds, and initial stage start numbers:

-3634.865	195873	6
-3634.865	392418	28
-3634.865	830392	35
-3634.865	915642	40
-3634.865	533738	11
-3634.865	76974	16
-3634.865	285380	1
-3634.865	939021	8
-3634.865	207896	25
-3634.865	unperturbed	0

- Best loglikelihood needs to be replicated several times

Results for LCA Measurement Model Continued

MODEL FIT INFORMATION

Number of Free Parameters	20
Loglikelihood	
H0 Value	-3634.865
Bayesian (BIC)	7407.885

FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES BASED ON THE ESTIMATED MODEL

Latent Classes		
1	262.95714	0.26296
2	291.27967	0.29128
3	445.76319	0.44576

Entropy = 0.681

Latent Class Probabilities and Logits

Average Latent Class Probabilities for Most Likely Latent Class Membership (Row) by Latent Class (Column)

	1	2	3
1	0.894	0.072	0.034
2	0.065	0.793	0.142
3	0.035	0.048	0.916

Classification Probabilities for the Most Likely Latent Class Membership (Column) by Latent Class (Row)

	1	2	3
1	0.864	0.079	0.058
2	0.063	0.866	0.071
3	0.019	0.101	0.880

Logits for the Classification Probabilities for the Most Likely Latent Class Membership (Column) by Latent Class (Row)

	1	2	3
1	2.706	0.310	0.000
2	-0.121	2.499	0.000
3	-3.826	-2.163	0.000

Results in Probability Scale for U=1 (as opposed to U=0)

- Class 1: High on all. Class 2: High on last 3. Class 3: Low on all

Variable	Class 1	Class 2	Class 3
U1	0.78	0.15	0.18
U2	0.81	0.15	0.22
U3	0.85	0.18	0.18
U4	0.73	**0.84**	0.21
U5	0.80	**0.80**	0.19
U6	0.77	**0.86**	0.22

- U1-U3 key for defining class 1
  - Only indicators with high probability for only class 1
  - See small P-values for class 1 vs the other two on next slide
- U4-U6 key for defining class 3
  - Only indicators differentiating between class 2 and class 3
  - See small P-values for class 3 vs the other two on next slide
- All U's needed for class 2
  - Some low, some high

Latent Class Indicator Discrimination Between Classes

EQUALITY TESTS OF PROBABILITIES
ACROSS THE LATENT CLASSES

	χ^2	P-Value	Df
U1			
Overall test	123.252	0.000	2
Class 1 vs. 2	83.642	0.000	1
Class 1 vs. 3	104.299	0.000	1
Class 2 vs. 3	0.440	0.507	1
U2			
Overall test	114.816	0.000	2
Class 1 vs. 2	76.826	0.000	1
Class 1 vs. 3	94.216	0.000	1
Class 2 vs. 3	1.850	0.174	1
U3			
Overall test	103.578	0.000	2
Class 1 vs. 2	72.540	0.000	1
Class 1 vs. 3	93.791	0.000	1
Class 2 vs. 3	0.005	0.946	1

	χ^2	P-Value	Df
U4			
Overall test	136.292	0.000	2
Class 1 vs. 2	3.783	0.052	1
Class 1 vs. 3	91.357	0.000	1
Class 2 vs. 3	83.173	0.000	1
U5			
Overall test	153.914	0.000	2
Class 1 vs. 2	0.001	0.979	1
Class 1 vs. 3	113.319	0.000	1
Class 2 vs. 3	95.138	0.000	1
U6			
Overall test	139.128	0.000	2
Class 1 vs. 2	2.828	0.093	1
Class 1 vs. 3	98.524	0.000	1
Class 2 vs. 3	74.422	0.000	1

- Shows for which classes which indicators discriminate well (low P-value) and discriminate poorly (high P-value)

Outline

- Background
- Measurement model analysis
- **Automatic multistep approaches**
  - Distal outcomes
  - Covariates
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

Automatic Multistep Approaches Using the AUXILIARY Option

- Single run analyzing the measurement model while adding the full model variables using the AUXILIARY option:
  - Distal outcome settings:
    - D3STEP, D3STEPC (for categorical distal), DU3STEP (unequal variances)
    - BCH, BCHC (for categorical distal)
    - Combinations of continuous and categorical distal outcomes
  - Covariate setting:
    - R3STEP
- Other settings for the AUXILIARY option listed in the last table of Web Note 21 not recommended

Automatic Approaches for Distal Outcomes: Pros and Cons of AUXILIARY D3STEP, D3STEPC, DU3STEP, BCH, BCHC

- Pros: Convenience
  - Single run
  - No need to save CPROBABILITIES and fix logits from the first to the last step or save BCH weights
- Cons: Limitations
  - The distal outcomes cannot be regressed on covariates
  - With several distal outcomes, the analysis is not done jointly for all of the outcomes but one at a time which means that missing data methodology is not used
    - Exception: TYPE = IMPUTATION as described in manual section
  - One latent class variable only

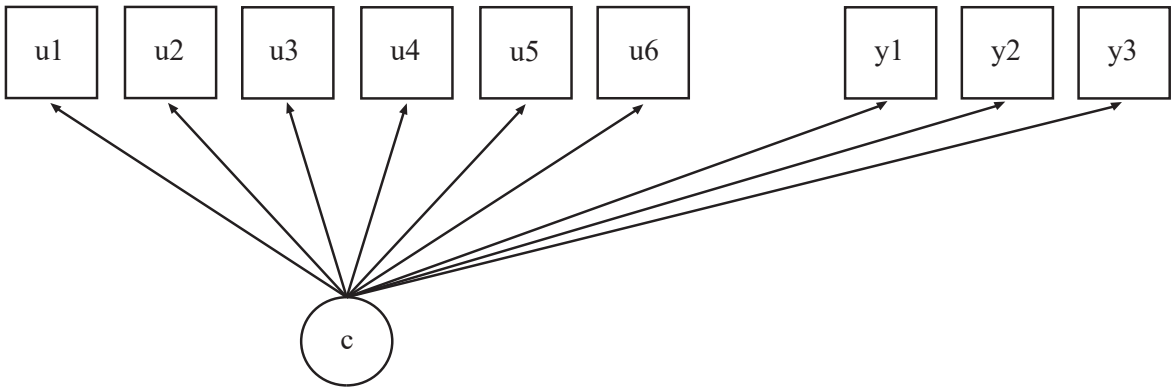
Automatic Approach for Covariates: Pros and Cons of AUXILIARY R3STEP

- Pros: Convenience
  - Single run
  - No need to save CPROBABILITIES and fix logits from the first to the last step
- Cons: Limitations
  - Observations with missing data on one or more of the covariates are deleted
    - Exception: TYPE = IMPUTATION as described in manual section
  - Direct effects from covariates to latent class indicators (DIF) cannot be included
  - One latent class variable only

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
  - **Distal outcomes**
  - Covariates analysis
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

Full Model: Distal Outcomes Y Added to the LCA Measurement Model



- Latent class indicators: u1-u6, binary
- Latent class variable: c
- Distal outcomes: y1 (continuous), y2 (binary), y3 (ordinal)
- The automatic approaches AUXILIARY BCH, BCHC will be shown here

Input for Auxiliary BCH

```
TITLE:      Auxiliary (BCH) for distal outcomes y1-y3

DATA:       FILE = distal.dat;

VARIABLE:   NAMES = u1-u6 x1 y2 y3 x2 y1 x3 class;
            USEVARIABLES = u1-u6;
            CATEGORICAL = u1-u6;
            MISSING = ALL(999);
            CLASSES = c(3);
            AUXILIARY = y1 (BCH) y2 (BCHC) y3 (BCHC);
            ! It is also possible to mix BCH options with D3STEP,
            ! D3STEPC, DU3STEP

ANALYSIS:   TYPE = MIXTURE;
            ESTIMATOR = MLR;
            STARTS = 40 10;
            PROC = 8;
```

No MODEL command needed

The u thresholds/probabilities vary across all classes by default

EQUALITY TESTS OF MEANS/PROBABILITIES ACROSS CLASSES

Y1
NUMBER OF DELETED OBSERVATIONS FOR THE AUXILIARY VARIABLE: 178
NUMBER OF OBSERVATIONS USED FOR THE AUXILIARY VARIABLE: 822

	Mean	S.E.
Class 1	-0.830	0.137
Class 2	0.337	0.140
Class 3	1.173	0.101

	Chi-Square	P-Value	Degrees of Freedom
Overall test	132.564	0.000	2
Class 1 vs. 2	31.083	0.000	1
Class 1 vs. 3	132.018	0.000	1
Class 2 vs. 3	20.047	0.000	1

AUXILIARY Tests for Binary Y2: ORs

EQUALITY TESTS OF MEANS/PROBABILITIES ACROSS CLASSES

Y2

NUMBER OF DELETED OBSERVATIONS FOR THE AUXILIARY VARIABLE: 187
NUMBER OF OBSERVATIONS USED FOR THE AUXILIARY VARIABLE: 813

	Prob	S.E.	Odds Ratio	S.E.	2.5% C.I.	97.5% C.I.
Class 1						
Category 1	0.324	0.037	1.000	0.000	1.000	1.000
Category 2	0.676	0.037	2.688	0.571	1.772	4.077
Class 2						
Category 1	0.600	0.040	1.000	0.000	1.000	1.000
Category 2	0.400	0.040	0.858	0.192	0.553	1.332
Class 3						
Category 1	0.563	0.030	1.000	0.000	1.000	1.000
Category 2	0.437	0.030	1.000	0.000	1.000	1.000

	Chi-Square	P-Value	Degrees of Freedom
Overall test	27.096	0.000	2
Class 1 vs. 2	20.295	0.000	1
Class 1 vs. 3	21.645	0.000	1
Class 2 vs. 3	0.465	0.495	1

AUXILIARY Tests for Ordinal Y3: ORs

NUMBER OF DELETED OBSERVATIONS FOR THE AUXILIARY VARIABLE: 0
NUMBER OF OBSERVATIONS USED FOR THE AUXILIARY VARIABLE: 1000

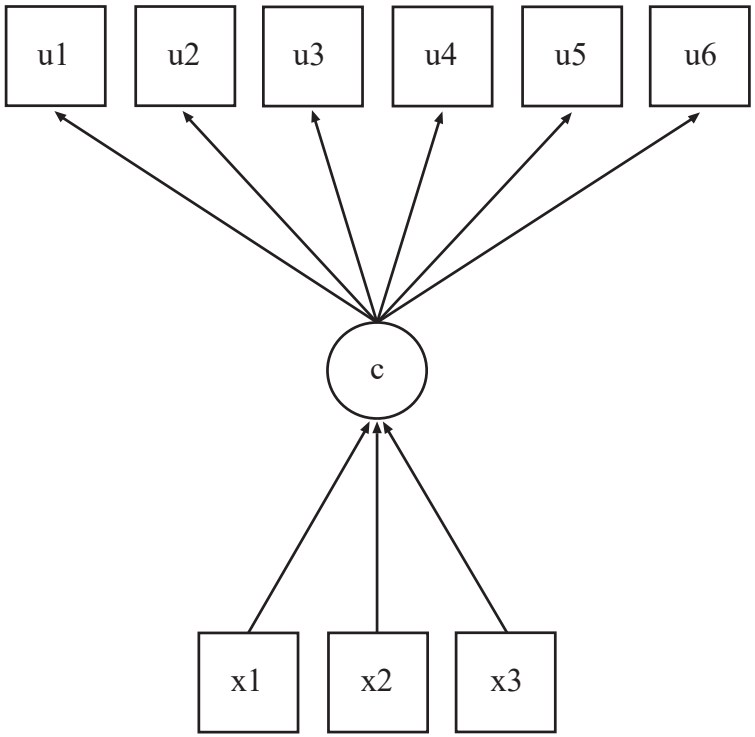
	Prob	S.E.	Odds Ratio	S.E.	2.5% C.I.	97.5% C.I.
Class 1						
Category 1	0.473	0.035	1.000	0.000	1.000	1.000
Category 2	0.033	0.013	0.998	0.182	0.698	1.429
Category 3	0.494	0.035	1.033	0.189	0.722	1.477
Class 2						
Category 1	0.412	0.036	1.000	0.000	1.000	1.000
Category 2	0.065	0.017	1.279	0.254	0.866	1.888
Category 3	0.523	0.036	1.157	0.227	0.787	1.701
Class 3						
Category 1	0.473	0.027	1.000	0.000	1.000	1.000
Category 2	0.041	0.011	1.000	0.000	1.000	1.000
Category 3	0.486	0.027	1.000	0.000	1.000	1.000

	Chi-Square	P-Value	Degrees of Freedom
Overall test	3.078	0.545	4
Class 1 vs. 2	2.680	0.262	2
Class 1 vs. 3	0.239	0.887	2
Class 2 vs. 3	2.243	0.326	2

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
  - Distal outcomes
  - **Covariates**
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

Full Model: Covariates X Added to the LCA Measurement Model



- Latent class indicators: u1-u6, binary
- Latent class variable: c
- Covariates: x1-x3
- Automatic approach: AUXILIARY R3STEP

Input for AUXILIARY R3STEP

```
TITLE:      Auxiliary (R3STEP) for covariates x1-x3

DATA:       FILE = distal.dat;

VARIABLE:   NAMES = u1-u6 x1 y2 y3 x2 y1 x3 class;
            USEVARIABLES = u1-u6;
            CATEGORICAL = u1-u6;
            MISSING = ALL(999);
            CLASSES = c(3);
            AUXILIARY = (R3STEP) x1 x2 x3;
            ! Note: x1, x2, x3 not consequtive on NAMES list
            ! so list function cannot be used
            ! (R3STEP) can alternatively be placed after
            ! each variable

ANALYSIS:   TYPE = MIXTURE;
            ESTIMATOR = MLR;
            STARTS = 40 10;
```

No MODEL command needed

The u thresholds/probabilities vary across all classes by default

Results for AUXILIARY R3STEP

WARNING: LISTWISE DELETION IS APPLIED TO THE AUXILIARY VARIABLES IN THE ANALYSIS. TO AVOID LISTWISE DELETION, DATA IMPUTATION CAN BE USED FOR THE AUXILIARY VARIABLES FOLLOWED BY ANALYSIS WITH TYPE=IMPUTATION.

NUMBER OF DELETED OBSERVATIONS: 332

NUMBER OF OBSERVATIONS USED: 668

Results for AUXILIARY R3STEP, Continued

ODDS RATIOS FOR CATEGORICAL LATENT VARIABLE MULTINOMIAL LOGISTIC REGRESSIONS USING THE 3-STEP (FIXED LOGITS) PROCEDURE

			95% C.I.	
	Estimate	S.E.	Lower 2.5%	Upper 2.5%
C#1 ON				
X1	0.257	0.070	0.151	0.438
X2	1.757	0.200	1.405	2.197
X3	0.561	0.099	0.398	0.792
C#2 ON				
X1	0.684	0.185	0.403	1.161
X2	1.313	0.167	1.024	1.684
X3	0.874	0.150	0.625	1.223

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- **Manual multistep approaches**
  - Distal outcomes
  - Distal outcomes and covariates with missing data
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

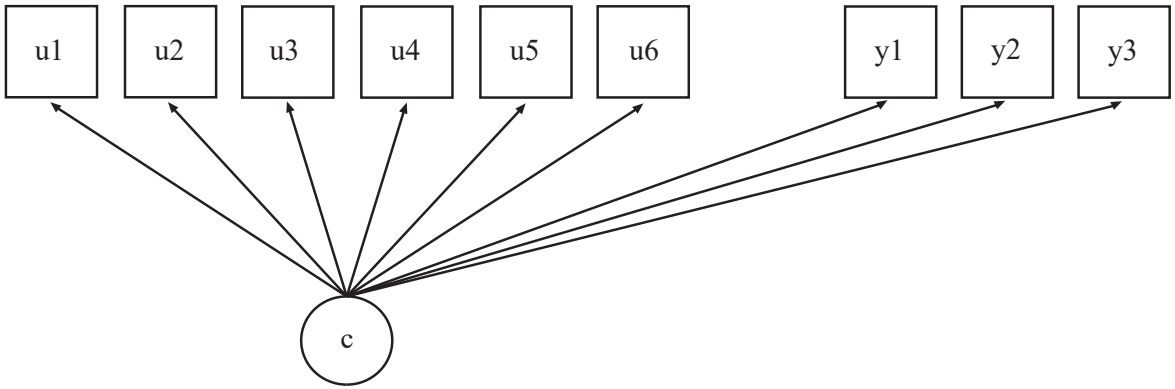
Manual Multistep Approaches

- Advantages of manual approaches:
  - Distal outcomes can be regressed on covariates
  - Observations with missing data on covariates do not have to be deleted
  - Direct effects from the covariates to the latent class indicators (measurement non-invariance, DIF) can be included
  - With several distal outcomes, the analysis is done jointly so that missing data methodology is used
  - More than one latent class variable can be handled (e.g., LTA)
- First step: Analysis of the measurement model
- Last step:
  - 3-step: Last step uses fixed logits from measurement model to correct for measurement error
  - BCH: Last step uses weights from measurement model to correct for measurement error
  - 2-step: Last step uses fixed measurement model parameters

Outline

- Background
- LCA measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
  - **Distal outcomes**
    - 3-step
    - BCH
    - 2-step
  - Distal outcomes and covariates with missing data
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

Full Model: Distal Outcomes Y Added to the LCA Measurement Model



- Latent class indicators: u1-u6, binary
- Latent class variable: c
- Distal outcomes:
  - y1: continuous, missing data
  - y2: binary, missing data
  - y3: ordinal, no missing data
- Joint analysis of y1, y2, y3:
  - Missing data methodology applied (FIML)

Input for 3-Step, First Step

TITLE:	Manual 3-step, first step Same as measurement model, but adding SAVEDATA of cprobabilities Entropy= 0.681
DATA:	**FILE = distal.dat;**
VARIABLE:	NAMES = u1-u6 x1 y2 y3 x2 y1 x3 class; USEVARIABLES = u1-u6; CATEGORICAL = u1-u6; MISSING = ALL(999); CLASSES = c(3); **AUXILIARY = y1 y2 y3;**
ANALYSIS:	TYPE = MIXTURE; STARTS = 40 10; PROCESSORS = 8;
SAVEDATA:	**SAVE = CPROB;** **FILE = cproby.dat;**

3-Step Saved Data

SAVEDATA INFORMATION

Save file
cproby.dat

Order and format of variables

U1	F10.3
U2	F10.3
U3	F10.3
U4	F10.3
U5	F10.3
U6	F10.3
Y1	F10.3
Y2	F10.3
Y3	F10.3
CPROB1	F10.3
CPROB2	F10.3
CPROB3	F10.3
MLCC	F10.3

Save file format	13F10.3
Save file record length	10000
Save missing symbol	*

Latent Class Probabilities and Logits

Average Latent Class Probabilities for Most Likely Latent Class Membership (Row) by Latent Class (Column)

	1	2	3
1	0.894	0.072	0.034
2	0.065	0.793	0.142
3	0.035	0.048	0.916

Classification Probabilities for the Most Likely Latent Class Membership (Column) by Latent Class (Row)

	1	2	3
1	0.864	0.079	0.058
2	0.063	0.866	0.071
3	0.019	0.101	0.880

Logits for the Classification Probabilities for the Most Likely Latent Class Membership (Column) by Latent Class (Row)

	1	2	3
1	2.706	0.310	0.000
2	-0.121	2.499	0.000
3	-3.826	-2.163	0.000

Input for 3-Step, Last Step

TITLE: Manual 3-step, last step for distal outcomes y1-y3

DATA: **FILE = cproby.dat**; ! from first step

VARIABLE:
NAMES = u1-u6 y1-y3 cprob1-cprob3 **n;**
! n is the same as MLCC
USEVARIABLES = y1-y3 n;
CATEGORICAL = y2 y3;
NOMINAL = n;
MISSING = *;
CLASSES = c(3);

ANALYSIS:
TYPE = MIXTURE;
STARTS = 0;
PROCESSORS = 8;

MODEL:
%c#1%
[n#1@2.706 n#2@0.310];
%c#2%
[n#1@-0.121 n#2@2.499];
%c#3%
[n#1@-3.826 n#2@-2.163];
! The y means/thresholds
! vary across the classes by default

Input for BCH, First Step

TITLE:	Manual BCH, first step
DATA:	FILE = distal.dat;
VARIABLE:	NAMES = u1-u6 x1 y2 y3 x2 y1 x3 class; USEVARIABLES = u1-u6; CATEGORICAL = u1-u6; MISSING = ALL(999); CLASSES = c(3); **AUXILIARY = y1 y2 y3;**
ANALYSIS:	TYPE = MIXTURE; STARTS = 40 10; PROCESSORS = 8;
SAVEDATA:	**SAVE = BCHWEIGHTS;** **FILE = bch.dat;**

BCH Saved Data

Save file	
bch.dat	
Order and format of variables	
U1	F10.3
U2	F10.3
U3	F10.3
U4	F10.3
U5	F10.3
U6	F10.3
Y1	F10.3
Y2	F10.3
Y3	F10.3
BCHW1	F10.3
BCHW2	F10.3
BCHW3	F10.3
Save file format	
12F10.3	
Save file record length	10000
Save missing symbol	*

Input for BCH, Last Step

TITLE:	Manual BCH, last step for distal outcomes y1-y3
DATA:	**FILE = bch.dat;**
VARIABLE:	NAMES = u1-u6 y1-y3 w1-w3; **USEVARIABLES = y1-y3;** CATEGORICAL = y2 y3; **MISSING = *;** CLASSES = c(3); **TRAINING = w1-w3(BCH);**
ANALYSIS:	TYPE = MIXTURE; **STARTS = 0;** PROCESSORS = 8;

The y means/thresholds vary across the classes by default

Input for 2-Step, First Step

TITLE:	Manual 2-step, first step
DATA:	FILE = distal.dat;
VARIABLE:	NAMES = u1-u6 x1 y2 y3 x2 y1 x3 class; USEVARIABLES = u1-u6; CATEGORICAL = u1-u6; MISSING = ALL(999); CLASSES = c(3);
ANALYSIS:	TYPE = MIXTURE; STARTS = 40 10; PROCESSORS = 8;
OUTPUT:	**SVALUES;**

MODEL COMMAND WITH FINAL ESTIMATES USED AS STARTING VALUES

```
%OVERALL%

[ c#1*-0.52780 ];
[ c#2*-0.42550 ];

%C#1%

[ u1$1*-1.28835 ];
[ u2$1*-1.41536 ];
[ u3$1*-1.74521 ];
[ u4$1*-1.00460 ];
[ u5$1*-1.36393 ];
[ u6$1*-1.17851 ];

%C#2%

[ u1$1*1.72069 ];
[ u2$1*1.73016 ];
[ u3$1*1.54348 ];
[ u4$1*-1.68227 ];
[ u5$1*-1.37222 ];
[ u6$1*-1.82568 ];

%C#3%

[ u1$1*1.52507 ];
[ u2$1*1.28543 ];
[ u3$1*1.52212 ];
[ u4$1*1.30734 ];
[ u5$1*1.42798 ];
[ u6$1*1.29213 ];
```

Use the Mplus Editor to replace * with @. Make sure no unwanted changes of * occur such as the missing data flag in MISSING = *;

Input for 2-Step, Last Step

```
TITLE:      Manual 2-step, last step

DATA:       FILE = distal.dat;

VARIABLE:   NAMES = u1-u6 x1 y2 y3 x2 y1 x3 class;
            USEVARIABLES = u1-u6 y1 y2 y3;
            CATEGORICAL = u1-u6 y2 y3;
            MISSING = ALL(999);
            CLASSES = c(3);

ANALYSIS:   TYPE = MIXTURE;
            STARTS = 0;
            PROCESSORS = 8;

MODEL:      ! Only measurement parameters
            %C#1%
            [ u1$1@-1.28835 ];
            [ u2$1@-1.41536 ];
            [ u3$1@-1.74521 ];
            [ u4$1@-1.00460 ];
            [ u5$1@-1.36393 ];
            [ u6$1@-1.17851 ];

            %C#2%

            [ u1$1@1.72069 ];
            [ u2$1@1.73016 ];
            [ u3$1@1.54348 ];
            [ u4$1@-1.68227 ];
            [ u5$1@-1.37222 ];
            [ u6$1@-1.82568 ];

            %C#3%

            [ u1$1@1.52507 ];
            [ u2$1@1.28543 ];
            [ u3$1@1.52212 ];
            [ u4$1@1.30734 ];
            [ u5$1@1.42798 ];
            [ u6$1@1.29213 ];
```

The y means/thresholds
vary across the classes by default

Odds Ratio Results when Using the Manual Approach

- The automatic approach produces odds ratios and their non-symmetric CIs for distal outcomes that are binary and ordinal
- Using the manual approach, these results are found under the heading LATENT CLASS INDICATOR ODDS RATIOS FOR THE LATENT CLASSES
    - These computations can also be done using MODEL CONSTRAINT as shown in Mplus Web Talk No. 2, slides 17-20

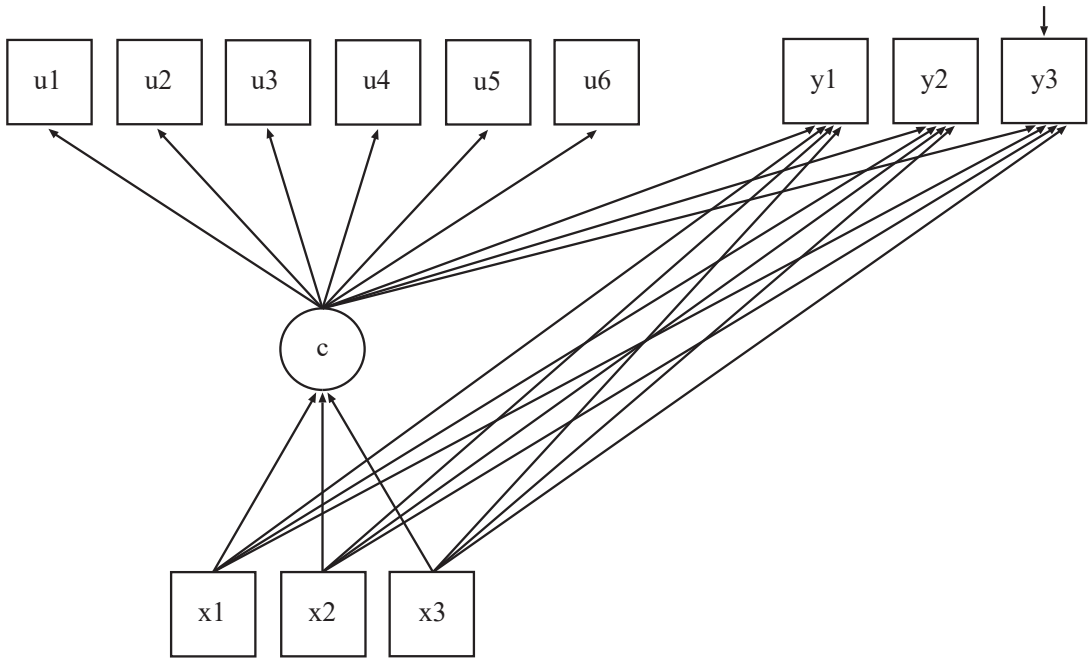
Chi-Square Testing when Using the Manual Approach

- The automatic approach produces chi-square testing of intercept/threshold differences across classes
- Using the manual approach, these are found under the heading EQUALITY TESTS OF MEANS/PROBABILITIES ACROSS THE LATENT CLASSES
  - Such testing is also available with covariates when the regressions of the distal outcomes on the covariates are not class-varying

Outline

- Background
- LCA measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
  - Distal outcomes
  - **Distal outcomes and covariates with missing data**
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

Full Model Based on an LCA Measurement Model



- Latent class indicators: u1-u6 binary
- Distal outcomes: y1 continuous, y2 (binary), y3 ordinal (3 categories). Missing data on y1 and y2 but not y3
- Covariates: x1 binary, x2-x3 continuous. Missing data on x1-x2
- Slopes for the regression of y on x not class varying: class-varying y intercepts/thresholds show effect of c on y
- The Further Topics section discusses class-varying y on x

Full Model, Continued

- Monte Carlo generated data, 1 replication, N = 1000, entropy = 0.740
- Missing data on covariates: Covariance coverage

	X1	X2	X3
X1	0.810		
X2	0.668	0.814	
X3	0.810	0.814	1.000

- 19% missing for both the x1 and x2 covariates
- In line with regression analysis, individuals with missing on one or more of the covariates are deleted:
  - Sample size drops from 1000 to 668, a 33% loss
- This loss can be avoided by bringing the covariates into the model, that is, estimating also their marginal distribution by covarying them using x1-x3 WITH x1-x3 or by using multiple imputation

Manual Multistep Approaches with Missing Data on X In the Last Step Analysis

- Covariates brought into the model:
  - ML and numerical integration
    - 3-step, BCH not available with numerical integration, 2-step
    - Pros: Likelihood-ratio testing available
    - Cons: Cannot handle many covariates with missing, can be somewhat unstable with large amounts of missing data, can be imprecise without many integration points, cannot specify categorical covariates
  - Bayes
    - 3-step, BCH not available due to weights, 2-step
    - Pros: Stable
    - Cons: No LRT, can be slow, cannot specify categorical covariates
- Multiple imputation:
  - 3-step, BCH, 2-step
  - Pros: Stable, can specify categorical covariates, can use non-model variables for missing data imputation
  - Cons: No LRT, slow with many covariates with missing
- 3-step demonstrated here for all 3 approaches

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
  - Distal outcomes
  - Distal outcomes and covariates with missing data
    - **ML and integration using 3-step**
    - Bayes using 3-step
    - Multiple imputation using 3-step
    - Combined imputation and final analysis using 3-step, BCH, 2-step
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

First Step of 3-Step

```
TITLE:      Covariates missing, 3-step, first step
            saving y and x

DATA:       FILE = distal.dat;

VARIABLE:   NAMES = u1-u6 x1 y2 y3 x2 y1 x3 classes;
            USEVARIABLES = u1-u6;
            CATEGORICAL = u1-u6;
            MISSING = ALL(999);
            CLASSES = c(3);
            ! Saving variables for last step:
            AUXILIARY = x1 x2 x3 y1 y2 y3;

ANALYSIS:   TYPE = MIXTURE;
            ESTIMATOR = MLR;
            STARTS = 40 10;

SAVEDATA:   FILE = cprobyx.dat;
            SAVE = CPROB;
```

SAVEDATA Information: 3-step Saved Data

Save file
cprobyx.dat
Order and format of variables

U1	F10.3
U2	F10.3
U3	F10.3
U4	F10.3
U5	F10.3
U6	F10.3
X1	F10.3
X2	F10.3
X3	F10.3
Y1	F10.3
Y2	F10.3
Y3	F10.3
CPROB1	F10.3
CPROB2	F10.3
CPROB3	F10.3
MLCC	F10.3

Save file format
16F10.3
Save file record length 10000
Save missing symbol *

Logits for 3-Step LCA

Logits for the Classification Probabilities for the Most Likely Latent Class Membership (Column) by Latent Class (Row)

1	2.706	0.310	0.000
2	-0.121	2.499	0.000
3	-3.826	-2.163	0.000

Last Step of 3-Step with Integration

```
TITLE:      Manual, covariates missing, 3-step,
            last step, integration

DATA:       FILE = cprobyx.dat; ! from first step

VARIABLE:   NAMES = u1-u6 x1-x3 y1-y3 cprob1-cprob3 n;
            USEVARIABLES = n x1-x3 y1-y3;
            NOMINAL = n;
            CATEGORICAL = y1 y2;
            MISSING = *;
            CLASSES = c(3);

ANALYSIS:   TYPE = MIXTURE;
            ESTIMATOR = MLR;
            STARTS = 0;
            ALGORITHM = INTEGRATION;
            ! 2-dimensions of integration due to missing on x1, x2
            INTEGRATION = MONTECARLO;
            ! MONTECARLO integration due to partially missing
            ! x variables (not latent variables)
            ! Default of 500 points. With missing data on many x's,
            ! MONTECARLO(5000) may be needed
```

Input continues on next slide

Last Step of 3-Step with Integration Continued

```
MODEL:      %OVERALL%
            c ON x1-x3;
            y1-y3 ON x1-x3;
            x1-x3 WITH x1-x3;
            ! Disadvantage: The binary x1 is treated
            ! as continuous-normal
            ! See Statistical Methodology section

            %c#1%
            [n#1@2.706 n#2@0.310];

            %c#2%
            [n#1@-0.121 n#2@2.499];

            %c#3%
            [n#1@-3.826 n#2@-2.163];

OUTPUT:     TECH8;
```

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
  - Distal outcomes
  - Distal outcomes and covariates with missing data
    - ML and integration: 3-step
    - **Bayes: 3-step**
    - Multiple imputation: 3-step
    - Combined imputation and final analysis: 3-step, BCH, 2-step
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

3-Step using Bayes in the Last Step
No Numerical Integration Needed

```
TITLE:      Manual, covariates missing, 3-step,
            last step, Bayes

DATA:       FILE = cprobyx.dat;

VARIABLE:   NAMES = u1-u6 x1-x3 y1-y3 cprob1-cprob3 n;
            USEVARIABLES = n x1-x3 y1-y3;
            NOMINAL = n;
            CATEGORICAL = y1 y2;
            MISSING = *;
            CLASSES = c(3);

ANALYSIS:   TYPE = MIXTURE;
            ESTIMATOR = BAYES;
            BITERATIONS = (5000);
            STARTS = 0;
```

```
MODEL:      %OVERALL%
            c ON x1-x3;
            y1-y3 ON x1-x3;
            x1-x3 WITH x1-x3;

            %c#1%
            [n#1@2.706 n#2@0.310];
            %c#2%
            [n#1@-0.121 n#2@2.499];
            %c#3%
            [n#1@-3.826 n#2@-2.163];

OUTPUT:     TECH1 TECH8;
            ! Effective sample size (ESS)
            ! calls for close to 20,000 iterations
```

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
  - Distal outcomes
  - Distal outcomes and covariates with missing data
    - ML and integration: 3-step
    - Bayes: 3-step
    - **Multiple imputation: 3-step**
    - Combined imputation and final analysis: 3-step, BCH, 2-step
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

Multiple Imputation to Handle Missing on X

- Multiple imputation not generally needed for missing data in the latent class indicators of the measurement model or for missing data in the distal outcomes - FIML sufficient
- Multiple imputation useful for missing data on covariates
  - Imputation model can bring in variables related to missing data but not in the model
  - No numerical integration required after imputation due to no missing data
  - 3-step
  - BCH
  - 2-step
- Asparouhov& Muthén (2025). Multiple imputation with Mplus. Technical Report.
  `https://statmodel.com/download/Imputations7.pdf`
  - Section 4 has17 general tips

3-Step With Multiple Imputation: Imputation Step with Data from First Step of 3-Step

TITLE: Covariates missing 3-step Imputation step
DATA: FILE = cprobyx.dat;

VARIABLE: NAMES = u1-u6 x1-x3 y1-y3 cprob1-cprob3 n;
USEVARIABLES = x1-x3 y1-y3; ! Variables used in
! the BASIC analysis to provide sample statistics and
! saved in addition to variables on the AUXILIARY list
AUXILIARY = n;
MISSING = *;

DATA IMPUTATION: **NAMES = u1-u6(C) x1(C) x2 x3 y1 y2(C) y3(C);**
IMPUTE = x1(C) x2;
! The imputations use information from the variables
! on the DATA IMPUTATION NAMES list and
! imputes values for variables on the IMPUTE list.
! NAMES list must include all USEV variables
NDATASETS = 100; THIN = 100;
SAVE = imp*.dat;

ANALYSIS: **TYPE = BASIC**; ! Sample statistics for USEV variables
! computed after multiple imputation done by Bayes
BITERATIONS = (1000); ! Monitor by TECH8

Specifications for Data Imputation

Number of imputed data sets	100
H1 imputation model type	COVARIANCE
Iteration intervals for thinning	100
Number of dependent variables	12
Number of independent variables	0
Number of parameters	82

- The number of variables (12) and H1 parameters (82) are determined by the DATA IMPUTATION NAMES list
  - NAMES = u1-u6(C) x1(C) x2 x3 y1 y2(C) y3(C)
- The number of H1 parameters is computed as:
  - Number of covariances/correlations = 12*(12-1)/2 = 66
  - Number of means and variances: 3 + 3 = 6
  - Number of thresholds: 10
- With many variables, the number of H1 parameters can get prohibitive

SAVEDATA Information for 3-Step Imputation Step

SAVEDATA INFORMATION	
Save file imp*.dat	
Order of variables	
X1 X2 X3 Y1 Y2 Y3 N	
Save file format	Free
Save file record length	10000
Save missing symbol	*

Last Step of 3-Step after Imputation

```
TITLE:      Last step of 3-step after imputation

DATA:       FILE = implist.dat;
            TYPE = IMPUTATION;

VARIABLE:   NAMES = x1-x3 y1-y3 n;
            USEVARIABLES = n x1-x3 y1-y3;
            NOMINAL = n;
            CATEGORICAL = y1 y2;
            MISSING = *;
            CLASSES = c(3);

ANALYSIS:   TYPE = MIXTURE;
            ESTIMATOR = MLR;
            STARTS = 0;

MODEL:      %OVERALL%
            c ON x1-x3;
            y1-y3 ON x1-x3;
            ! Note: No x part because
            ! the x's have no missing data

            %c#1%
            [n#1@2.706 n#2@0.310];
            %c#2%
            [n#1@-0.121n#2@2.499];
            %c#3%
            [n#1@-3.826 n#2@-2.163];

OUTPUT:     TECH8;
```

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
  - Distal outcomes
  - Distal outcomes and covariates with missing data
    - ML and integration: 3-step
    - Bayes: 3-step
    - Multiple imputation: 3-step
    - **Combined imputation and final analysis: 3-step, BCH, 2-step**
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

Combined Approach

- Stage A: Measurement model analysis, saving data for all 3 methods: 3-step, BCH, 2-step
- Stage B: Combined imputation and final analysis

- Advantage: Keeps the number of datasets to a minimum
- Multiple imputations:
  - Can be saved and used for many final analyses with DATA: TYPE=IMPUTATION
  - Multiple imputation can be skipped if little or no missing data on covariates, or replaced by integration or Bayes

Analysis Stages for Combined Approach

A. Measurement Model Analysis

Input Specifications	Variables Saved
USEV = u1-u6	u1-u6
AUXILIARY (for final stage)	x1-x3, y1-y3
CPROB (for 3-step)	cprob1-cprob3, n
BCHWEIGHTS (for BCH)	w1-w1
SVALUES (for 2-step)	

B. Final Analysis (Multiple imputation + Full model analysis)

DATA IMPUTATION for missing on X (can be replaced by integration)

3-Step	BCH	2-Step
USEV = n, x1-x3, y1-y3	USEV = x1-x3, y1-y3	USEV = u1-u6, x1-x3, y1-y3
Logits from Stage A	TRAINING = w1-w3(BCH)	SVALUES from Stage A
(manually added)		(manually added)

Combined Approach: Stage A, Measurement Model

TITLE: Combined approach, stage A, Measurement Model
First step for all 3 methods

DATA: FILE = distal.dat;

VARIABLE: NAMES = u1-u6 x1 y2 y3 x2 y1 x3 class;
USEVARIABLES = u1-u6;
MISSING = ALL(999);
CATEGORICAL = u1-u6;
CLASSES = c(3);
AUXILIARY = x1 x2 x3 y1 y2 y3;

ANALYSIS: TYPE = MIXTURE;
STARTS = 80 20;

SAVEDATA: **FILE = stageA.dat;**
SAVE = CPROB BCHWEIGHTS;

OUTPUT: **SVALUES;**

- Needs reordering of classes: SVALUES(2 1 3);

SAVEDATA Information: Stage A Measurement Analysis

Save file stagea.dat Order and format of variables	
U1	F10.3
U2	F10.3
U3	F10.3
U4	F10.3
U5	F10.3
U6	F10.3
X1	F10.3
X2	F10.3
X3	F10.3
Y1	F10.3
Y2	F10.3
Y3	F10.3
BCHW1	F10.3
BCHW2	F10.3
BCHW3	F10.3
CPROB1	F10.3
CPROB2	F10.3
CPROB3	F10.3
MLCC	F10.3

Combined Approach: Stage B, Final Step for 3-Step

TITLE: Combined Approach: Stage B,
Final Step for 3-Step
Imputation + Analysis in one step

DATA: **FILE = stageA.dat;**

VARIABLE: NAMES = u1-u6 x1-x3 y1-y3 w1-w3 cp1-cp3 n;
USEVARIABLES = n x1-x3 y1-y3; ! Refers to
! variables in the MODEL command
NOMINAL = n;
CATEGORICAL = y2 y3;
MISSING = *;
CLASSES= c(3);

DATA IMPUTATION: **NAMES = u1-u6(C) x1(C) x2 x3**
y1 y2(C) y3(C);
IMPUTE = x1(C) x2;
NDATASETS = 100;
THIN = 10;
SAVE = imputed*.dat; ! not needed

ANALYSIS: TYPE = MIXTURE;
ESTIMATOR = MLR;
STARTS = 0;

MODEL: %OVERALL%
c ON x1-x3;
y1-y3 ON x1-x3;
! Note: No x part because
! the x's no longer
! have missing data

%c#1%
[n#1@2.706 n#2@0.310];
%c#2%
[n#1@-0.121 n#2@2.499];
%c#3%
[n#1@-3.826 n#2@-2.163];

OUTPUT: TECH8;

Combined Approach: Stage B, Final Step for BCH

TITLE: Combined Approach: Stage B,
Final Step for BCH
Imputation + Analysis in one step

DATA: FILE = stageA.dat;

VARIABLE: NAMES = u1-u6 x1-x3 y1-y3 w1-w3 cp1-cp3 n;
USEVARIABLES = x1-x3 y1-y3;
CATEGORICAL = y2 y3;
MISSING = *;
CLASSES = c(3);
TRAINING = w1-w3(BCH);

DATA IMPUTATION: **NAMES = u1-u6(C) x1(C) x2 x3 y1 y2(C) y3(C);**
IMPUTE = x1(C) x2;
NDATASETS = 100;
THIN = 10;
SAVE = imputed*.dat; ! not needed

ANALYSIS: TYPE = MIXTURE;
ESTIMATOR = MLR;
STARTS = 0;

MODEL: %OVERALL%
c ON x1-x3;
y1-y3 ON x1-x3;

OUTPUT: TECH8;

Combined Approach: Stage B, Final Analysis for 2-Step

TITLE: Final step for 2-step based on Combined runs
Imputation + Analysis in one step

DATA: FILE = stageA.dat;

VARIABLE:
NAMES = u1-u6 x1-x3 y1-y3 w1-w3 cp1-cp3 n;
USEVARIABLES = u1-u6 x1-x3 y1-y3;
MISSING = *;
CATEGORICAL = u1-u6 y2 y3;
CLASSES = c(3);

DATA IMPUTATION:
NAMES = u1-u6(C) x1(C) x2 x3 y1 y2(C) y3(C);
IMPUTE = x1(C) x2;
! the imputation uses information from the variables
! on the DATA IMPUTATION NAMES list and
! imputes values for variables on the IMPUTE list
NDATASETS = 100;
THIN = 10;
SAVE = imputed*.dat; ! Not needed

Input continues on the next slide

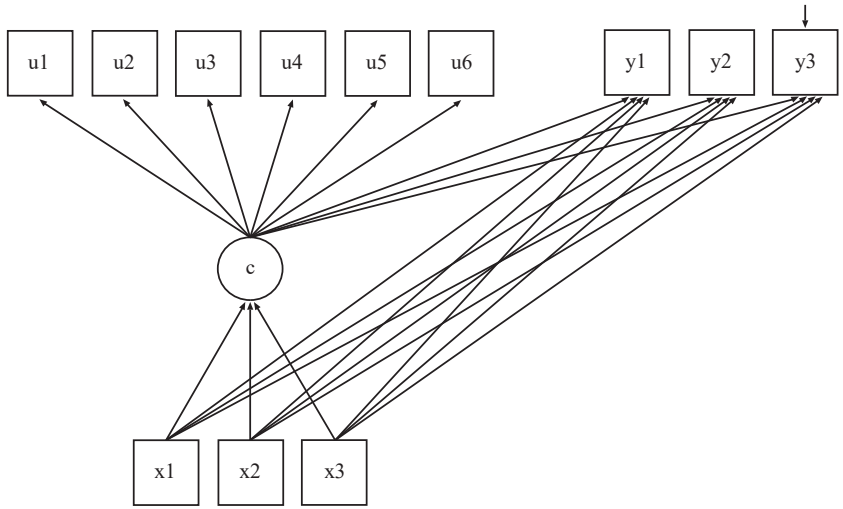
Stage B, Final Analysis for 2-Step Continued

```
ANALYSIS:   TYPE = MIXTURE;
            ESTIMATOR = MLR;
            STARTS = 0;
            ! Can add BITER for the imputation

MODEL:      %OVERALL%
            c ON x1-x3;
            y1-y3 ON x1-x3;
            ! Note: No x part because
            ! the x's have no missing data

            %C#1%

            [ u1$1*-1.28835 ];
            [ u2$1*-1.41536 ];
            [ u3$1*-1.74521 ];
            [ u4$1*-1.00460 ];
            [ u5$1*-1.36393 ];
            [ u6$1*-1.17851 ];

            %C#2%

            [ u1$1*1.72069 ];
            [ u2$1*1.73016 ];
            [ u3$1*1.54348 ];
            [ u4$1*-1.68227 ];
            [ u5$1*-1.37222 ];
            [ u6$1*-1.82568 ];

            %C#3%

            [ u1$1*1.52507 ];
            [ u2$1*1.28543 ];
            [ u3$1*1.52212 ];
            [ u4$1*1.30734 ];
            [ u5$1*1.42798 ];
            [ u6$1*1.29213 ];
```

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
- **Results comparing manual approaches**
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references

Results Comparing Manual Approaches: Guide to Output



- Where in the output are key results found?
  - X →Y: Y ON X
    - Continuous Y: Linear regression slopes (MODEL RESULTS)
    - Categorical Y: Logistic regression odds ratios
  - C →Y: [Y], [Y$1]
    - Continuous Y: Linear regression intercepts (MODEL RESULTS)
    - Categorical Y: Odds ratios
  - X →C: C ON X
    - Multinomial logistic regression odds ratios
- Example: BCH

X →Y Results for Final Step of BCH

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	Rate of Missing
Latent Class 1					
Y1 ON					
X1	0.593	0.101	5.863	0.000	0.166
X2	0.553	0.043	12.811	0.000	0.130
X3	0.530	0.059	8.971	0.000	0.108

LOGISTIC REGRESSION ODDS RATIO RESULTS FOR OBSERVED VARIABLES

	Estimate	S.E.	95% C.I. Lower 2.5%	95% C.I. Upper 2.5%
Latent Class 1				
Y2 ON				
X1	2.008	0.431	1.318	3.059
X2	1.501	0.139	1.251	1.800
X3	1.670	0.218	1.293	2.156

- Same for all classes

C →Y Results for Final Step of BCH: Continuous Y

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	Rate of Missing
Latent Class 1					
Intercepts					
Y1	-1.014	0.096	-10.563	0.000	0.094
Latent Class 2					
Intercepts					
Y1	-0.164	0.105	-1.556	0.120	0.083
Latent Class 3					
Intercepts					
Y1	0.972	0.097	9.975	0.000	0.067

C → Y Results for Final Step of BCH: Categorical Y

LATENT CLASS INDICATOR ODDS RATIOS FOR THE LATENT CLASSES

	Estimate	S.E.	95% C.I. Lower 2.5%	95% C.I. Upper 2.5%
Latent Class 1 Compared to Latent Class 2				
Y2				
Category > 1	5.527	1.723	3.000	10.183
Y3				
Category > 1	0.849	0.219	0.512	1.407
Category > 2	1.017	0.259	0.617	1.677
Latent Class 1 Compared to Latent Class 3				
Y2				
Category > 1	3.900	1.081	2.265	6.715
Y3				
Category > 1	0.969	0.217	0.625	1.502
Category > 2	1.007	0.228	0.646	1.570
Latent Class 2 Compared to Latent Class 3				
Y2				
Category > 1	0.706	0.187	0.420	1.188
Y3				
Category > 1	1.142	0.263	0.726	1.795
Category > 2	0.991	0.226	0.634	1.549

X →C Results for Final Step of BCH

LOGISTIC REGRESSION ODDS RATIO
RESULTS FOR LATENT CLASS VARIABLES

	Estimate	S.E.	95% C.I. Lower 2.5%	95% C.I. Upper 2.5%
C#1 ON				
X1	0.295	0.072	0.183	0.475
X2	1.700	0.181	1.379	2.095
X3	0.695	0.103	0.520	0.928
C#2 ON				
X1	0.620	0.153	0.383	1.005
X2	1.328	0.147	1.069	1.651
X3	0.956	0.141	0.716	1.277

Reminder of Available Multistep Approaches with Missing Data on Covariates

- ML and numerical integration
  - 3-step
  - BCH: Not available with numerical integration
  - 2-step
- Bayes
  - 3-step
  - BCH: Not available due to weights
  - 2-step
- Multiple imputation
  - 3-step
  - BCH
  - 2-step
- Combined imputation and final analysis
  - 3-step
  - BCH
  - 2-step

Basis for Comparison of Multistep Approaches with Missing Data on Covariates

- Comparison of multistep results can be made visavis the full model for indicators, covariates, and distal outcomes:
  - Correct full model that generated the data (typically unknown) using Monte Carlo simulation
    - The data generating MODEL POPULATION specifies the probabilities of missing data on x1, x2 as functions of observed x3 values; see Appendix slide 135
    - The MODEL that analyzes the generated data correctly specifies x1 x2 ON x3; see Appendix slide 137
  - 1-step analysis of the data generated by the correct model brings x1, x2, x3 into the model using WITH instead of ON because the missing data mechanism is unknown
    - x1-x3 WITH x1-x3
    - This is an approximation of the correct model but one that more closely represents real-data analysis
- A key comparison is with the class probabilities of the measurement model

1-Step ML with Integration

TITLE: 1-step ML with integration

DATA: FILE = distal.dat;

VARIABLE:
names = u1-u6 x1 y2 y3 x2 y1 x3 class;
USEVARIABLES = u1-u6 x1 x2 x3 y1 y2 y3;
CATEGORICAL = u1-u6 y2 y3;
MISSING = ALL(999);
CLASSES = c(3);

ANALYSIS:
TYPE = MIXTURE;
STARTS = 20 4;
ALGORITHM = INTEGRATION;
INTEGRATION = MONTECARLO;
PROCESSORS = 8;

MODEL:
%OVERALL%
c ON x1-x3;
y1-y3 ON x1-x3;
x1-x3 WITH x1-x3;
! 2-dimensional integration due to missing on x1, x2
! Disadvantage: The binary x1 is treated as continuous

LCA Class Probabilities: ML, Integration

ML, integration	
Correct full model (x1 x2 ON x3; slide 137)	0.263, 0.318, 0.419
1-step (x1-x3 WITH x1-x3)	0.264, 0.316, 0.420
3-step, first step (measurement model)	0.263, 0.291, 0.446
Last step (x1-x3 WITH x1-x3)	0.266, 0.291, 0.443
BCH cannot do integration	
2-step, first step (measurement model)	0.263, 0.291, 0.446
Last step (x1-x3 WITH x1-x3)	0.264, 0.293, 0.443

LCA Class Probabilities: Bayes and Imputation

Bayes in last step	
3-step, first step (measurement model)	0.263, 0.291, 0.446
Last step (x1-x3 WITH x1-x3)	0.266, 0.294, 0.439
BCH step 2 cannot be done by Bayes	
2-step, first step (measurement model)	0.263, 0.291, 0.446
Last step (x1-x3 WITH x1-x3)	0.274, 0.283, 0.442
Multiple imputation using combined approach	
3-step, first step (measurement model)	0.263, 0.291, 0.446
Last step	0.266, 0.296, 0.438
BCH, first step (measurement model)	0.263, 0.291, 0.446
Last step	**0.263, 0.291, 0.446**
2-step, first step (measurement model)	0.263, 0.291, 0.446
Last step	0.265, 0.296, 0.439

LCA: 1-Step with Integration

Class probabilities: 0.264, 0.316, 0.420

	$X \rightarrow Y$		$C \rightarrow Y$	
Covariate	Y1 Slopes	Y2 OR's	Y1 Intercepts	Y2 OR's
X1	0.556*	1.840	C#1 -1.019*	3.000 (1 vs 3)*
X2	0.580*	1.529*	C#2 -0.098	0.685 (2 vs 3)
X3	0.516*	1.638*	C#3 1.055*	

$X \rightarrow C$

Covariate	C#1 OR's (1 vs 3)	C#2 OR's (2 vs 3)
X1	0.254*	0.579*
X2	1.823*	1.393*
X3	0.660*	0.928

LCA: 3-Step with Imputation, Combined Approach, ML

Class probabilities: 0.266, 0.296, 0.438

	$X \rightarrow Y$		$C \rightarrow Y$	
Covariate	Y1 Slopes	Y2 OR's	Y1 Intercepts	Y2 OR's
X1	0.555*	1.941*	C#1 -0.983*	3.562 (1 vs 3)
X2	0.563*	1.522*	C#2 -0.135	0.685 (2 vs 3)
X3	0.534*	1.645*	C#3 1.018*	

$X \rightarrow C$

Covariate	C#1 OR's (1 vs 3)	C#2 OR's (2 vs 3)
X1	0.281*	0.572*
X2	1.724*	1.388*
X3	0.701*	0.926

LCA: BCH with Imputation, Combined Approach, ML

Class probabilities: 0.263, 0.291, 0.446

	$X \rightarrow Y$		$C \rightarrow Y$	
Covariate	Y1 Slopes	Y2 OR's	Y1 Intercepts	Y2 OR's
X1	0.593*	2.008*	C#1 -1.014*	3.900 (1 vs 3)*
X2	0.553*	1.501*	C#2 -0.164	0.706 (2 vs 3)
X3	0.530*	1.670*	C#3 0.972*	

$X \rightarrow C$

Covariate	C#1 OR's (1 vs 3)	C#2 OR's (2 vs 3)
X1	0.295*	0.620
X2	1.700*	1.328*
X3	0.695*	0.956

LCA: 2-Step with Imputation, Combined Approach, ML

Class probabilities: 0.265, 0.296, 0.439

	$X \rightarrow Y$		$C \rightarrow Y$	
Covariate	Y1 Slopes	Y2 OR's	Y1 Intercepts	Y2 OR's
X1	0.570*	1.826*	C#1 -1.006*	2.948 (1 vs 3)*
X2	0.562*	1.524*	C#2 -0.116	0.669 (2 vs 3)
X3	0.529*	1.640*	C#3 1.002*	

$X \rightarrow C$

Covariate	C#1 OR's (1 vs 3)	C#2 OR's (2 vs 3)
X1	0.290*	0.574*
X2	1.740*	1.351*
X3	0.674*	0.964

LCA Results Conclusions

- Results differ somewhat across multistep approaches even for a model that is (largely) correct
  - Differences are small and do not change significance of estimates
- Results are based on Monte-Carlo generated data from a single replication
  - A literature summary of Monte Carlo studies with many replications is given in the Statistical Methodology section showing that the choice of best multistep approach depends on the circumstances some of which are unknowable
- For any given data set, it may be warranted to study results from all the approaches and see which results do and do not hold up

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
- Results comparing manual approaches
- **Real-data example using antisocial behavior measures**
- Further topics
- Statistical methodology, appendices, references

Real-Data Example: Antisocial Behavior

- ASB data:
  - 17 antisocial behavior items collected in the 1980 National Longitudinal Survey of Youth for respondents between the ages of 16 and 23 together with a set of background variables
  - The ASB items assessed the frequency of various behaviors during the past year, here dichotomized as 0 vs > 0 times
  - A sample of 7,326 respondents has complete data on the antisocial behavior items and the background variables
- ASB analyses:
  - SEM (MIMIC) 4-factor analysis (Muthén, 2025)
  - Latent class analysis from Mplus Short Course Topic 5, slides 91-118
    - 4-class and 5-class LCA of the 17 latent class indicators
    - 5 classes: High, property offense, drug, person offense, normative (low, except for pot)
  - ASB data available for practice

5-Class Results in Probability Scale for Highest Category

- Class probabilities: 0.019, 0.117, 0.172, 0.261, 0.431 (Entropy = 0.741)

Variable	High	Property	Drug	Person	Low
PROPERTY	0.95	**0.65**	0.13	0.23	0.02
FIGHT	0.85	0.65	0.15	**0.50**	0.08
SHOPLIFT	0.99	**0.73**	0.38	0.27	0.06
LT50	0.91	**0.57**	0.24	0.17	0.04
GT50	0.81	0.19	0.02	0.02	0.00
FORCE	0.53	0.15	0.01	0.07	0.00
THREAT	0.91	0.76	0.31	**0.65**	0.08
INJURE	0.61	0.32	0.05	0.16	0.00
POT	0.97	0.86	**0.95**	0.33	0.20
DRUG	0.83	0.48	**0.54**	0.02	0.01
SOLDPOT	0.75	0.38	0.22	0.00	0.00
SOLDDRUG	0.35	0.07	0.04	0.00	0.00
CON	0.81	**0.53**	0.18	0.34	0.07
AUTO	0.64	**0.24**	0.06	0.09	0.01
BLDG	0.78	**0.26**	0.02	0.03	0.00
GOODS	0.90	**0.50**	0.07	0.09	0.00
GAMBLING	0.28	0.09	0.01	0.02	0.00

Input for ASB Analysis with 17 Indicators, 9 Covariates, 2 Distal Outcomes, and 5 Classes: **1-Step Analysis**

```
TITLE:      ASB, 17 indicators, 9 covariates,
            2 distals, 1-step analysis

DATA:       FILE = asbfree.dat;

VARIABLE:   NAMES = property fight shoplift lt50
            gt50 force threat injure pot drug
            soldpot solddrug con auto bldg goods
            gambling dsm1-dsm22 male black hisp
            single divorce dropout college onset f1
            f2 f3 age94 cohort dep abuse;

            USEVARIABLES =
            ! 17 latent class indicators:
            property-gambling
            ! 9 covariates:
            male black hisp single divorce dropout
            college onset age94
            ! 2 distal outcomes
            dep abuse;
            CATEGORICAL = property-gambling;

            CLASSES = c(5);

ANALYSIS:   TYPE = MIXTURE;
            ESTIMATOR = ML;
            STARTS = 800 400;
            PROCESSORS = 12;

MODEL:      %OVERALL%

            c ON male-age94;
            dep abuse ON male-age94;
            dep WITH abuse;

            ! property-gambling thresholds
            ! and dep abuse means
            ! vary across classes by default
```

1-Step vs Multistep Analysis

- 1-step fails: Best log likelihood cannot be replicated even at STARTS = 1200 400
  - Likely cause: Data on the distal outcomes confound the 5-class solution based on the latent class indicators
- This makes multistep approaches not only preferrable but necessary:
  - 3-step, BCH, 2-step
  - BCH demonstrated here

Input for ASB Analysis with 17 Indicators, 9 Covariates, 2 Distal Outcomes, and 5 Classes: First Step Combined

```
TITLE:      ASB, 17 indicators, combined first step.
            Entropy = 0.741

DATA:       FILE = asbfree.dat;

VARIABLE:   NAMES = property fight shoplift lt50
            gt50 force threat injure pot drug
            soldpot solddrug con auto bldg goods
            gambling dsm1-dsm22 male black hisp
            single divorce dropout college onset f1
            f2 f3 age94 cohort dep abuse;

            USEVARIABLES =
            property-gambling;
            ! male black hisp single divorce dropout
            ! college onset age94 dep abuse;

            AUXILIARY = male black hisp single
            divorce dropout college onset age94
            dep abuse;

            CATEGORICAL = property-gambling;

            CLASSES = c(5);

ANALYSIS:   TYPE = MIXTURE;
            ESTIMATOR = ML;
            STARTS = 800 400;
            PROCESSORS = 12;

SAVEDATA:   FILE = 17-9-2.dat;
            SAVE = CPROB BCHWEIGHTS;
OUTPUT:     SVALUES;
```

Input for ASB Analysis with 17 Indicators, 9 Covariates, 2 Distal Outcomes, and 5 Classes: Last Step, BCH

TITLE: ASB, 17 indicators, 9 covariates, 2 distals, combined last step, BCH

DATA: **FILE = 17-9-2.dat;**

VARIABLE: NAMES = property fight shoplift lt50 gt50 force threat injure pot drug soldpot solddrug con auto bldg goods gambling male black hisp single divorce dropout college onset age94 dep abuse w1-w5 cprob1-cprob5 n;

USEVARIABLES = male black hisp single divorce dropout college onset age94 dep abuse w1-w5;

TRAINING = w1-w5(BCH);

CLASSES = c(5);

ANALYSIS: TYPE = MIXTURE;
ESTIMATOR = ML;
STARTS = 0;
PROCESSORS = 12;

c ON male-age94;
dep abuse ON male-age94;
dep WITH abuse;

ASB Analysis Summary

- 1-step: Failure, best log likelihood cannot be replicated
- First step (measurement model - latent class indicators only):
  - Class probabilities: 0.019, 0.117, 0.172, 0.261, 0.431
    - Agreement with Short Course results
    - High, property offense, drug, person offense, normative (low)
- 3-Step, last step (adding covariates and distal outcomes):
  - Class probabilities: 0.021, 0.094, 0.737, 0.138, 0.009
    - Failure to match the measurement model probabilities
- BCH, last step (adding covariates and distal outcomes):
  - Class probabilities: 0.019, 0.118, 0.172, 0.260, 0.431
    - Agreement with measurement model
- 2-Step, last step (adding covariates and distal outcomes):
  - Class probabilities: 0.014, 0.116, 0.214, 0.648, 0.008
    - Failure to match the measurement model probabilities
    - STARTS > 0 gives a better and replicated logL with class prob's that also don't match well

ASB Conclusions

- The superiority of BCH is in line with simulation findings (see overview of Monte Carlo studies)
  - Class-specific histograms show strong within-class non-normality for the distal outcomes which violates the assumption of within-class normality, causing bias for 3-step and 2-step
- Distal outcomes with strong floor effects were treated as continuous with class-invariant variances
  - Would the comparison of methods turn out differently if the variances were allowed to vary across classes?
  - Would the comparison of methods turn out differently if the distal outcomes were dichotomized and treated as binary?
  - Should the distal outcomes be treated as censored, two-part?

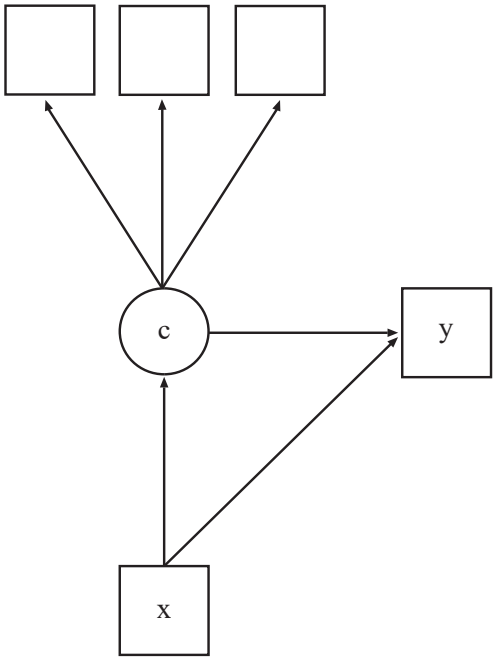
ASB Conclusions, Continued

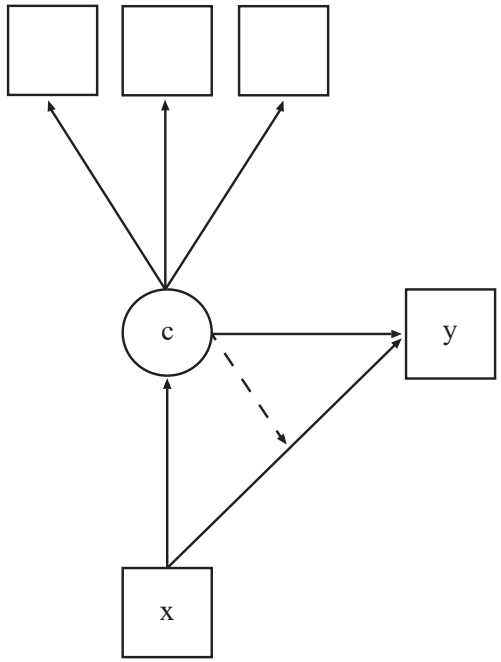
- ASB is a general population survey so that considerable heterogeneity among individuals can be expected
  - Measurement non-invariance (DIF) is likely for some latent class indicators as a function of some covariates - direct effects from covariates to indicators
  - Ignoring direct effects causes bias, especially for C ON X
  - Searching for, accounting for, and evaluating direct effects is treated in the web talk Using Mplus to Investigate Direct Effects in Latent Class Analysis

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- **Further topics**
  - Class-varying regression
  - Last step considerations
  - CPROBABILITIES with multiple imputation
- Statistical methodology, appendices, references

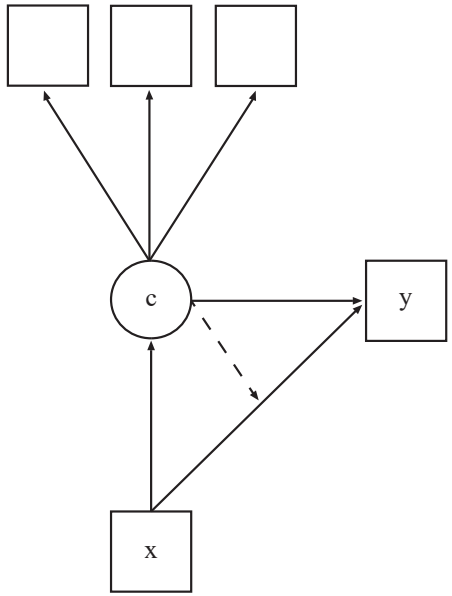
Class-Invariant vs Class-Varying Regression





- The model on the left has two main effects:
  - $x \rightarrow y$: y on x regression coefficient
  - $c \rightarrow y$: class-varying y intercepts/thresholds
- The model on the right has an interaction
  - The broken arrow from c to the arrow from x to y indicates a class-varying coefficient in the regression of y on x

Class-Varying Regression Representing an Interaction



- The y on x coefficient that varies as a function of the c variable implies an interaction between x and c in their influence on y
- With an interaction, the influence of c on y is no longer a main effect that can be represented by class-varying y intercepts/thresholds (compare ANCOVA)
- See the Statistical Methodology section for details

Class-Invariant vs Class-Varying Y ON X Input Excerpts for Three Classes

- Latent class indicator means/thresholds vary across classes by default

Class-invariant case:

```
%OVERALL%
y ON x;
```

Class-varying case:

```
%OVERALL%
y ON x;
%C#1%
y ON x;
%C#2%
y ON x;
%C#3%
y ON x;
```

- Class-varying regression can cause non-convergence unless the sample size is large and the class sizes are not small
- How do you test whether class-varying regression is needed?

Testing Class-Varying Regression: Omnibus Test and Pairwise Tests for Three Classes

```
MODEL:      %OVERALL%
            y ON x;

            %c#1%
            y ON x (p1);
            %c#2%
            y ON x (p2);
            %c#3%
            y ON x (p3);
```

```
MODEL TEST:         ! Omnibus test
                    0 = p2-p1;
                    0 = p3-p1;

MODEL CONSTRAINT:   ! Pairwise tests
                    NEW(diff21 diff31 diff32);
                    diff21 = p2-p1;
                    diff31 = p3-p1;
                    diff32 = p3-p2;
```

- In addition, BIC can be compared for the class-invariant and class-varying models
  - Number of parameters and BIC for the ASB example:
    - Class-invariant Y ON X: 71, **60,142**
    - Class-varying Y ON X: 143, 60,450

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
  - Class-varying regression
  - **Last step considerations**
  - CPROBABILITIES with multiple imputation
- Statistical methodology, appendices, references

Last Step Considerations: STARTS = 0

- The latent classes are determined in the first step
- STARTS = 0 is used in the last step because the last step does not involve a mixture analysis needing random starting values to find the optimal maximum of the log likelihood
- If STARTS $>$ 0 in the last step leads to a better log likelihood than STARTS = 0, this can mean that
  - Multistep fails:
    - The solution gives class probabilities in worse agreement with those of the first step than STARTS = 0
    - The auxiliary distal outcome model is mis-specified, such as needing its own latent class variable
  - Multistep needs STARTS $>$ 0
    - The solution gives better class probability agreement with those of the first step than STARTS = 0 because the automatic start values of STARTS = 0 are not good enough for the complexity of the auxiliary model

Last Step Considerations: Within-Class Correlation and Missing Data for the Distal Outcomes

- The distal outcomes correlate due to being influenced by the latent class variable
- The addition of within-class correlation using WITH for categorical distals is not available using ML but can be modeled using Bayes
  - Adding within-class correlation for distal outcomes:
    - May improve the estimation with respect to missing data on the distal outcomes (FIML)
    - May obtain significant correlations due to omitted covariates
    - May lead to worse class probability agreement with first step
- An alternative to using within-class correlations to improve missing data handling is to use multiple imputation of the distal outcomes
  - Less reason to use within-class WITH

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
  - Class-varying regression
  - Last step considerations
  - **CPROBABILITIES with multiple imputation**
- Statistical methodology, appendices, references

CPROBABILITIES

- An individual's estimated probability for each latent class
  - Example: 6 individuals, 3 latent classes:

	CPROB1	CPROB2	CPROB3
i = 1	**0.5**	0.3	0.2
i = 2	0.1	0.1	**0.8**
i = 3	0.1	**0.6**	0.3
i = 4	0.3	**0.4**	0.3
i = 5	0.3	**0.7**	0.0
i = 6	**0.5**	0.2	0.3

- Computed using two sources:
  - The individual's observed value for each variable
  - The estimated parameters of the model (based on all individuals)

CPROBABILITIES from Final Analysis for 2-Step After Multiple Imputation

TITLE: Final step for 2-step based on Combined runs
Imputation + Analysis in one step

DATA: FILE = stageA.dat;

VARIABLE: NAMES = u1-u6 x1-x3 y1-y3 w1-w3 cp1-cp3 n;
USEVARIABLES = u1-u6 x1-x3 y1-y3;
MISSING = *;
CATEGORICAL = u1-u6 y2 y3;
CLASSES = c(3);

DATA IMPUTATION: **NAMES = u1-u6(C) x1(C) x2 x3 y1 y2(C) y3(C);**
IMPUTE = x1(C) x2;
! the imputation uses information from the variables
! on the DATA IMPUTATION NAMES list and
! imputes values for variables on the IMPUTE list
NDATASETS = 100;
THIN = 100;
SAVE = imputed*.dat;

Input continues on the next slide

CPROB's from Final Analysis for 2-Step Continued

```
ANALYSIS:    TYPE = MIXTURE;
             ESTIMATOR = MLR;
             STARTS = 0;
             ! Can add BITER for the imputation

MODEL:       %OVERALL%
             c ON x1-x3;
             y1-y3 ON x1-x3;
             ! Note: No x part because
             ! the x's have no missing data

             %C#1%

             [ u1$1*-1.28835 ];
             [ u2$1*-1.41536 ];
             [ u3$1*-1.74521 ];
             [ u4$1*-1.00460 ];
             [ u5$1*-1.36393 ];
             [ u6$1*-1.17851 ];

             %C#2%

             [ u1$1*1.72069 ];
             [ u2$1*1.73016 ];
             [ u3$1*1.54348 ];
             [ u4$1*-1.68227 ];
             [ u5$1*-1.37222 ];
             [ u6$1*-1.82568 ];

             %C#3%

             [ u1$1*1.52507 ];
             [ u2$1*1.28543 ];
             [ u3$1*1.52212 ];
             [ u4$1*1.30734 ];
             [ u5$1*1.42798 ];
             [ u6$1*1.29213 ];

SAVEDATA:    FILE = impcprob.dat;
             SAVE = CPROB;
```

CPROBABILITIES with Multiple Imputation

CPROB's for 2-step in the final analysis

	CPROB's	MLC
i=1, imputation 1:	0.006 0.631 0.364	2
imputation 2:	0.006 0.619 0.374	2
imputation 3:	0.006 0.606 0.388	2
i=1, average:	0.006 0.618 0.376	2

- Imputation results obtained from SAVE = imputed*.dat
- Average results obtained from SAVEDATA SAVE = CPROB

Comparison: CPROB's from Measurement Model (stageA.dat)

	CPROB's	MLC
i=1	0.009 0.673 0.317	2

CPROBABILITIES: Choosing Between 3-Step, BCH, and 2-Step

- 3-step: Not recommended
  - Uses information from the measurement model in the form of the nominal most likely class variable with fixed logits
  - Not all of the information from the individual latent class indicators is captured by most likely class
- BCH: Mplus does not provide CPROBABILITIES
  - Uses information from the measurement model in the form of the BCH weights - not straightforward to get to the CPROB's
  - Not all of the information from the individual latent class indicators is captured by the BCH weights
- 2-step: Recommended
  - In addition to the y-x information in the final analysis step, 2-step uses information from the measurement model in the form of the fixed measurement parameters for the latent class indicators

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- **Statistical methodology, appendices, references**
  - **Posterior probabilities and MLC for 3-step and BCH**
  - Missing data on a binary covariate
  - Why class-varying regression is an interaction
  - Overview of Monte Carlo studies for 3-step, BCH, and 2-step
  - Appendix: Generating missing data on covariates

Posterior Probabilities and Most Likely Class (MLC)

- Posterior probabilities are computed for each individual and latent class
- The CPROB option of the SAVEDATA command saves them
- Example: 6 individuals, 3 latent classes (MLC bolded):

	1	2	3
i = 1	**0.5**	0.3	0.2
i = 2	0.1	0.1	**0.8**
i = 3	0.1	**0.6**	0.3
i = 4	0.3	**0.4**	0.3
i = 5	0.3	**0.7**	0.0
i = 6	**0.5**	0.2	0.3

- A summary is given in a table with average class probabilities for MLC (row) by class (column): P(C = c | N = n) where N is the MLC vble
- Using Bayes Theorem, this produces P(N = n | C = c) which is used by both 3-step and BCH capturing the measurement error in MLC

Computations Based on Posterior Probabilities

	Posterior prob's		
	Class		
	1	2	3
i = 1	**0.5**	0.3	0.2
i = 2	0.1	0.1	**0.8**
i = 3	0.1	**0.6**	0.3
i = 4	0.3	**0.4**	0.3
i = 5	0.3	**0.7**	0.0
i = 6	**0.5**	0.2	0.3
P(C):	0.30	0.38	0.32

P(C = c | N = n)	
N C	
n=1, c=1:	(0.5+0.5)/2
n=1, c=2:	(0.3+0.2)/2
n=1, c=3:	(0.2+0.3)/2
n=2, c=1:	(0.1+0.3+0.3)/3
n=2, c=2:	(0.6+0.4+0.7)/3
n=2, c=3:	(0.3+0.3+0)/3
n=3, c=1:	0.1/1
n=3, c=2:	0.1/1
n=3, c=3:	0.8/1

		P(C = c | N = n)		
		c		
		1	2	3
n	1	0.50	0.25	0.25
	2	0.23	0.57	0.20
	3	0.10	0.10	0.80

P(N=n | C=c) Obtained via Bayes Theorem

P(N=n | C=c) = P(C=c | N=n)*P(N=n)/P(C=c)
where P(C=c | N=n) was obtained on the previous slide

	P(N = n)		P(C = c)
n=1:	2/6	c=1:	0.30
n=2:	3/6	c=2:	0.38
n=3:	1/6	c=3:	0.32

P(N = n | C = c)

		n		
		1	2	3
	1	0.50*(2/6)/0.30	0.23*(3/6)/0.30	0.10*(1/6)/0.30
c	2	0.25*(2/6)/0.38	0.57*(3/6)/0.38	0.10*(1/6)/0.38
	3	0.25*(2/6)/0.32	0.20*(3/6)/0.32	0.80*(1/6)/0.32

Measurement Error in MLC

$P(N = n \mid C = c)$

	N=1	N=2	N=3
C=1:	**0.56**	0.38	0.06
C=2:	0.22	**0.75**	0.04
C=3:	0.26	0.31	**0.42**

- The output refers to this table as Classification Probabilities for the Most Likely Latent Class Membership (Column) by Latent Class (Row)
- The P(N | C) table represents a multinomial regression of an observed nominal indicator N on a latent variable C (cf factor analysis)
- The P(N | C) table shows that MLC is not a perfect indicator of latent class membership but can be seen as having measurement error

Taking MLC Measurement Error Into Account Using Logits for the Last Step of 3-Step

- The measurement error is translated into logits for each latent class and category of the nominal variable

P(N | C)

	N=1	N=2	N=3
C=1:	**0.56**	0.38	0.06
C=2:	0.22	**0.75**	0.04
C=3:	0.26	0.31	**0.42**

log[P(N=n | C=c)/ P(N=Last | C=c)]

	N=1	N=2	N=3
C=1:	**2.23**	1.85	0
C=2:	-1.23	**2.93**	0
C=3:	-0.48	-0.30	**0**

- Correction for MLC measurement error in the last step analysis:
  - The nominal N variable is added to the analysis variables and given fixed logits

Taking MLC Measurement Error Into Account Using Weights for BCH

- Vermunt (2010) considers BCH for LCA where the full model adds covariates x
- Consider the elements of matrices E, A, and D
  - $E : e_{in} = P(N = n|x_i)$, the biased MLC relationship with x
  - $A : a_{ic} = P(C = c|x_i)$, the true (unbiased) relationship with x
  - $D : d_{cn} = P(N = n|C = c)$, the MLC measurement error
- Note that $P(N = n|x_i) = \sum_c P(C = c|x_i)\ P(N = n|C = c)$
  - The summation over the classes corresponds to the matrix product $A \times D$, so that $E = A \times D$ (the dimension of D is $C \times C$)
- Because $E = A \times D$, the true (unbiased) relationship A is obtained as $A = E \times D^{-1}$, that is, correcting the biased relationship of E using the inverse of D as weights in the E regression of N on x

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references
  - Posterior probabilities and MLC for 3-step and BCH
  - **Missing data on a binary covariate**
  - Why class-varying regression is an interaction
  - Overview of Monte Carlo studies for 3-step, BCH, and 2-step
  - Appendix: Generating missing data on covariates

Missing Data on a Binary Covariate Using ML

- Consider the regression of Y on X1 and X2 where X1 is binary with missing data and the Xs are brought into the model using WITH:
  - Y ON X1 X2;
  - X1 WITH X2;
- The Xs are now treated as DVs, ML is used, and X1 is not on the CATEGORICAL list because ML cannot handle WITH involving categorical variables
- EM algorithm:
  - The E step does not recognize that X1 is binary but uses a continuous-variable normality assumption for it so that expectations for continuous, not binary variables are produced
  - The M step regresses Y on such continuous X1 values (except for the non-missing X1 observations which are kept as binary)
  - The missing data literature indicates that this approximation is sufficient to not produce an important bias in the regression
    - Schafer (1997), Demirtas et al., (2008), Muthén et al. (2016)

Missing Data on a Binary Covariate. The Bayes Alternative

- Chapter 10 of the Muthén et al. (2016) book shows that Bayes can handle this case better because it allows X1 to be put on the CATEGORICAL list and allows using WITH
  - Simulations show that this gives slightly better results when you have a lot of missing data and when the missing data is NMAR
- This Bayes approach is, however, not yet available for mixture models in Mplus because it requires C ON X1 with missing X1

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references
  - Posterior probabilities and MLC for 3-step and BCH
  - Missing data on a binary covariate
  - **Why class-varying regression is an interaction**
  - Overview of Monte Carlo studies for 3-step, BCH, and 2-step
  - Appendix: Generating missing data on covariates

Why is a Class-Varying Regression an Interaction? A Reminder of Interaction with Linear Regression

- Interaction with linear regression for observed covariates x and c:

$$\begin{aligned} y &= a + b_1 * x + b_2 * c + b_3 x * c + e, \\ &= a + (b_1 + b_3 * c) * x + b_2 * c + e, \end{aligned}$$

  where b_1 and b_2 are the main effects of x and c, respectively, and b_3 is the interaction effect ($b_1 + b_3 * c$ is called a moderator function, or evaluated at a specific c value, a simple slope)
  - We will focus on the model's 4 parameters: a, b_1, b_2, b_3
- How does this translate to the mixture case?

Why is a Class-Varying Regression an Interaction?

- Interaction with linear regression where c is observed:

$$y = a + b_1 * x + b_2 * c + b_3 x * c + e, \quad (1)$$
$$= a + (b_1 + b_3 * c) * x + b_2 * c + e, \quad (2)$$

- Interaction with mixtures where c is a latent class variable
  - For example, two classes scored 0/1 (like a dummy vble; with more classes, there are more dummy vbles):
    - In line with (2), b_1 is the effect of x on y for the first class (c = 0) and $b_1 + b_3$ the effect of x on y for the second class (c = 1)
    - The interaction is the difference in effects, that is, b_3
- Mixture notation: $y = a_c + b_c * x + e$
  - This has 4 parameters as with linear regression, 2 a's, 2 b's
  - No interaction case: $y = a_c + b * x + e$. The class-invariant b allows the effect of c on y to be summarized by the intercepts a_c
  - With interaction, that is, b_c, the effect of c on y can no longer be summarized by the intercepts a_c (compare ANCOVA)

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references
  - Posterior probabilities and MLC for 3-step and BCH
  - Missing data on a binary covariate
  - Why class-varying regression is an interaction
  - **Overview of Monte Carlo studies for 3-step, BCH, and 2-step**
  - Appendix: Generating missing data on covariates

Brief Overview of Monte Carlo Studies of 2-Step, 3-Step, and BCH

- C ON X:
  - Vermunt (2010):
    - 3-step $\approx$ BCH in terms of bias
    - 3-step has less variability than BCH
  - Bakk-Kuha (2018):
    - 2-step slightly better than 3-step and $\approx$ BCH in terms of bias
    - 2-step has less variability than BCH
  - Asparouhov-Muthén (Web Note 21):
    - 2-step better than BCH for RI-LTA

Brief Overview of Monte Carlo Studies, Continued

- Y ON X:
  - Bakk-Kuha (2018):
    - 2-step ≈ 3-step ≈ BCH except for very low entropy
    - 2-step and 3-step have slightly less variability than BCH
- Y ON X with unequal Y variances across classes
  - Bakk-Vermunt (2016), Asparouhov-Muthén (Web Note 21):
    - BCH has less bias than 3-step
    - 3-step allowing unequal variances best
- Y ON X with non-normal Y within class:
  - Bakk-Vermunt (2016): BCH has less bias than 3-step
  - Bakk-Kuha (2018):
    - BCH best
    - 2-step ≈ 3-step
  - Asparouhov-Muthén (Web Note 21):
    - BCH best
    - 2-step bad, 3-step very bad

Brief Overview of Monte Carlo Studies, Continued

- C ON X with missing data on binary X
  - Depaoli, Jia, Visser (2025):
    - Only 3-step investigated
    - MCAR and MAR missingness as a function of a U indicator
    - FIML $\approx$ Bayes $\approx$ MI
    - No discernable bias of FIML or Bayes for C ON binary X with missing data

- FIML:
  - Using WITH among Xs, ignoring that an X is binary
- Bayes:
  - Using WITH among Xs, ignoring that an X is binary
  - Mplus default of diffuse priors. Informative priors reduce MSE
- MI:
  - Multiple imputation of Xs using X and U information before last step
  - Some bias with a high degree of missingness and strong effect of X on C

Outline

- Background
- Measurement model analysis
- Automatic multistep approaches
- Manual multistep approaches
- Results comparing manual approaches
- Real-data example using antisocial behavior measures
- Further topics
- Statistical methodology, appendices, references
  - Posterior probabilities and MLC for 3-step and BCH
  - Missing data on a binary covariate
  - Why class-varying regression is an interaction
  - Overview of Monte Carlo studies for 3-step, BCH, and 2-step
  - **Appendix: Generating missing data on covariates**

Appendix: Generating Missing Data on Covariates

The Monte Carlo generation of data uses the logistic regression specification for the probability of missing on x1, x2 (full input on next slide):

```
MODEL MISSING:    %OVERALL%
                  [x1*-1.5 x2*-1.5];
                  x1 x2 ON x3*0.4;
```

- Missingness on x1, x2 as a function of x3:
  - MAR missingness as opposed to MCAR
- At the mean zero for x3, the logit intercept of -1.5 translates to a probability of missing on x1, x2 of $1/(1+e^{1.5}) = 0.18$
- One standard deviation above the mean of x3, the logit of -1.5+0.4 translates to a probability of missing = 0.25
- Deleting individuals with missing on one or more of the x1, x2 covariates: Sample size for the generated data drops from 1000 to 668, a 33% loss

Appendix: Input for Monte Carlo Data Generation of LCA

```
TITLE:      Missing on X and Distal:
            Entropy = 0.740.

            x1: binary, missing
            x2: cont's, missing
            x3: cont's, not missing
            y1: cont's, missing
            y2 binary, missing
            y3 ordinal, not missing

MONTECARLO: NAMES = u1-u6 x1 x2 x3 y1 y2 y3;
            NOBSERVATIONS = 1000;
            NREP = 1;
            SAVE = distal.dat;
            CLASSES = c(3);
            GENCLASSES = c(3);
            GENERATE = u1-u6(1) x1(1)
            y2(1) y3(2);
            CATEGORICAL = u1-u6 x1 y2 y3;
            MISSING = x1 x2 y1 y2;

            ! seed = 347;

MODEL MISSING: %OVERALL%
               [x1*-1.5 x2*-1.5];
               x1 x2 ON x3*0.4;
               [y1*-1.5 y2*-1.5];

ANALYSIS:   TYPE = MIXTURE;
            ALGORITHM =
            INTEGRATION;
            INTEGRATION =
            MONTECARLO;

MODEL
POPULATION: %OVERALL%
            [x1$1*0]; x1 ON x3*1; x3*1;
            [x2*0]; x2 ON x3*1; x2*1;
            c#1 ON x1*0.5 x2*-0.3 x3*0.2;
            c#2 ON x1*-.5 x2*.3 x3*-.2;

            y1-y3 ON x1-x3*0.5;
            y1*1;
```

Input for Monte Carlo Data Generation, Continued

```
%C#1%
[u1$1-u6$1*1.386];
[y1*1 y2$1*0.5 y3$1*0.2 y3$2*0.4];

%C#2%
[u1$1-u6$1*-1.386];
[y1*-1 y2$1*-0.5 y3$1*-0.2 y3$2*0];


%C#3%
[u1$1-u3$1*1.386];
[u4$1-u6$1*-1.386];
[y1*0 y2$1*0.5 y3$1*-0.2 y3$2*0.2];

MODEL:  %OVERALL%
        [x1$1*0]; x1 ON x3*1; x3*1;
        [x2*0]; x2 ON x3*1; x2*1;
        c#1 ON x1*0.5 x2*-0.3 x3*0.2;
        c#2 ON x1*-.5 x2*.3 x3*-.2;
```

```
y1-y3 ON x1-x3*0.5;
y1*1;

%C#1%
[u1$1-u6$1*1.386];
[y1*1 y2$1*0.5 y3$1*0.2 y3$2*0.4];

%C#2%
[u1$1-u6$1*-1.386];
[y1*-1 y2$1*-0.5 y3$1*-0.2 y3$2*0];

%C#3%
[u1$1-u3$1*1.386];
[u4$1-u6$1*-1.386];
[y1*0 y2$1*0.5 y3$1*-0.2 y3$2*0.2];
```

Analysis of the Correct Full LCA Model Knowing the Missing Data Mechanism

```
TITLE:      Correct full model specified as the data were generated

DATA:       FILE = distal.dat;

VARIABLE:   NAMES = u1-u6 x1 y2 y3 x2 y1 x3 class;
            USEVARIABLES = u1-u6 x1 x2 x3 y1 y2 y3;
            CATEGORICAL = u1-u6 x1 y2 y3;
            MISSING = ALL(999);
            CLASSES = c(3);

ANALYSIS:   TYPE = MIXTURE;
            STARTS = 80 20;
            ALGORITHM = INTEGRATION;
            INTEGRATION = MONTECARLO;
            PROC = 8;
MODEL:      %OVERALL%
            ! x1-x3 with x1-x3; not allowed with ML due to categ. x1
            ! Instead, model it the way the missingness was generated,
            ! that is, as a function of x3 (MAR):
            x1 x2 ON x3;
            c ON x1-x3;
            y1-y3 ON x1-x3;
```

References

- Asparouhov & Muthén (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. Structural Equation Modeling: A Multidisciplinary Journal, 21:3, 329-341. The posted version corrects several typos in the published version. An earlier version of this paper was posted as **Mplus Web Notes: No. 15**.
- Asparouhov & Muthén (2021). Auxiliary variables in mixture modeling: Using the BCH method in Mplus to estimate a distal outcome model and an arbitrary secondary model. **Mplus Web Notes: No 21**.
- Bakk, Tekle, & Vermunt (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. Sociological Methodology, 43, 272-311.
- Bakk & Kuha (2018). Two-step estimation of models between latent classes and external variables. Psychometrika, 83, 871-892.
- Bakk & Vermunt (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. Structural Equation Modeling, 23, 20-31.

References, Continued

- Bolck, Croon, & Hagenaars (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. Political Analysis, 12, 3-27.
- Depaoli, Jia & Visser (2025). Addressing missing data in latent class analysis when using a three-step estimation approach. Structural Equation Modeling, 32, 287-303.
- Muthén, B. (2025). Mplus: An overview of its unique analysis capabilities. Forthcoming in the Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences: Volume Three.
- Muthén, Muthén & Asparouhov (2016). Regression and Mediation Analysis Using Mplus. `https://www.statmodel.com/Mplus_Book.shtml`
- Vermunt (2010). Latent class modeling with covariates: Two improved three-step approaches. Political Analysis, 18, 450-469.