IMMERSE Training

June 5-8, 2003

University of California, Santa Barbara

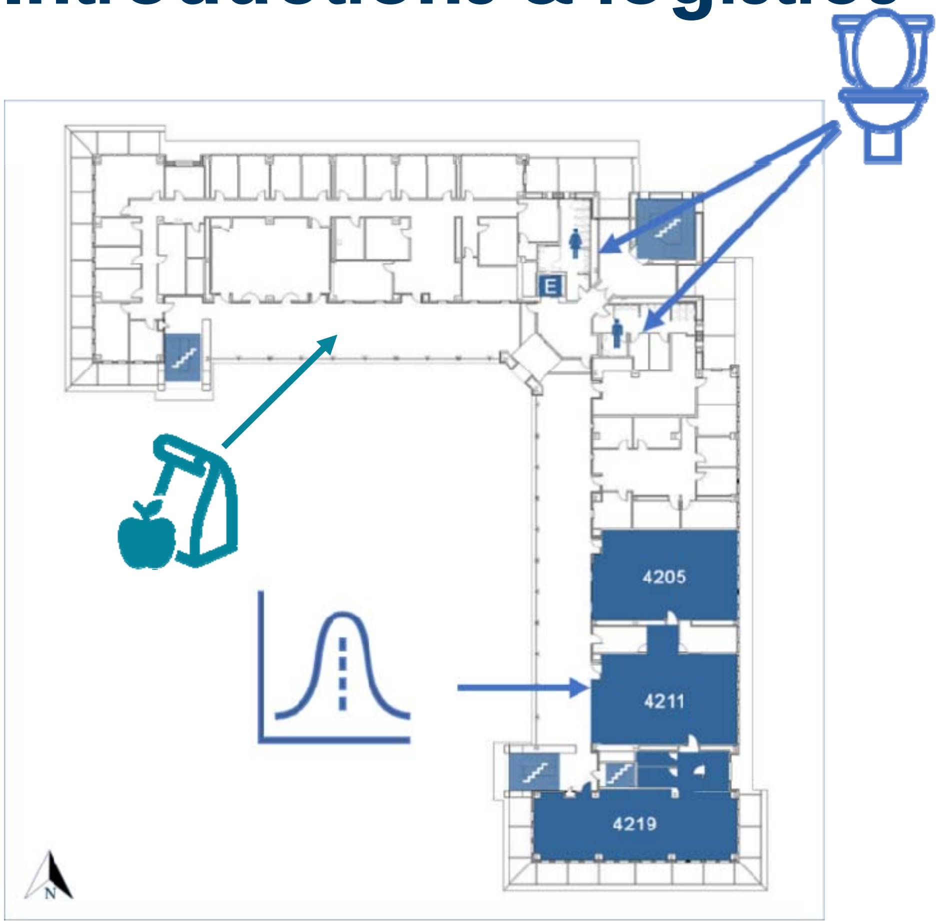
IMMERSE Project – IES funded Training Grant (R305B220021)

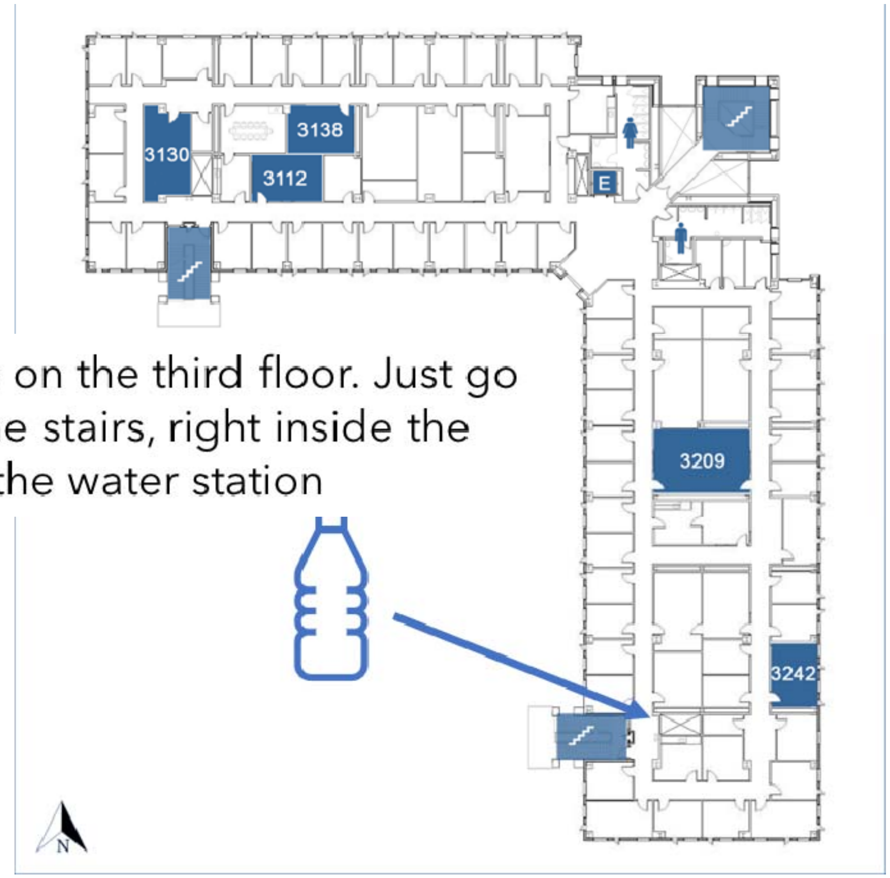
UC SANTA BARBARA

Introductions

Housekeeping

Introductions & logistics





Water is on the third floor. Just go down the stairs, right inside the door is the water station

Tentative Daily Schedule

9:00: Start

9:00-12:30: Class/Lab/Break(s)
12:30-1:30: Lunch
1:30-5:00: Class/Lab/Break(s)/Consultations
5:00: End

IMMERSE Project – IES funded Training Grant (R305B220021)

Housekeeping

- Please don't share our slide deck without permission
- Feel free to interrupt and ask questions
  - Parking lot
- Snack Hall
- Room door locks (go two doors down to ITG to get let in)
- Optional Activities
  - Yoga/order food (meet at 5:45 in lobby, walk from there)
  - Funk Zone
  - Lunch on Wednesday

Thank you IES!

IMMERSE Project – IES funded Training Grant (R305B220021)

Mixture Modeling and Latent Class Analysis: Day 1

Please do not copy or distribute without permission.

IMMERSE Project – IES funded Training Grant (R305B220021)

UC SANTA BARBARA

Primary Goal for the Next Three Days...

- Lay the foundation for principled applications of mixture models in your own research through:
  - A conceptual deep-dive into latent class analysis (because much of the modeling process generalizes to more complex finite mixture models, including longitudinal mixture models).
  - Collective, active learning to construct your own understanding of mixture modeling best-practices and pitfalls.
  - Guided lab activities to build coding and trouble-shooting skills with model specification, estimation, and evaluation.

Remember: This is a year-long training program. We know from experience that rushing through the fundamental content in order the get to more advanced mixture models in these three days will leave gaps and cracks in the foundation that will compromise achieving your ultimate training objectives.

IMMERSE Project – IES funded Training Grant (R305B220021)

Today's Agenda

- Pre-Training Day 1 Recap
- LCA "Big Picture" Example
- The Latent Class Measurement Model
- Evaluating Measurement Model Quality
- Sidebar: Categorical Variables in Mplus
- LCA Example in Mplus
- Mixture Model Estimation
- Latent Class Enumeration
  - Tests of model fit
  - Model fit indices
  - Classification precision
  - Model usefulness

IMMERSE Project – IES funded Training Grant (R305B220021)

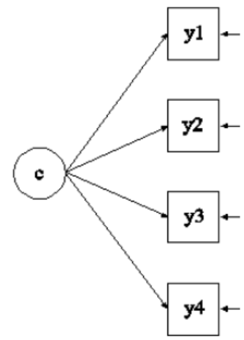
Pre-Training Day 1 Recap

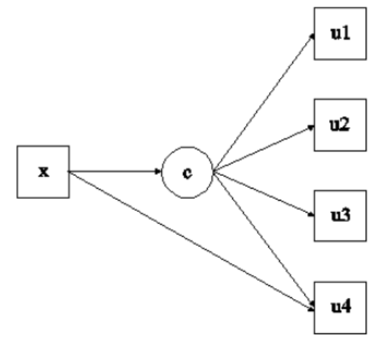
UC SANTA BARBARA

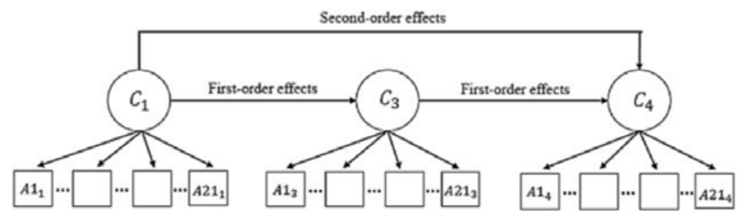
Finite Mixture Modeling Family

Some members include…

- Cross-sectional:
  - **Latent class analysis (LCA)**
  - **Latent profile analysis (LPA)**
  - Regression mixture models
  - Factor mixture models (FMM)
  - And more….
- Longitudinal:
  - Growth mixture models (GMM)
  - **Latent transition models (LTA)**
  - Survival mixture analysis (SMA)
  - And more…





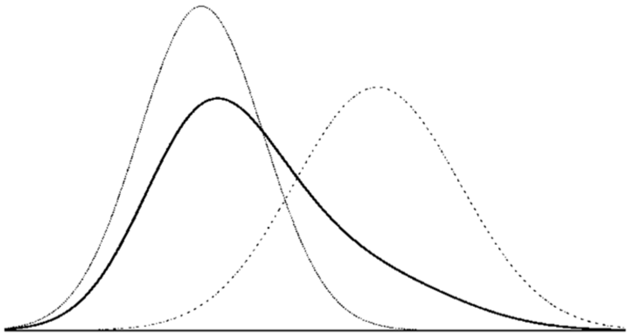


IMMERSE Project – IES funded Training Grant (R305B220021)

	Continuous latent variables	Categorical latent variables	Hybrids
Cross-sectional models	Factor analysis Structural equation models IRT	Latent Class Analysis Latent Profile Analysis Regression Mixture Models	Factor mixture analysis
Longitudinal models	Growth analysis (random effects)	Latent Transition Analysis Latent Class Growth Analysis	Growth mixture analysis

Promise of (Latent Variable) Mixture Modeling

- Free from standard assumption of population homogeneity
- Allows for unobserved population heterogeneity
- Free from antiquated/arbitrary/non-validated grouping/categorizing (typically based on single measures)
- Models that embrace the multidimensional and intersectional
- Person-centered approach (in comparison to variable-centered approaches)



IMMERSE Project – IES funded Training Grant (R305B220021)

Typical Research Questions

- RQ1: How many latent groups of students learners are there in Kindergarten? Does parent SES and preschool attendance predict who will be in each group?

- RQ2: Are there different patterns of student motivation?

- RQ3: What do the four classes of teacher attitudes towards technology look like?

Mixture models: Lauded by some

- Theoretical models that conceptualize individual differences at the latent level as differences in *kind*, that consider typologies or taxonomies, map directly onto analytic latent class models.
- Mixture models give us a great deal of flexibility in terms of how we characterize population heterogeneity and individual differences with respect to a latent phenomenon.
- Can help avoid serious distortions that can results from ignoring population heterogeneity if it is, indeed, present.

IMMERSE Project – IES funded Training Grant (R305B220021)

Mixture models: Impugned by others

- Latent classes or mixtures may not reflect the **Truth**.
- The empirically extracted latent classes depend upon the within- and between-class model specification and the joint distribution of the indicators. Thus, the resultant classes may diverge markedly from the underlying "True" latent structure in the population.
- The current criticisms and skepticisms of mixture modeling are nearly identical to those of path analysis and SEM in the second half of the 20th century because some of the same bad modeling practices have reappeared:
  - "Nobody pays much attention to the assumptions, and the technology tends to overwhelm common sense." (Friedman, 1987)

Don’t cut off your latent classes to spite your data

- Any model is, at best, an approximation to reality.
- “All models are wrong, but some are useful”. (George Box)
- We *can* evaluate model-theory consistency.
- We *can* evaluate model-data consistency.
- There are many alternative ways of thinking about(and parameterizing) **relationships in a variable system** and if mixture modeling can be useful in empirically distinguishing between or among alternative perspectives, then they provide important information.

IMMERSE Project – IES funded Training Grant (R305B220021)

Remember...

with mixture modeling comes great responsibility.

- All modeling requires careful and thoughtful attention to details of the data, specification, and results.
- Mixture modeling requires ***more*** than most!
  - Unstable model estimation require close attention
  - Solutions are error prone
  - Best practices are still being established

Invest in your modeling process

- Set yourself up for success using well-articulated and theory-informed research goals
- Make sure your analytic approach is right for your goals and for your data!
- MplusAutomation will make modeling efficient and replicable
- Look at your output files
- Look for warnings and errors
- Slow down and be thoughtful
- Allow time for interpretation and meaning-making

Applied Latent Class Analysis Example

The “Big Picture”

Mixture Model Building Steps

1. Data screening and descriptives. — Pre-Training
2. Class enumeration process (without covariates—except as auxiliary option).
3. Select final unconditional model (this is your measurement model).

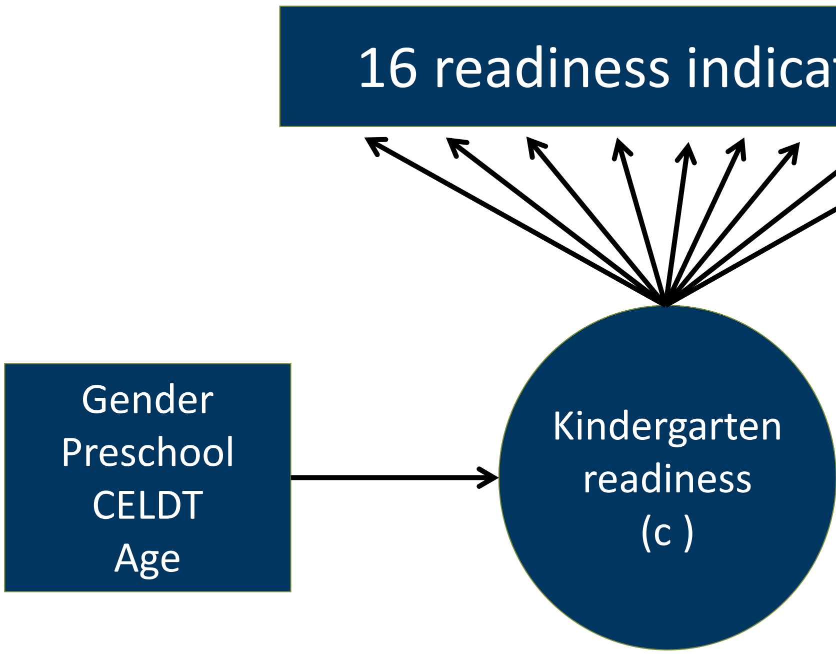
   (Steps 2–3: Day 1)
4. Add potential predictors (and check for measurement invariance).
5. Add potential distal outcomes.

   (Steps 4–5: Day 2)

IMMERSE Project – IES funded Training Grant (R305B220021)

UC SANTA BARBARA

Latent Classes of Kindergarten Readiness



Kindergarten Student Entrance Profile (KSEP)

A. Seeks adult help when appropriate

B. Engages in cooperative play activities with peers

C. Exhibits impulse control and self-regulation

D. Stays with or repeats a task

E. Separates appropriately from caregiver

F. Is enthusiastic and curious in approaching new activities

G. Follows rules when participating in routine activities

H. Uses tools with increasing precision

I. Demonstrates general coordination

J. Demonstrates sense of own body in relation to others

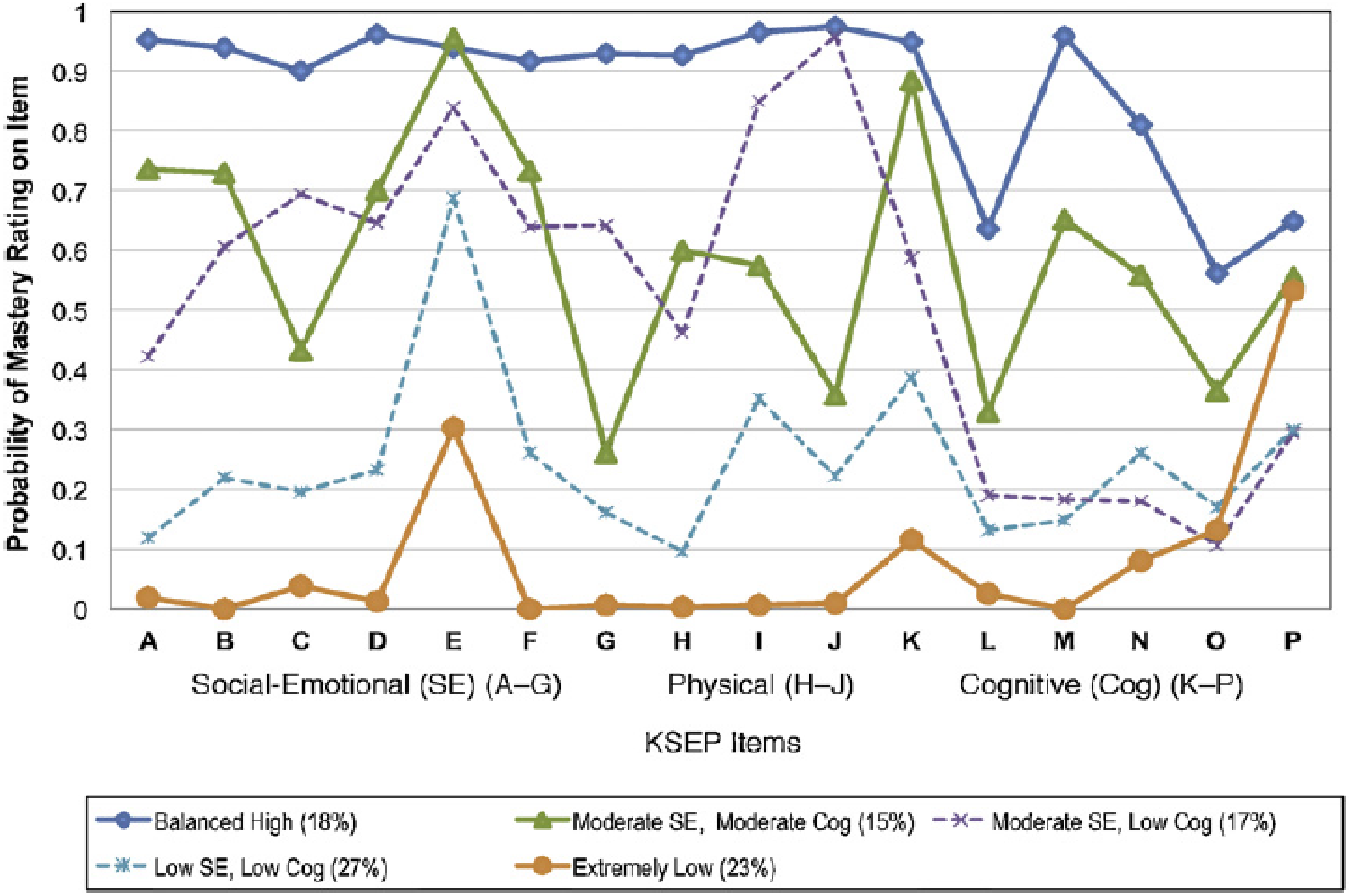
K. Recognizes own written name

L. Writes own name

M. Demonstrates expressive abilities

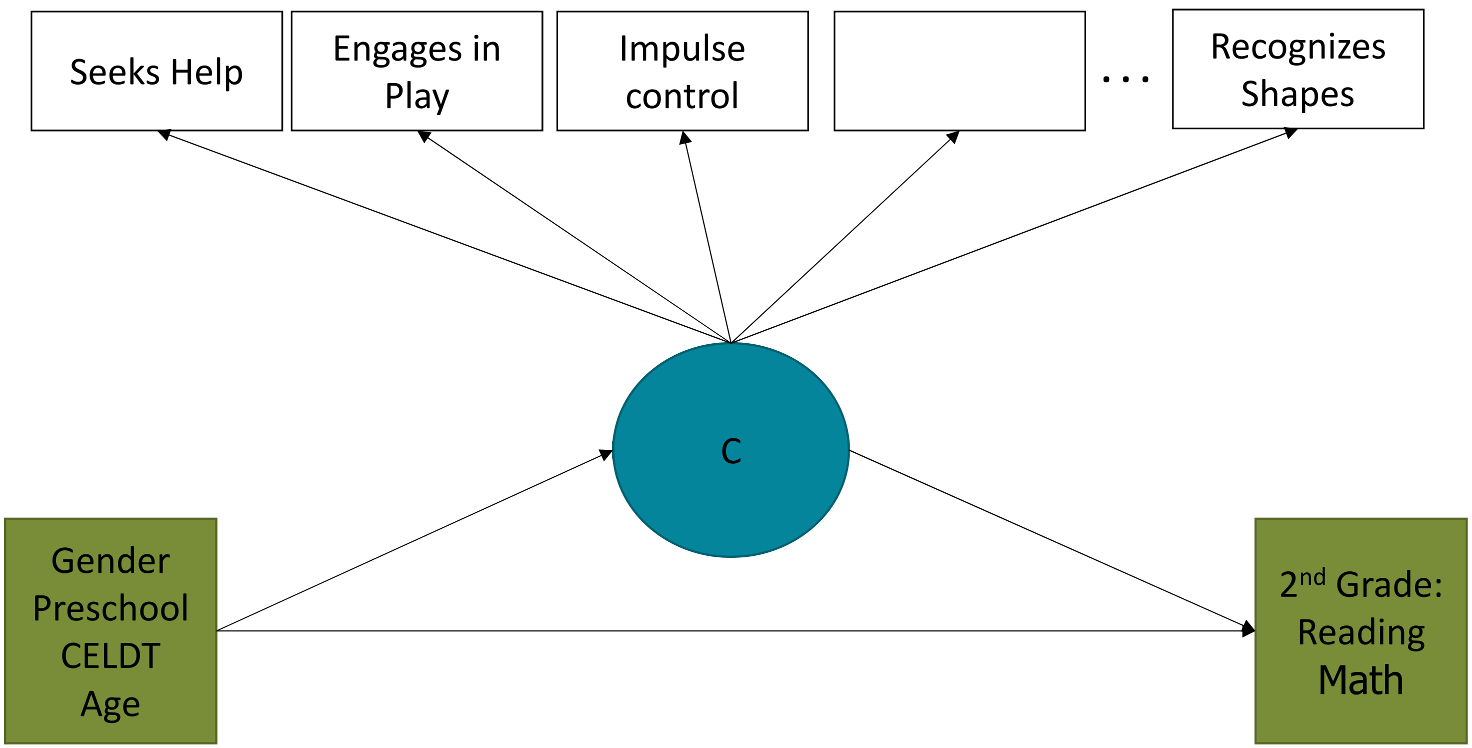
N. Understands that numbers represent quantity

O. Recognizes colors



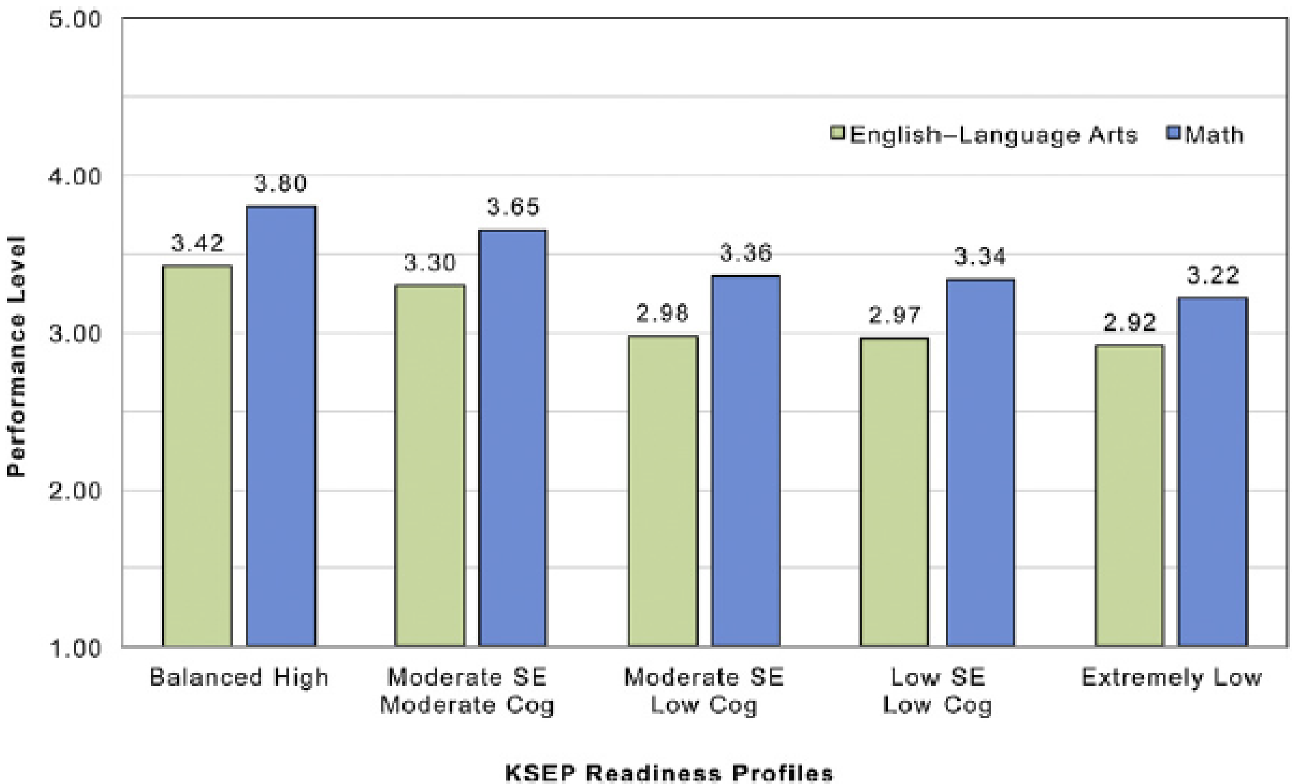
Quirk, M., Nylund-Gibson, K., & Furlong, M. (2013). Exploring patterns of Latino/a children's school readiness at kindergarten entry and their relations with grade 2 achievement. *Early Childhood Research Quarterly, 28*(2), 437-449.

LCA with Auxiliary Variables



IMMERSE Project – IES funded Training Grant (R305B220021)

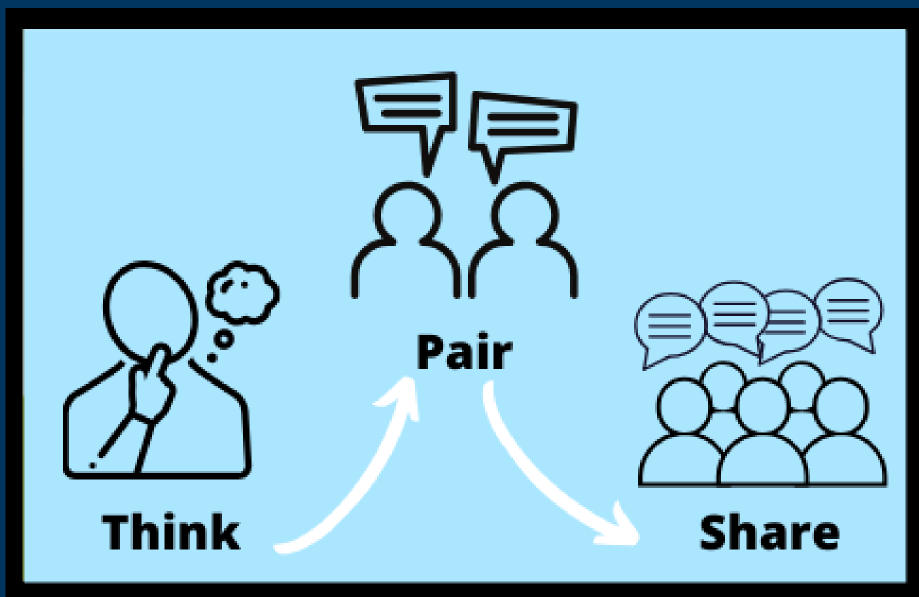
Distal Outcome Comparisons



Quirk, M., Nylund-Gibson, K., & Furlong, M. (2013). Exploring patterns of Latino/a children's school readiness at kindergarten entry and their relations with grade 2 achievement. *Early Childhood Research Quarterly, 28*(2), 437-449.

UC SANTA BARBARA

The Latent Class Measurement Model

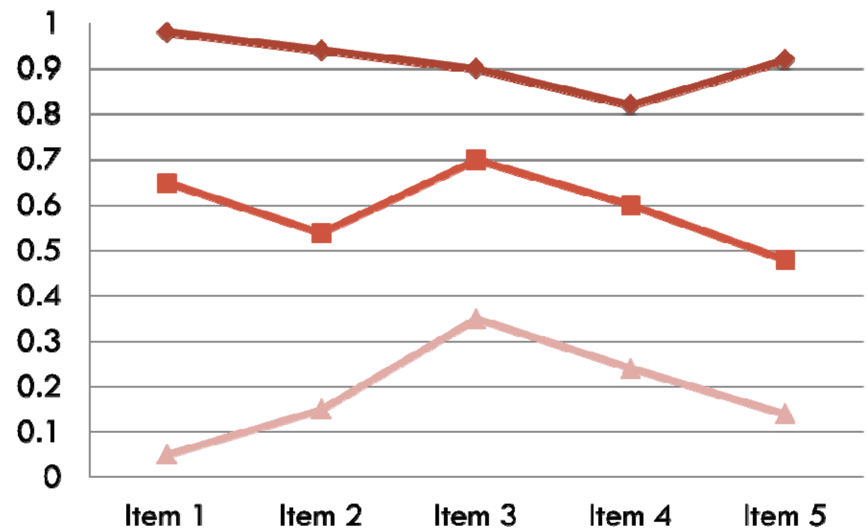


- What is a latent variable?
- What is a manifest or indicator variable?
- What is a latent variable(LV) measurement model?
- How do we select indicator variables for a latent variable measurement model?
- How might data equity figure into indicator selection?

Selection of Latent Class Indicators (items)

Items that all measure the same trait/ characteristic will sometimes result in ordered classes

Items that load together on a single factor (**one dimension**) will usually result in ordered classes.



What might be the rationale for using a latent class model over a factor/IRT measurement model in this case?

IMMERSE Project – IES funded Training Grant (R305B220021)

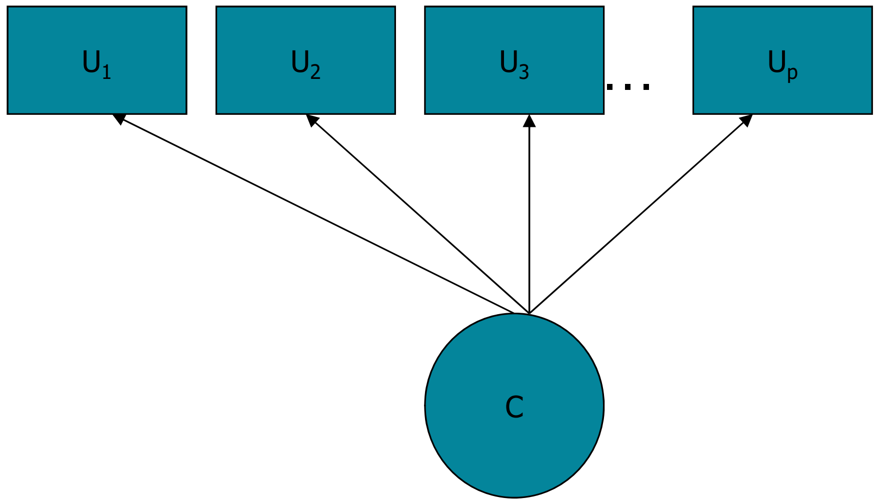
Selection of Latent Class Indicators (items)

- Consider face validity
- Consider content validity
- How many items representing each dimension/(sub)domain of the latent class construct are needed?
  - Should there be the same number of indicators for each (sub)domain?
- Do all items need to have the same measurement scale/modality?

LCA vs. Factor Analysis

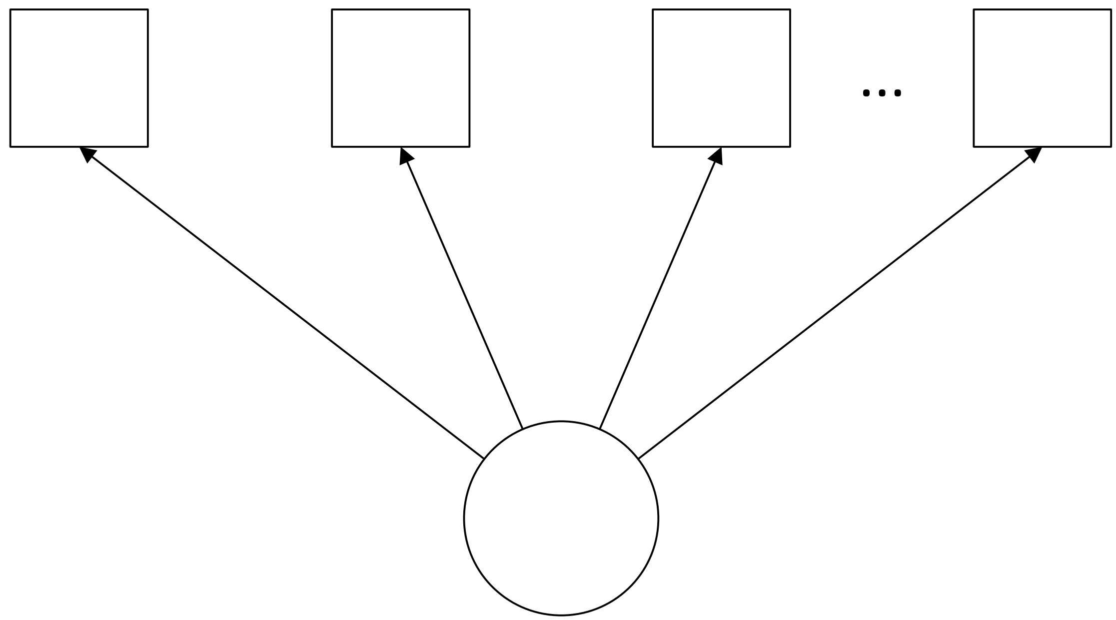
Latent Class Analysis	Factor Analysis
✓Groups people or observations	✓Groups variables
✓Accommodates continuous and/or categorical outcomes (traditionally categorical)	✓Accommodates continuous and/or categorical outcomes (traditionally continuous)
✓Latent variable is categorical (multinomial distribution)	✓Latent variable is continuous (usually assumed to be normally distributed)
✓Person-centered approach	✓Variable-centered approach

The Traditional LCA Model

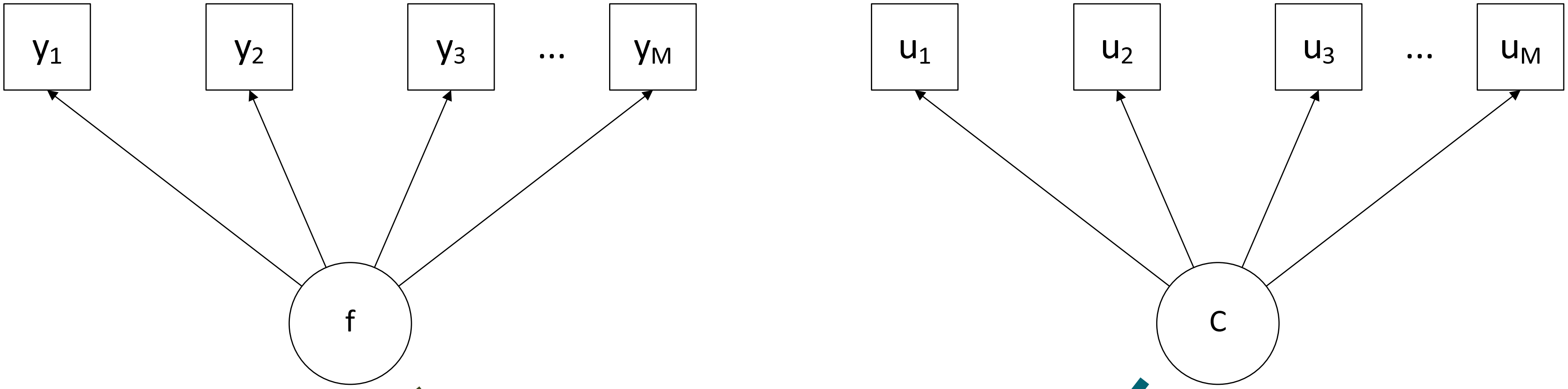


- Categorical items/indicators (u's)
- Categorical latent class variable (c)
- Cross-sectional data

Factor Model or Mixture Model?



Factor Model or Mixture Model?



This is why LCA is sometimes presented as the categorical analogue to factor analysis

LCA vs. Factor Analysis

Latent Class Analysis	Factor Analysis
✓**Person-centered approach**	✓**Variable-centered approach**

Don't let this language cause you to lose sight of the fact that *both* are latent **variable** measurement models, specifying stochastic relationships between an unobserved latent variables and multiple observed indicators. Many of the principled practices for establishing and evaluating factor and IRT measurement models do translate to latent class and latent profile measurement models.

IMMERSE Project – IES funded Training Grant (R305B220021)

Applications of LCA

- Binary test items as multiple indicators for an underlying 2-level categorical latent variable representing profiles of Mastery and Non-mastery.

- DSM-IV symptom checklist (diagnostic criteria) for depression.

This might be a situation that calls for a *confirmatory* latent class analysis approach. Most applied LCAs are exploratory. Should they be?

IMMERSE Project – IES funded Training Grant (R305B220021)

Example Data

Student	Item 1	Item 2	Item 3	Item 4	Sum
1	1	1	1	1	4
2	0	0	0	0	0
3	1	0	1	0	2
4	0	1	0	1	2
5	0	0	1	1	2
6	1	1	1	0	3
7	1	1	0	1	3

Naïve approach

- Create a cut-point based on the sum score, e.g., clinical depression if satisfying 5 or more of the 9 symptoms; mastery defined as 80% of items correctly answered.

- Problems
  - Treats all items the same, e.g., doesn't take into account that some items may be more "difficult" than others
  - Doesn't take into account measurement error, e.g., some with Mastery status may still make a careless error.

LCA Approach

- Characterizes groups of individuals based on response patterns for multiple indicators.
- Class membership “explains” observed covariation between indicators.
- Allows for measurement error in that class-specific item probabilities may be between zero and one.
- Allows comparisons of indicator sensitivity and specificity to identify items that best differentiate the classes
- Estimates the prevalence of each class in the population
- Estimates class membership probabilities for individuals

LCA vs. Factor Analysis Measurement Parameters

Latent Class Analysis	Factor Analysis
✓Item probabilities for each item in each class (a.k.a., class-specific or conditional [on class membership] item probabilities)	✓Factor Loadings ✓Item intercepts

IMMERSE Project – IES funded Training Grant (R305B220021)

LCA vs. Factor Analysis Structural Parameters

Latent Class Analysis	Factor Analysis
✓Class probabilities	✓Factor Means
• ~~Number of classes~~	✓Factor Variances
	• ~~Number of factors~~

IMMERSE Project – IES funded Training Grant (R305B220021)

Latent Class Measurement Model

- LCA model with *r* observed binary items, *u*, has a categorical latent variable *c* with *K* classes (*c = k; k = 1, 2,…, K*). The marginal item probability for item an $u_j = 1$,

$$P\left(u_j = 1\right) = \sum_{k=1}^{K} P(c = k) P(u_j = 1 \mid c = k)$$

IMMERSE Project – IES funded Training Grant (R305B220021)

LCA: Two sets of parameters

$$P\left(u_j = 1\right) = \sum_{k=1}^{K} P(c = k) P(u_j = 1 | c = k)$$

STRUCTURAL: Population proportion of each class.

This is the relative class size: How big each class is.

E.g., class 1 is 45% of the population

MEASUREMENT: Conditional item probabilities

These are the parameters that define classes. Think of these like the factor loading, as they are how the items relate to the latent variable

Latent Class Measurement Model

- Assuming **conditional independence**, the joint probability of all the *r* observed items is

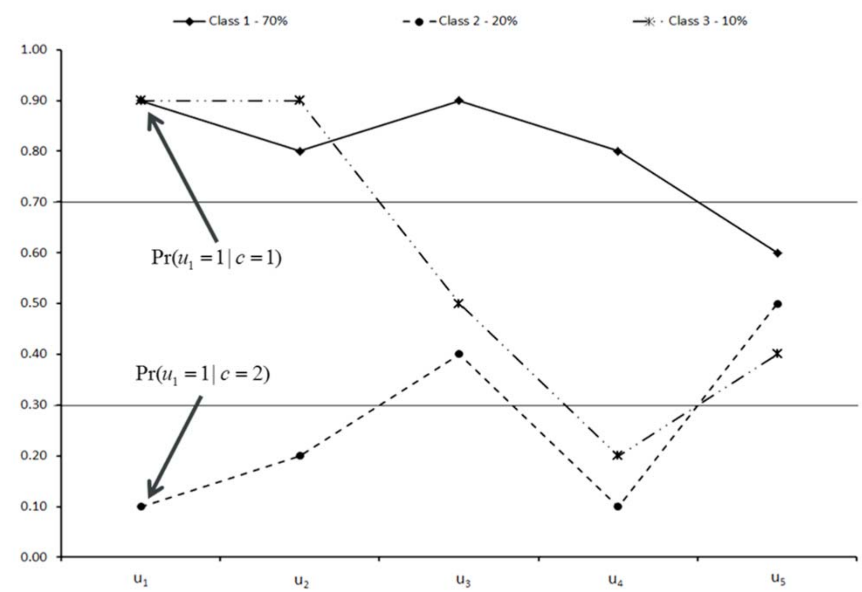
$$P\left(u_1, u_2, \ldots, u_r\right) = \sum_{k=1}^{K} P(c=k) P(u_1 | c=k) P(u_2 | c=k) \ldots P(u_r | c=k)$$

This is the *default* in Mplus! What does this assumption mean? Is it necessary for the model to be identified? When, if ever, would we relax this assumption?

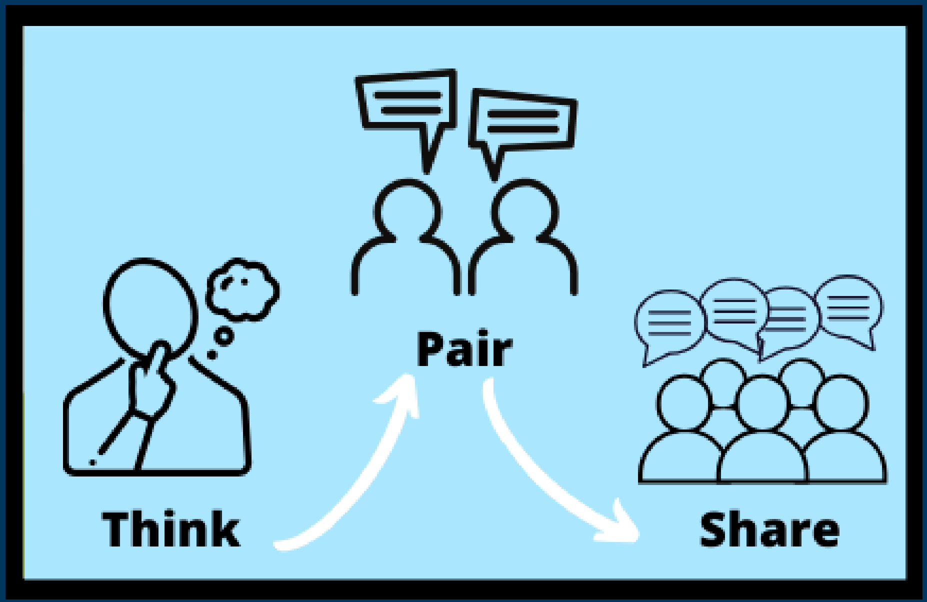
IMMERSE Project – IES funded Training Grant (R305B220021)

LCA: Conditional Item Probabilities

- For binary items, we often plot these in item probability plots.



Evaluating the Quality of the Measurement Model



- What does it mean for an LV measurement model to be "good"?
- How do we evaluate the quality of items in a factor or IRT model? What might be the analogue in an LCA with binary items?
- How do we evaluate the quality of the factor or IRT scores? What might be the analogue in an LCA with binary items?

IMMERSE Project – IES funded Training Grant (R305B220021)

Overall Measurement Model Quality

- Overall fit to the data, i.e., model-data consistency.
  - If the measurement model does not fit the data, then there isn't much point in getting into reliability and validity evaluations. [Remember: If the model doesn't fit the data, it is definitely a wrong model.]
  - More on this in the section on class enumeration…
- Relative fit to the data compared to alternative models.
  - If there is an alternative model that fits the data nearly-as-well or better than the model at-hand, it must be eliminated using logic and theory—the data cannot be The Decider.
  - **More on this in the section on class enumeration…**

IMMERSE Project – IES funded Training Grant (R305B220021)

Item-level Measurement Quality

Ideally, we'd like our LCA measurement model to have both these features:

- **Class homogeneity**

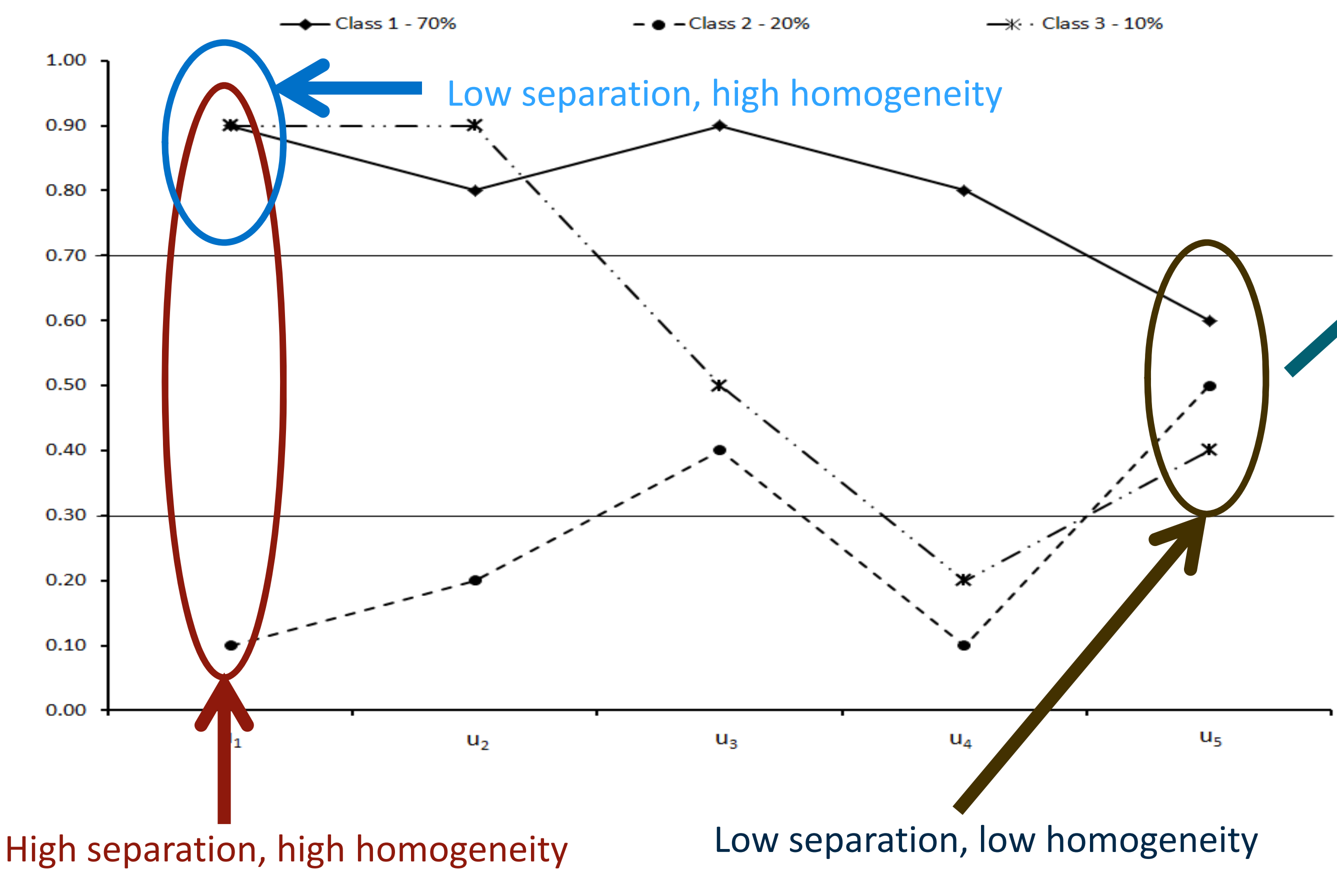
  Individuals within a given class are similar to each other with respect to item responses, e.g., for binary items, class-specific response probabilities above .70 or below .30 indicate high homogeneity.

- **Class separation**

  Individual across two classes are dissimilar with respect to item responses, e.g., for binary items, odds ratios (ORs) of item endorsements between two classes >5 or <.2 indicate high separation.

IMMERSE Project – IES funded Training Grant (R305B220021)

Conditional Item Probability Plots



Does this mean u_5 is a bad item? Should it be removed as an indicator in the measurement model? What if there was a class that had low homogeneity for *all* items?

NOTE: Just as it is nonsensical to have highly valid but unreliable factors, there cannot be latent classes with high separation and low homogeneity.

	$\hat{\omega}_{m\|k}$				$\hat{OR}_{m\|j,m\|k}$		
Item	Class 1 (70%)	Class 2 (20%)	Class 3 (10%)		Class 1 vs. 2	Class 1 vs. 3	Class 2 vs. 3
u_1	**.90***	**.10**	**.90**		**81.00****	1.00	**0.01**
u_2	**.80**	**.20**	**.90**		**16.00**	0.44	**0.03**
u_3	**.90**	.40	.50		**13.50**	**9.00**	0.67
u_4	**.80**	**.10**	**.20**		**36.00**	**16.00**	0.44
u_5	.60	.50	.40		1.50	2.25	1.50

* Item probabilities >.7 or <.3 are bolded to indicate a high degree of class homogeneity.
** Odds ratios >5 or <.2 are bolded to indicate a high degree of class separation.

Adapted from : Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods* (pp. 551–611). New York, NY: Oxford University Press.

$$\widehat{OR}_{1v2} = \frac{P(u_1 = 1|c_1) / P(u_1 = 0|c_1)}{P(u_1 = 1|c_2) / P(u_1 = 0|c_2)}$$

	$\hat{\omega}_{m\|k}$				$\hat{OR}_{m\|j,m\|k}$		
Item	Class 1 (70%)	Class 2 (20%)	Class 3 (10%)		Class 1 vs. 2	Class 1 vs. 3	Class 2 vs. 3
u_1	**.90***	**.10**	**.90**		**81.00****	1.00	**0.01**
u_2	**.80**	**.20**	**.90**		**16.00**	0.44	**0.03**
u_3	**.90**	.40	.50		**13.50**	**9.00**	0.67
u_4	**.80**	**.10**	**.20**		**36.00**	**16.00**	0.44
u_5	.60	.50	.40		1.50	2.25	1.50

* Item probabilities >.7 or <.3 are bolded to indicate a high degree of class homogeneity.

** Odds ratios >5 or <.2 are bolded to indicate a high degree of class separation.

$$\frac{.9/.1}{.1/.9} = 81$$

$$\frac{.1/.9}{.9/.1} = .01$$

Adapted from : Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods* (pp. 551–611). New York, NY: Oxford University Press.

Latent Class “Reliability”

- Factor determinacy, often interpreted at the reliability of the factor scores, is the estimated correlation between the estimated factor scores and the “true” factor scores (also defined as the coefficient of multiple determination—multiple R^2—between the items and the factor scores).
- What is the categorical analogue?...Posterior class probabilities and classification precision.

Posterior Class Probabilities

- The model-estimated values for each individual's probabilities of being in each of the latent classes (i.e., the probability of "true" class membership) based on the maximum likelihood parameter estimates and the individual's observed responses on the indicator variables.

$$\hat{p}_{ik} = \hat{\Pr}(c_i = k \mid \mathbf{u}_i, \hat{\boldsymbol{\theta}}) = \frac{\hat{\Pr}(\mathbf{u}_i \mid c_i = k, \hat{\boldsymbol{\theta}}) \cdot \hat{\Pr}(c_i = k)}{\hat{\Pr}(\mathbf{u}_i)},$$

- A single summary measure (like the factor determinacy) to summarize the posterior class probabilities across all latent classes and all individuals in the sample: Relative Entropy.

Relative Entropy

- An index that summarizes the overall precision (i.e., reliability) of classification for the whole sample across all the latent classes:

$$E_K = 1 - \frac{\sum_{i=1}^{n}\sum_{k=1}^{K}\left[-\hat{p}_{ik}\ln(\hat{p}_{ik})\right]}{n\log(K)}.$$

- When posterior classification is no better than random guessing, E_K=0, and when there is perfect posterior classification for all individuals in the sample, E_K =1.

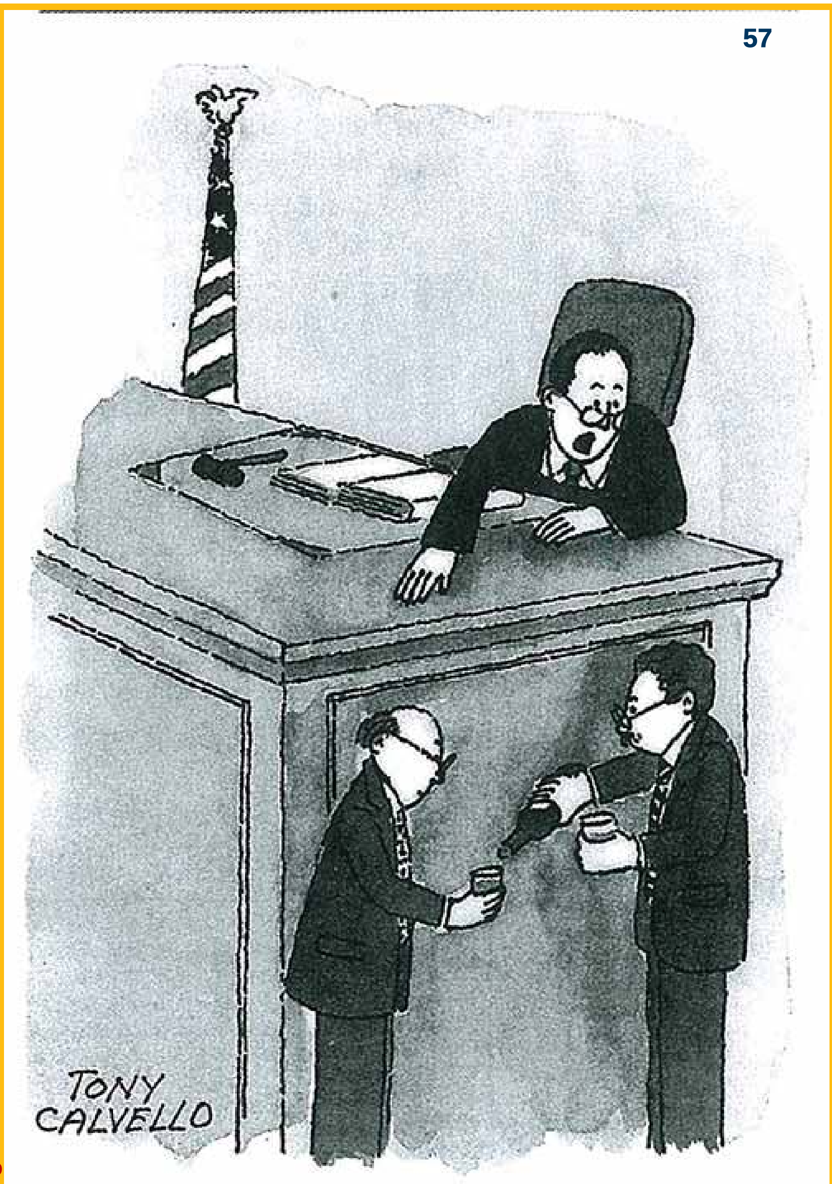
Relative Entropy (cont'd)

- Since even when E_K is close to 1.00 there can be a high degree of latent class assignment error for particular individuals, and since posterior classification uncertainty may increase simply by chance for models with more latent classes, **E_K was never intended nor should it be used for model selection during the class enumeration process.** (NOTE: A mixture model with low entropy could still fit the data well.)
- However, values near zero may indicate that the latent classes are not sufficiently well-separated for the classes that have been estimated. Thus, E_K may be used to identity problematic over-extraction of latent classes and may also be used to judge the utility of the latent class analysis directly applied to a particular set of indicators to produce empirically, highly-differentiated groups in the sample.
- **More on this in the section on class enumeration…**

IMMERSE Project – IES funded Training Grant (R305B220021)

SIDEBAR:

- Mplus parameterization and syntax for observed **categorical *endogenous/dependent* variables**



VARIABLE: !Mplus command

Names are	names of the variables in the order in which they appear in the data set;
UseVariables are	names of variables to be included in model;
Categorical are	names of ordered categorical ***dependent*** variables (binary/ordinal);
Nominal are	names of unordered categorical ***dependent*** variables (multinomial);
Count are	names of count ***dependent*** variables(Poisson default);

IMMERSE Project – IES funded Training Grant (R305B220021)

Categorical Independent/Exogenous Variables

Why?

If you have categorical exogenous/independent variables (e.g., covariates), you include them in your model the same as you would in a linear regression, e.g., dummy variables, contrasts, etc.

DO NOT identify them as "categorical" in the VARIABLE command in Mplus. You can create dummy variables using the DEFINE command within Mplus or outside of Mplus (e.g., in R) before creating the Mplus data file.

IMMERSE Project – IES funded Training Grant (R305B220021)

Category Coding

- The estimation of the model for binary or ordered categorical dependent variables uses zero to denote the lowest category, one to denote the second lowest category, etc.
- If the variables are not coded this way in the data, they are automatically recoded.
- The original data file is not overwritten but when data are saved using the SAVEDATA option, the recoded categories are saved.
- Mplus codes the lowest category as "0" but refers to it in the output as "Category 1".

IMMERSE Project – IES funded Training Grant (R305B220021)

Variations on CATEGORICAL

- Categorical = u1-u3;
  - By default, the number of categories for each variable is determined from the data. (Max categories is 10.)
- Categorical = u1-u3(*);
  - The categories of each variable are to be recoded using the categories found in the data for the set of variables rather than for each variable.
    - [This is useful when a response category is not observed on a particular variable.]
- Categorical = u1-u3 (1-5);
  - (1-5) is the set of categories allowed for a variable or set of variables.
- Categorical = u1-u3 (*) | u4-u6 (2-4) | u7-u9
  - Allows different options for different variables

NOMINAL Option

- By default, the number of categories is determined from the data.
- Nominal variables cannot have more than 10 categories.
- (Re)coding of nominal dependent variables is the same as for ordinal.
- The *last* (i.e., highest label) category is the reference category in the multinomial logistic regression parameterization.
  - There is not an override option. If you want a different category as the reference, you much recode the data so that the desired reference category has the highest value label.

IMMERSE Project – IES funded Training Grant (R305B220021)

Latent Response Variable Parameterization

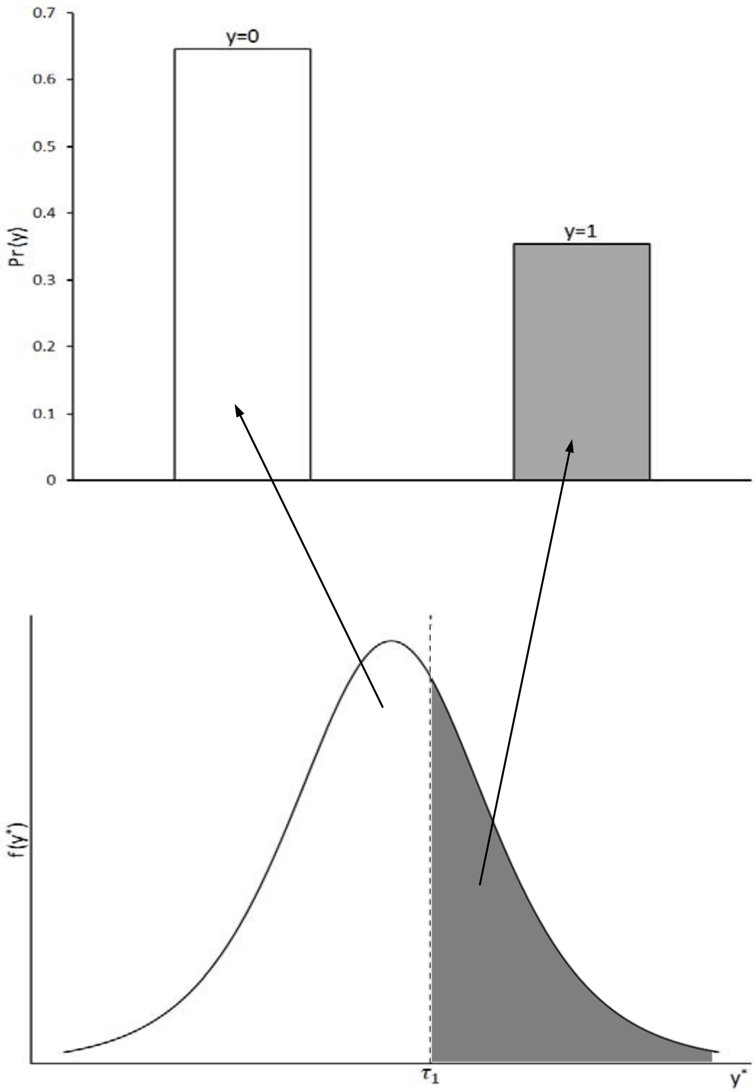
- Mplus parameterizes the (conditional) distributions of all ***endogenous/dependent*** binary and ordinal observed variables using the latent response variable (LRV) formulation.
- This is a flexible (and equivalent!) alternative parameterization (to working on a probability scale) that easily integrates into a larger (latent) variable system.
- The LRV approach assumes that a *distinct* latent continuous response variable, y*, ranging from (-∞ , +∞), has generated *each* observed, categorical variable, y.

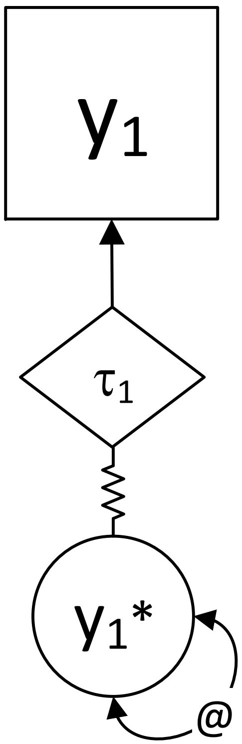
Binary Observed Variable w/ LRV Parameterization

- Assume that a latent variable, y*, ranging from (-∞ , +∞), has generated an observed variable, y, which is binary.

$$y_i = \begin{cases} 1 \text{ if } y_i^* > \tau_1 \\ 0 \text{ if } y_i^* \le \tau_1 \end{cases},$$

- τ_1 (tau one) is what Mplus calls the "threshold" for *y* and refers to as "[y$1]" in the Mplus MODEL syntax.
- If y* (or the errors thereof in a conditional model) is assumed to have a standard *logistic* distribution, then the LRV model will be equivalent to a generalized linear model using a **logit** link function.





What threshold value would correspond to a very small probability for Y = 1, say <.001? What threshold value would correspond to a very large probability for Y = 1, say >.999?

```
. logit selfinjury2 sexmin

Logistic regression                             Number of obs   =      18442
                                                LR chi2(1)      =     422.79
                                                Prob > chi2     =     0.0000
Log likelihood = -12394.484                     Pseudo R2       =     0.0168

------------------------------------------------------------------------------
 selfinjury2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sexmin |   1.109201   .0559836    19.81   0.000     .9994753    1.218927
       _cons |  -.3730469   .0156711   -23.80   0.000    -.4037617   -.3423321
------------------------------------------------------------------------------
```

```
. ologit selfinjury2 sexmin

Ordered logistic regression                     Number of obs   =      18442
                                                LR chi2(1)      =     422.79
                                                Prob > chi2     =     0.0000
Log likelihood = -12394.484                     Pseudo R2       =     0.0168

------------------------------------------------------------------------------
 selfinjury2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sexmin |   1.109201   .0559836    19.81   0.000     .9994753    1.218927
-------------+----------------------------------------------------------------
       /cut1 |   .3730469   .0156711                      .3423321    .4037617
------------------------------------------------------------------------------
```

```
. logit selfinjury2 sexmin

Logistic regression                               Number of obs   =      18442
                                                  LR chi2(1)      =     422.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -12394.484                       Pseudo R2       =     0.0168

------------------------------------------------------------------------------
 selfinjury2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sexmin |   1.109201   .0559836    19.81   0.000     .9994753    1.218927
       _cons |  -.3730469   .0156711   -23.80   0.000    -.4037617   -.3423321
------------------------------------------------------------------------------
```

```
. ologit selfinjury2 sexmin

Ordered logistic regression                       Number of obs   =      18442
                                                  LR chi2(1)      =     422.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -12394.484                       Pseudo R2       =     0.0168

------------------------------------------------------------------------------
 selfinjury2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sexmin |   1.109201   .0559836    19.81   0.000     .9994753    1.218927
-------------+----------------------------------------------------------------
       /cut1 |   .3730469   .0156711                      .3423321    .4037617
------------------------------------------------------------------------------
```


IMMERSE Project – IES funded Training Grant (R305B220021)

Interpreting Binary Logistic Regression Parameters

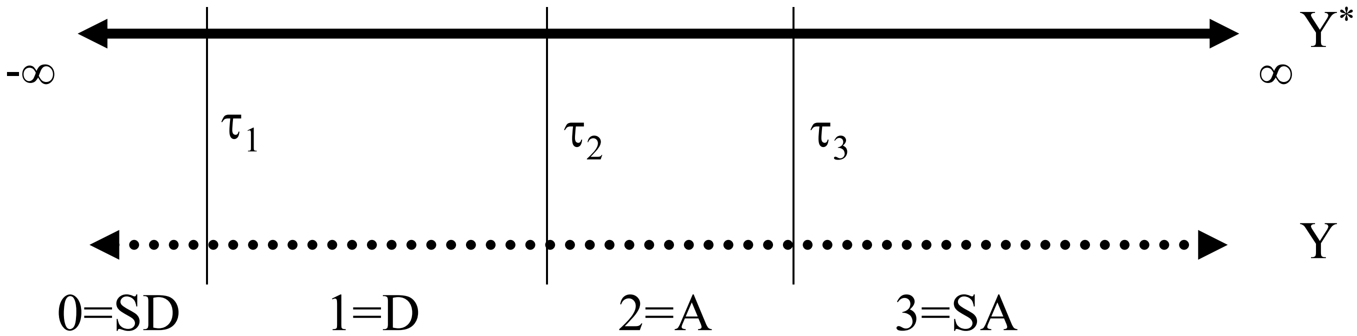
$$\text{logit}\left(Y\right) = \log\left(\frac{\Pr(Y=1\mid x)}{1-\Pr(Y=1\mid x)}\right) = -\tau_1 + \beta_1 x$$

- $-\tau_1$ represents the log($\text{odds}_{Y|x}$) when x=0.
- $-\tau_1 = \beta_0$ (intercept) from the traditional logistic regression, i.e., the estimated "threshold" in Mplus is simply (-1) x (logit intercept).
- $\beta_1 = \beta_1$ from the traditional logistic regression (i.e., log odds ratio (OR) for Y corresponding to a positive one-unit difference in x).
- NOTE: In the unconditional LCA (i.e., without covariates or distal outcomes), Mplus provides ***both*** the threshold estimates AND the model-estimated, class-specific item probabilities.

IMMERSE Project – IES funded Training Grant (R305B220021)

Ordinal Observed Variable w/ LRV Parameterization

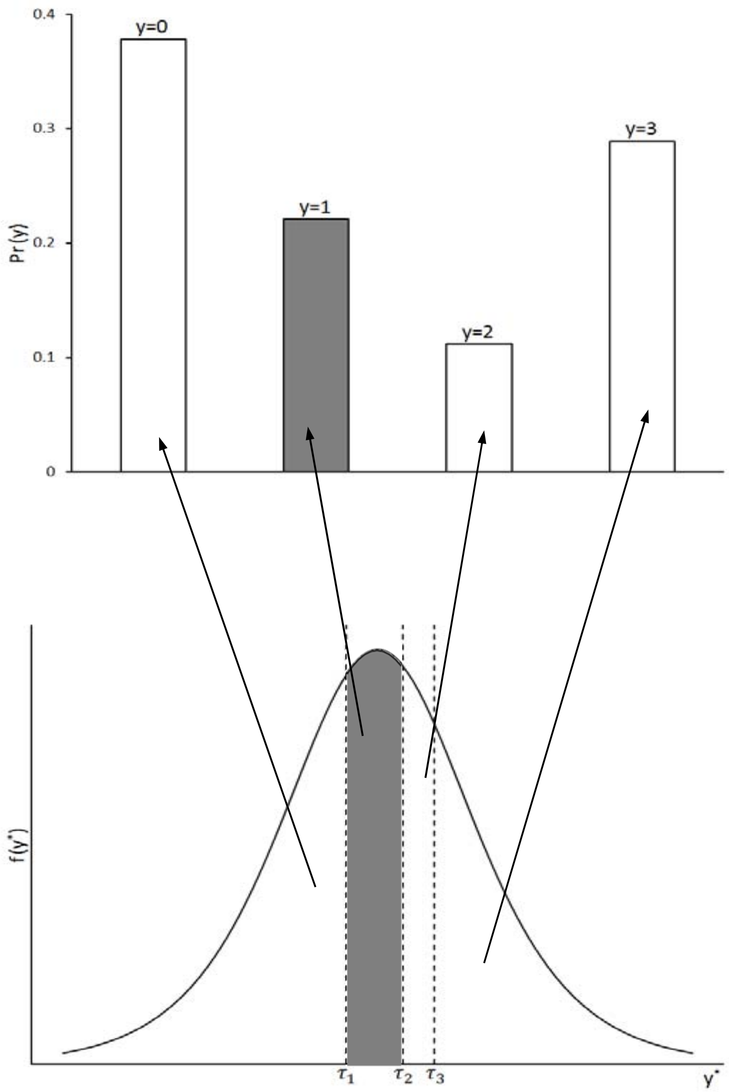
- Assume that a latent variable, y*, ranging from (-∞ , +∞), has generated an observed variable, y, which is ordinal. For example:

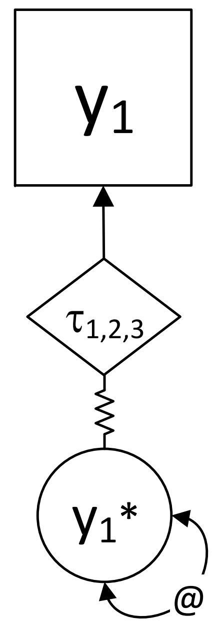


$$y_i = \begin{cases} 0 = SD & \text{if } \tau_0 = -\infty \le y_i^* < \tau_1 \\ 1 = \mathrm{D} & \text{if } \tau_1 \le y_i^* < \tau_2 \\ 2 = \mathrm{A} & \text{if } \tau_2 \le y_i^* < \tau_3 \\ 3 = \mathrm{SA} & \text{if } \tau_3 \le y_i^* < \tau_4 \end{cases}$$

Ordinal Observed Variable w/ LRV Parameterization

- For ordinal categorical dependent variables, there are as many thresholds as there are categories minus one (1). The thresholds are referred to in the MODEL command by adding a dollar sign ($) followed by a number to the variable name.
  - E.g., the two thresholds for a three-category ordinal variable, u2, are referred to as [u2$1] and [u2$2].
- If y* (or the errors thereof in a conditional model) is assumed to have a standard *logistic* distribution, then the LRV model will be equivalent to a cumulative log odds model using a **logit** link function.





$$\log\left(\frac{\Pr(Y > j)}{\Pr(Y \leq j)}\right) = -\tau_{j+1}$$

IMMERSE Project – IES funded Training Grant (R305B220021)

LCA: Conditional Item Probabilities

- Using the latent response variable formulation, Mplus will estimate class-specific thresholds for each categorical item.
- In the case of a binary u_j, coded 0/1, the conditional probability is simply the inverse logit of the negative threshold.*
- Recall:

$$\text{logit}\left(Y\right) = \log\left(\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right) = -\tau_1$$

- Solving for Pr(Y=1):

$$\Pr\left(Y=1\right) = \frac{\exp\left(-\tau_1\right)}{1+\exp\left(-\tau_1\right)} = \frac{1}{1+\exp\left(\tau_1\right)}$$

Generalizing:

$$\Pr(u_j = 1 \mid c = k) = \frac{1}{1+\exp(\tau_{jk})}$$

*Don't worry, Mplus does the calculations and provides the model-estimated item probabilities (along with the thresholds) in the output.

LCA: Class Proportions

- Mplus parameterizes the distribution of the latent class variable in terms of an unconditional (i.e., intercepts only, no predictors) multinomial regression using the *last* class as the reference class.
- Instead of estimating K-1 class proportions, Mplus estimates the K-1 intercepts/means of the unconditional multinomial regression and then calculates the model-estimated class proportions using be taking the inverse multinomial logit of the intercepts.
- As with the thresholds and item probabilities, Mplus provides both the multinomial intercepts and class proportions in the output.
- Mplus parameterizes the relationship between covariates and the latent class variables as a *conditional* multinomial regression. More on this when we get to predictors and outcomes of latent class membership.

Why are only K-1 class proportions estimated if there are K latent classes in the model?

LCA Example in Mplus

Example: Longitudinal Study of American Life (LSAF)
(formerly the Longitudinal Study of American Youth)

- A national longitudinal study funded by the National Science Foundation(NSF)
- Designed to investigate the development of students learning and achievement, particularly related to math, science, and technology and to examine the relationship of those student outcomes across middle and high school to post-secondary education and early career choices.
- More information can be found out http://lsay.org/index.html

Research Aim

- Describe/explore heterogeneity in self-reported attitudes towards math among high school students using a finite number of discrete math disposition profiles.

or

- Develop a taxonomy of math dispositions based on self-reported attitudes towards math.

IMMERSE Project – IES funded Training Grant (R305B220021)

Survey Prompt: "Now we would like you to tell us how you feel about math and science. Please indicate for you feel about each of the following statements."	Total sample (n_T = 2675)	
	f	rf
1) I enjoy math.	1784	.67
2) I am good at math.	1850	.69
3) I usually understand what we are doing in math.	2020	.76
4) Doing math often makes me nervous or upset.	1546	.59
5) I often get scared when I open my math book see a page of problems.	1821	.69
6) Math is useful in everyday problems.	1835	.70
7) Math helps a person think logically.	1686	.64
8) It is important to know math to get a good job.	1947	.74
9) I will use math in many ways as an adult.	1858	.70

Adapted from : Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods* (pp. 551–611). New York, NY: Oxford University Press.

```
DATA:

file is lsaysubset.txt;


VARIABLE:

Names are rand LSAYID REGION URBAN GENDER MOTHED FATHED RACETH SEDEX1
SEDEX2 SEDEX3 SEDEX4 SEDEX5 SEDEX6 KMHPH1 KMHPH2 KMHPH3 KMHPH4 KMHPH5
KMHPH6 KMHSCPH1 KMHSCPH2 KMHSCPH3 KMHSCPH4 KMHSCPH5 KMHSCPH6 KSCPH1 KSCPH2
KSCPH3 KSCPH4KSCPH5 KSCPH6 HSCRE1 HSCRE5 ABIOIRT APHYIRT ASCIIRTN AMTHIRTN
CBIOIRT CPHYIRT CSCIIRTN CALGIRT CQLTIRT CMTHIRTN EBIOIRT EPHYIRT ESCIIRTN
EQLTIRT EMTHIRTN GBIOIRT GPHYIRT GSCIIRTN GBASIRT GALGIRT GQLTIRT GMTHIRTN
IBIOIRT IPHYIRT.. .;

UseVariables = ca28ar ca28br ca28cr ca28er ca28gr
      ca28hr ca28ir ca28kr ca28lr;

Categorical = ca28ar ca28br ca28cr ca28er  ca28gr
      ca28hr ca28ir ca28kr ca28lr;
```

```
UseVariables = ca28ar ca28br ca28cr ca28er  ca28gr
      ca28hr ca28ir ca28kr ca28lr;

Categorical = ca28ar ca28br ca28cr ca28er  ca28gr
      ca28hr ca28ir ca28kr ca28lr;

missing=all(9999);

classes= c(5);

ANALYSIS:

type = mixture;
starts=500 100;
processors=4;

MODEL:
!Next slide
```

```
MODEL:
  %overall%

    [c#1 c#2 c#3 c#4];

    [ca28ar$1 ca28br$1 ca28cr$1 ca28er$1 ca28gr$1
     ca28hr$1 ca28ir$1 ca28kr$1 ca28lr$1 ];

  %c#1%
    [ca28ar$1 ca28br$1 ca28cr$1 ca28er$1 ca28gr$1
     ca28hr$1 ca28ir$1 ca28kr$1 ca28lr$1 ];

  %c#2%
    [ca28ar$1 ca28br$1 ca28cr$1 ca28er$1 ca28gr$1
     ca28hr$1 ca28ir$1 ca28kr$1 ca28lr$1 ];
.
.
.
  %c#5%
    [ca28ar$1 ca28br$1 ca28cr$1 ca28er$1 ca28gr$1
    ca28hr$1 ca28ir$1 ca28kr$1 ca28lr$1 ];

OUTPUT: tech1 tech10 tech11 residual;

PLOT: type=plot3;
  series = ca28ar ca28br ca28cr ca28er  ca28gr ca28hr
       ca28ir ca28kr ca28lr(*);
```

Note:
With categorical indicators, the following model statement would produce the same result!

```
MODEL:
!empty
```

How is that even possible? What does it mean Mplus does automatically for type = mixture?

Output

```
SUMMARY OF ANALYSIS
.
.
.
Observed dependent variables

  Binary and ordered categorical (ordinal)
   CA28AR      CA28BR      CA28CR      CA28ER      CA28GR      CA28HR
   CA28IR      CA28KR      CA28LR

Categorical latent variables
   C
.
.
.
Link                                                    LOGIT
```

LSAY Example: logits to probabilities

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Latent Class 1				
Thresholds				
CA28AR$1	-2.122	0.185	-11.442	0.000
CA28BR$1	-2.539	0.242	-10.514	0.000
CA28CR$1	-3.081	0.291	-10.577	0.000
CA28ER$1	-1.791	0.371	-4.825	0.000
CA28GR$1	-15.000	0.000	999.000	999.000
CA28HR$1	-2.498	0.262	-9.533	0.000
CA28IR$1	-1.839	0.188	-9.781	0.000
CA28KR$1	-2.876	0.324	-8.866	0.000
CA28LR$1	-2.723	0.310	-8.775	0.000

RESULTS IN PROBABILITY SCALE				
Latent Class 1				
CA28AR				
Category 1	0.107	0.018	6.039	0.000
Category 2	0.893	0.018	50.392	0.000

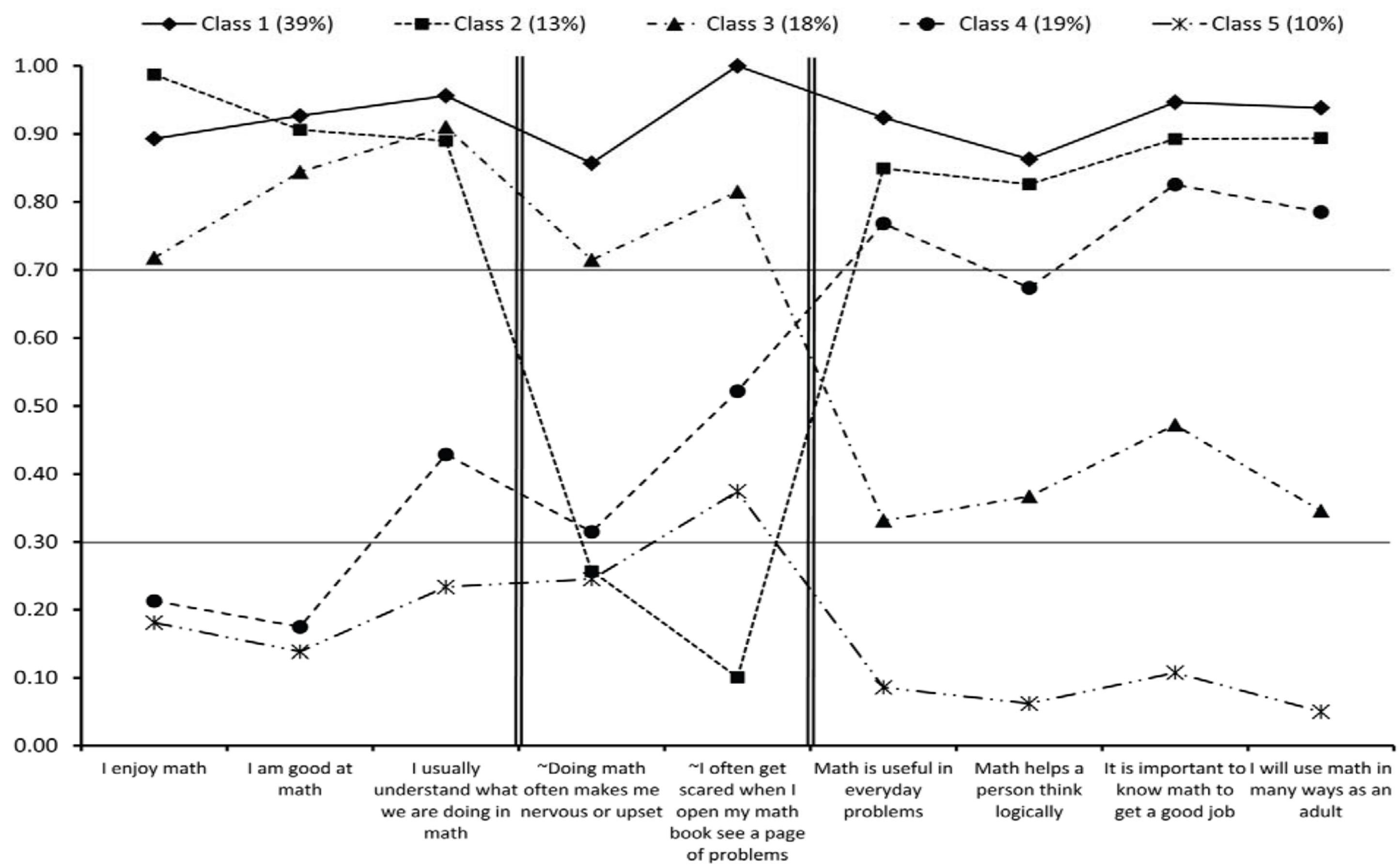
Mplus provides the item *thresholds* on the negative logit and probability scales : $\frac{1}{1+\exp(-2.122)} = 0.893$

LSAY Example: logits to probabilities

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Categorical Latent Variables				
Means				
C#1	1.321	0.195	6.776	0.000
C#2	0.216	0.376	0.574	0.566
C#3	0.555	0.227	2.447	0.014
C#4	0.597	0.211	2.821	0.005

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

Latent Classes		
1	525.13598	0.39248
2	173.96909	0.13002
3	244.13155	0.18246
4	254.57820	0.19027
5	140.18517	0.10477



1-Pro-math without anxiety, 2-Pro-math with anxiety, 3- Math Lover,
4- I don't like math but I know it's good for me, 5- Anti-Math with anxiety

Adapted from : Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods* (pp. 551–611). New York, NY: Oxford University Press.

LAB

Instructions:

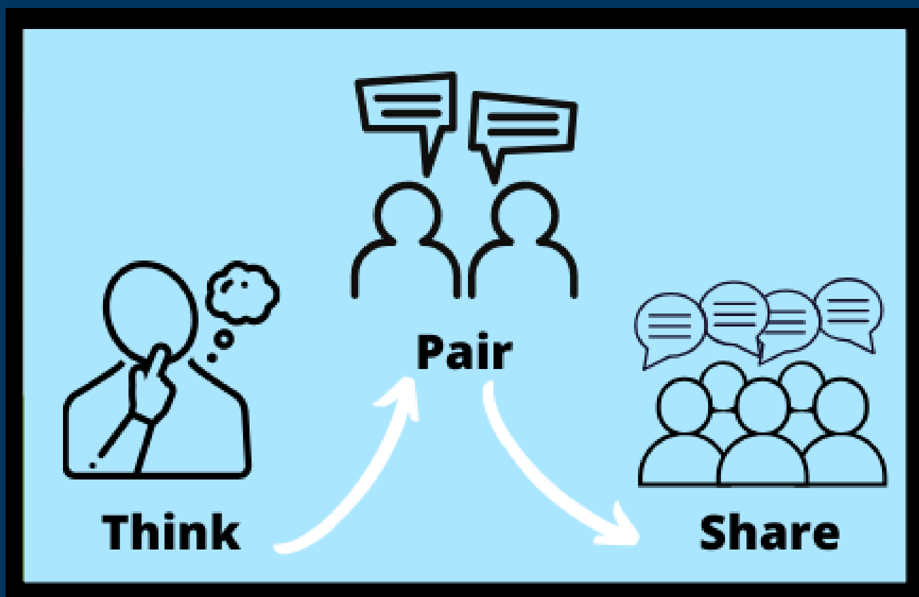
- Quick stretch break
- Material for lab will be on the Github webpage

Lunch (12:30-1:30pm)

Location: 4th Floor Terrace

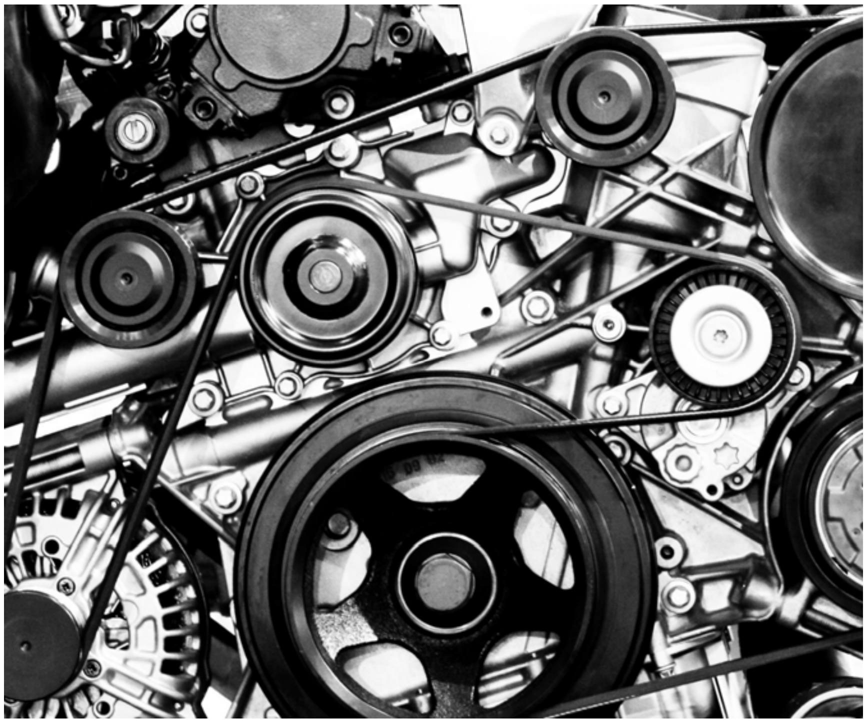
IMMERSE Project – IES funded Training Grant (R305B220021)

Mixture Model Estimation



- What is a sampling distribution for an estimate?
- What does is mean for an estimator to be unbiased?
- What are the missing-completely-at-random (MCAR) and missing-at-random (MAR) assumptions?
- How does MCAR/MAR relate to latent variable models when you have complete data (i.e., no missingness) on all observed variables in your dataset?

Getting "under the hood" of the Mplus engine



- Mplus uses maximum likelihood (ML) estimation for mixture models.
- For every analysis "type" in the ANALYSIS command, Mplus has an "estimator" default.
- For **Type = Mixture**, **Estimator = MLR** is the default.

UC SANTA BARBARA

What is Maximum Likelihood (ML) Estimation?

- A statistical method used to estimate the parameters of a probability distribution based on observed data.
- The goal of ML estimation is to find the values of the (population) parameters that maximize the **likelihood function**. The values are the maximum likelihood estimates (MLEs).
- The **likelihood function** expresses the probability of obtaining the observed data given the parameter values.

IMMERSE Project – IES funded Training Grant (R305B220021)

Why is ML the “go-to” estimation method?

- ML estimation provides estimates that are asymptotically unbiased, which means that as the sample size increases, the estimates converge to the true parameter values (i.e., means of the sampling distributions of the MLEs are equal to the true parameter values).
- ML estimation provides estimates that are asymptotically efficient, which means that as the sample size increases, the estimates have the smallest possible standard errors—in other words, have the best precision—among all other unbiased estimators (i.e., the variances of the sampling distributions of the MLEs are the smallest compared to the sampling distributions for all other type of estimates).

IMMERSE Project – IES funded Training Grant (R305B220021)

ML Estimation Steps

Specify the likelihood function for the data, which (assuming independent observations) is the *product* of the likelihood functions for each participant.

Take the natural logarithm—$\log_e(\ )$—of the likelihood function to simplify the calculations and convert the product to a sum.

- The log of a product is the sum of the logs.

Maximize the loglikelihood (LL) function with respect to the parameters by finding the values that make the function as large as possible.

- Usually, we need a special algorithm to do this step. The Mplus default is the EM algorithm.
- The values that maximize the loglikelihood are the <u>*same*</u> as the values that maximize the likelihood.

The parameter values that maximize the likelihood function are the maximum likelihood estimates (MLEs).

- ML estimation assumes: (1) The data are independent and identically distributed (i.i.d.) and (2) the chosen probability distribution is the correct model for the data.

Generic Mixture Model Likelihood Function

$$L_i = f(\mathbf{y}_i \mid \mathbf{\Omega}) = \sum_{k=1}^{K} \Pr\left(c = k\right) f_k(\mathbf{y}_i \mid \mathbf{\Omega}_k)$$

$$L = \prod_{i=1}^{n} f(\mathbf{y}_i \mid \mathbf{\Omega})$$

$$LL = \log_e\left(\prod_{i=1}^{n} f(\mathbf{y}_i \mid \mathbf{\Omega})\right) = \sum_{i=1}^{n} \log_e\left(f(\mathbf{y}_i \mid \mathbf{\Omega})\right)$$

Full Information Maximum Likelihood (FIML)

- All Mplus ML estimators are FIML estimators.
- FIML estimation is an extension of ML estimation that takes into account missing data and provides more efficient and unbiased parameter estimates compared to traditional methods that ignore (i.e., listwise-delete) observations with missing data.
    - But *only* if the missing data is MCAR or MAR.
- Key idea: Use a joint likelihood function ("complete data" likelihood) that incorporates the observed data and a model for the missing data mechanism.
    - The complete data likelihood function is maximized by summing the observed data likelihood over all possible values of the missing data, i.e., integrating-out the missingness.
    - FIML estimates the parameters that provide the best fit to the observed and missing data *simultaneously* (under MAR!).

Why is FIML the "go-to" estimation for LV models?

- With FIML, latent variables can be respecified as observed variables with missing data for *everyone*, and, thus, MCAR!
- FIML can also handle missingness on the actual observed endogenous/dependent variables as long as missingness is MCAR or MAR.
- Missingness on exogenous/independent variables?
  - Mplus will listwise-delete! (Pay attention to warnings and analysis sample size).
  - The best option is multiple imputation (MI) *unless* there is so little missingness that the listwise deletion results in a negligible decrease to the analysis sample size.
  - Caution: Be careful not to impute missingness on latent class indicators or distal outcomes if the imputation model does not include mixtures.
- Any downsides? Yes…misspecification in one part of the model can propagate to other parts of the model, resulting in biased estimates.

`Estimator = MLR` Default for `Type = Mixture`

- MLR is the Mplus abbreviation for *Robust ML* estimation.
- From the Mplus User's Guide: "MLR – maximum likelihood parameter estimates with standard errors and a chi-square test statistic (when applicable) that are robust to non-normality…"
- In regular ML estimation, the standard errors are computed *after* the MLEs are obtain using the inverse of the Fisher Information matrix evaluated at the MLEs. This results in asymptotic approximations to MLEs' standard errors.
- MLR in Mplus is designed to downplay the influence of outliers, problematic observations, and/or small latent class sizes relative to the distributional and asymptotic assumptions applied with `Estimator = ML`.

IMMERSE Project – IES funded Training Grant (R305B220021)

Algorithm = EM Default for Estimator = MLR

- EM is the abbreviation for the Expectation-Maximization (EM) algorithm.
- This is the Mplus default algorithm for ML estimation with missing or incomplete data (i.e., nearly all latent variable models, including all mixture models, unless requiring Algorithm = Integration, which Mplus will prompt if needed).
- The EM algorithm is particularly useful when dealing with latent variables or missing data that hinder the direct application of traditional maximum likelihood estimation (i.e., closed-form, multivariate calculus-based maximization) .

Basics of the EM Algorithm

1) **Initialization**: Start with an initial set of estimates for the parameters.
2) **E-step** (Expectation step): Given the current parameter estimates, calculate the expected values of the latent variables or unobserved data. This is done by computing the posterior probabilities of the missing data given the observed data and current parameter estimates.
   - Estimate the missing or latent variable values based on the current parameter estimates, effectively completing the data.
3) **M-step** (Maximization step): Maximize the expected log-likelihood function by updating the parameter estimates based on the expected values obtained in the E-step.
   - Maximize the loglikelihood function using the completed data to update the parameter estimates
4) **Iteration**: Repeat Steps 2 and 3 until convergence, which is typically determined by a predefined convergence criterion (e.g., small changes in the log-likelihood or parameter estimates).

Where does this "initial set" even come from?

Stopping Criteria for the EM Algorithm

- Iterations
  - MIterations = ___ ;
- Convergence
  - LogCriterion = ___ ;
  - RLogCriterion = ___;
  - Convergence = ___ ;

```
Optimization Specifications for the EM Algorithm
  Maximum number of iterations                                 500
  Convergence criteria
    Loglikelihood change                                 0.100D-06
    Relative loglikelihood change                        0.100D-06
    Derivative                                           0.100D-05
```

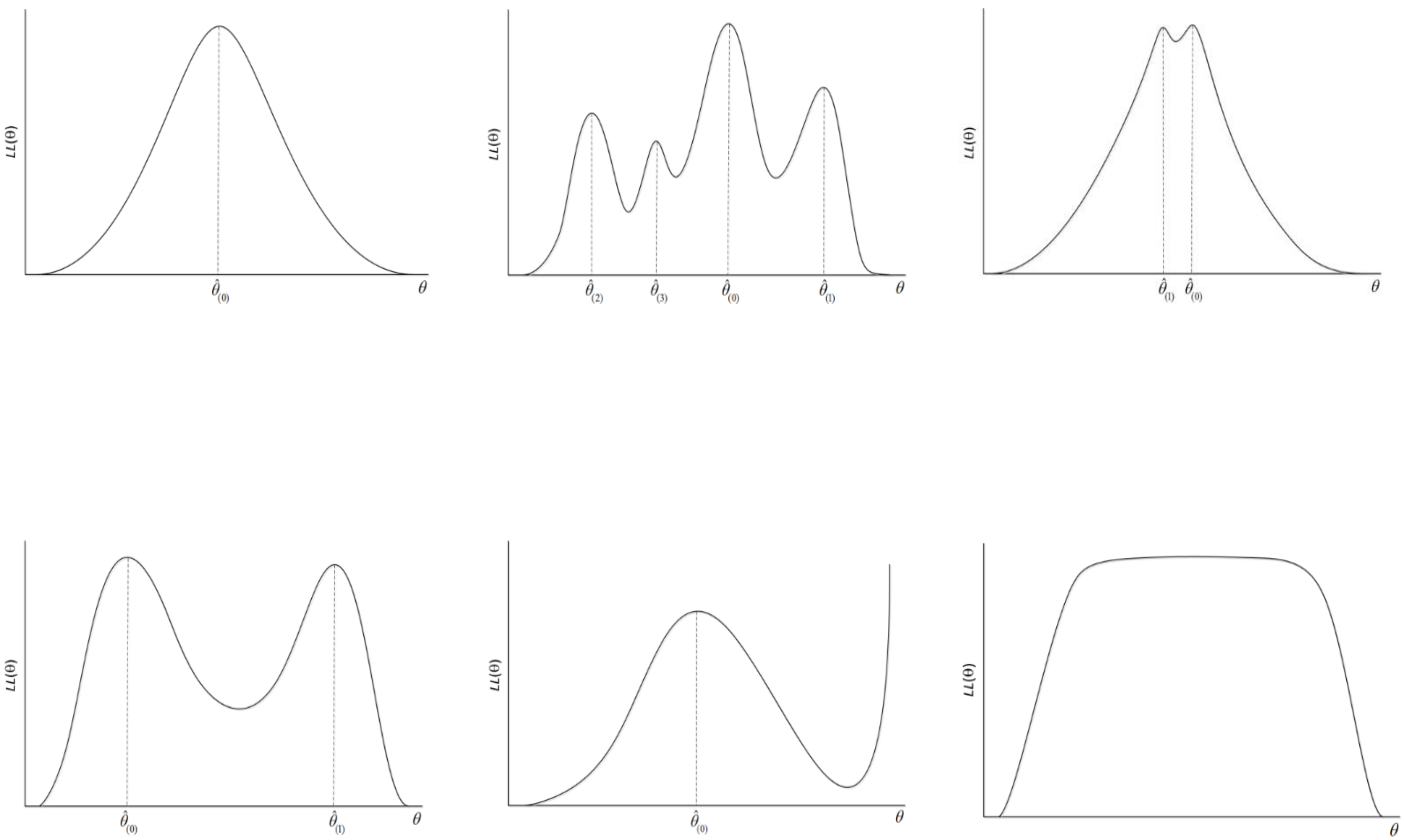
The challenges of FIML via EM

- ML estimation for mixture models can present statistical and empirical challenges that must be addressed during the application of mixture modeling:
  - The estimation may fail to converge even if the model is theoretically identified.
  - If the estimation algorithm does converge, since the log likelihood surface for mixtures is often multimodal, there is no way to prove the solution is a global rather than local maximum.

What does it mean for a model to be identified? How does empirical identification differ from theoretical identification? Can Mplus tell the difference?

What is the difference between a local and global maximum for a function?

IMMERSE Project – IES funded Training Grant (R305B220021)


LL(θ)
θ̂(0)
θ̂(2)
θ̂(3)
θ̂(1)
θ

Random Start Values (EM Step 1)

- Use multiple random sets of (initialization) starting values with the estimation algorithm—it is recommended that a minimum of 50 to 100 sets of extensively, randomly varied starting values are used (Hipp & Bauer, 2006) but more may be necessary to observe satisfactory replication of the best maximum log likelihood value, particularly as you increase the number of classes.

- Recommendations for a more thorough investigation of multiple solutions when there are more than two classes:

  ANALYSIS: Starts = 100 20;

  or with many classes

  ANALYSIS: Starts = 500 100;

```
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:

          -121464.387  544048           87
          -121464.387  153942           31
          -121464.387  551639           55
          -121464.387  575700           100
          -121464.387  471398           74
          -121464.387  848163           47
          -121464.387  259507           53
          -121464.387  462953           7
          -121464.387  903420           5
          -121464.387  85462            51
          -121464.387  696773           80
          -121464.387  569131           26
          -121464.387  533738           11
          -121464.387  813779           92
          -121464.387  120506           45
          -121464.387  352277           42

4 perturbed starting value run(s) did not converge or were rejected in the third stage.


THE BEST LOGLIKELIHOOD VALUE HAS BEEN REPLICATED.  RERUN WITH AT LEAST TWICE THE
RANDOM STARTS TO CHECK THAT THE BEST LOGLIKELIHOOD IS STILL OBTAINED AND REPLICATED.
```

```
Analysis:
  type = mixture;
  starts = 100 20;
  processors = 4;
```

We asked for 20 sets to go to the final stage.

16 of the 20 converged on the exact same value of the loglikelihood and 4 did not converge.

This is a good result but at least one more run is necessary.

Mplus will print this message ***every*** time.

```
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:
```

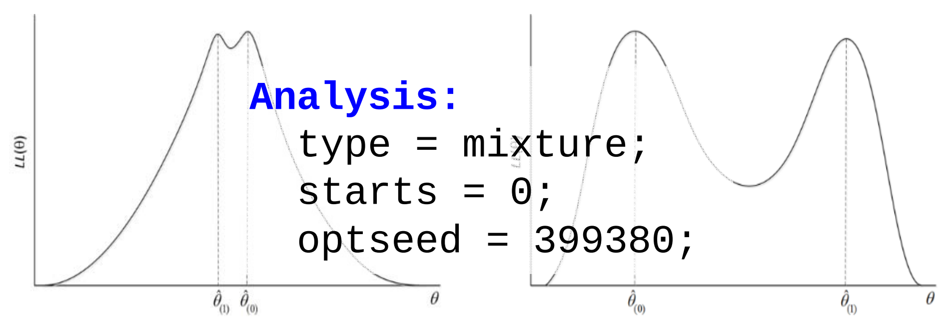
-721.236	188498	258
-721.236	930872	277
-721.306	399380	436
-721.306	399671	13
-721.375	369602	146
-721.375	273992	349
-721.375	937225	394
-721.375	780698	337
-721.375	748692	204
-721.390	965994	396
-721.390	55115	408
-721.390	751153	110
-721.390	319144	176
-721.390	392766	331
-721.390	358488	264
-721.531	534864	307
-721.556	608460	244
-721.556	618760	489
-721.583	502532	445
-721.583	341041	34
-721.583	471040	403
-721.583	499150	216
-721.583	850840	232
-721.583	164305	128
-721.583	168762	200
-721.583	794236	127
-721.583	360419	356
-721.583	464179	106
-721.827	898745	466
-722.000	211281	292
-722.034	902779	122

```
THE BEST LOGLIKELIHOOD VALUE HAS BEEN REPLICATED.  RERUN WITH AT LEAST TWICE THE
RANDOM STARTS TO CHECK THAT THE BEST LOGLIKELIHOOD IS STILL OBTAINED AND REPLICATED.
```

```
Analysis:
  type = mixture;
  starts = 500 100;
  processors = 4;
```

In this situation, there are a lot of solutions that converged right next to each other.

Need to rerun with more random starts… perhaps, starts = 2000 500;



```
Analysis:
  type = mixture;
  starts = 0;
  optseed = 399380;
```

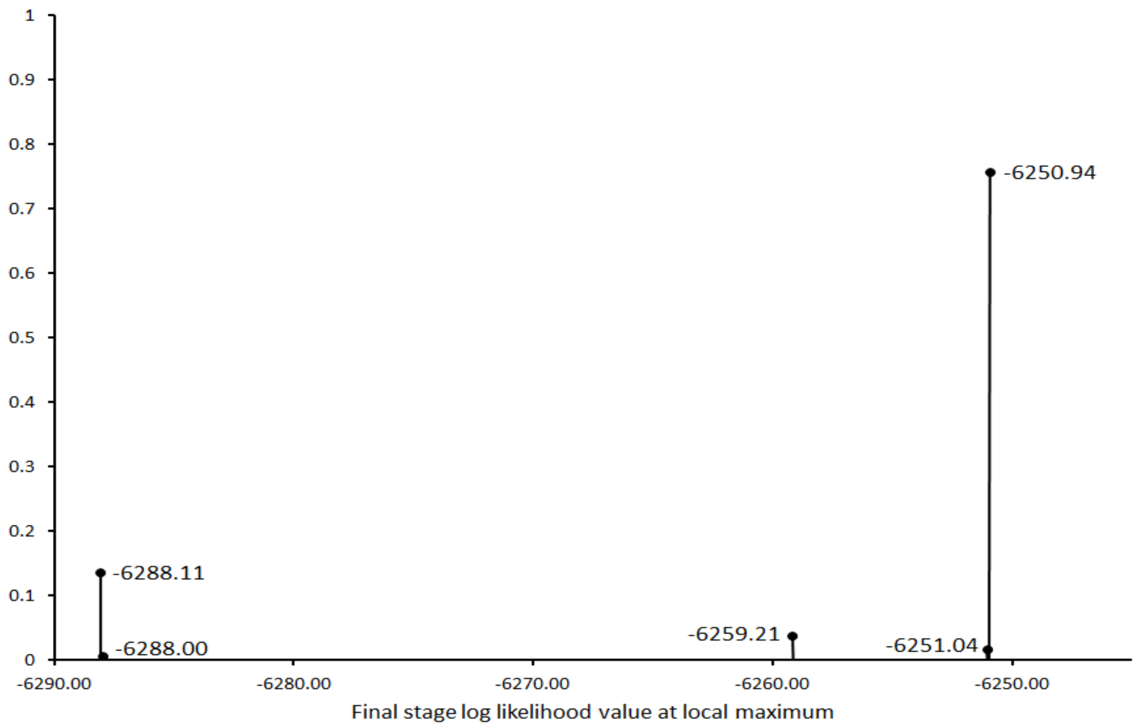

UC SANTA BARBARA

Rerunning with more random starts

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

           -6250.939  650371           14
           -6250.939  27071            15
           -6250.939  unperturbed      0
           -6250.939  127215           9
```

- If the unperturbed set of start values is not one of the sets to get to the best LL value, don't use it again as the perturbation point of origin.
- ALWAYS ask for `OUTPUT: svalues;` you can copy-and-paste these into a new input file so that the start value random perturbations are done starting at the best solution from the previous run.



Note: LL replication is *neither* necessary or sufficient for a given solution to be the global maximum.

I spy with my little eye…

- The number and proportion of sets of random starting values that converge to proper solution (as failure to consistently converge can indicate weak identification);
- The number and proportion of replicated maximum likelihood values for each local and the apparent global solution (as a high frequency of replication of the apparent global solution across the sets of random starting values increases confidence that the "best" solution found is the true maximum likelihood solution);
- The condition number. It is computed as the ratio of the smallest to largest eigenvalue of the information matrix estimate based on the maximum likelihood solution. A low condition number, less than 10^{-6}, may indicate singularity (or near singularity) of the information matrix and, hence, model non-identification (or empirical underidentification)
- The smallest estimated class proportion and estimated class size among all the latent classes estimated in the model (as a class proportion near zero can be a sign of class collapsing and class over-extraction).

Model Estimation Tracking

n=				# response patterns =								
1338				32		Final starting value sets converging		LL replication		Smallest class		
Model	**K=**	**Best LL**	**npar**	**Starts =**		**f**	**%**	**f**	**%**	**f**	**%**	**Condition Number**
1-class	1	-4102.59	5	100	50	50	100%	50	100%	1338	100%	2.00E-01
2-class	2	-3647.76	11	100	50	50	100%	50	100%	424	32%	5.26E-02
3-class	3	-3546.64	17	100	50	50	100%	43	86%	236	18%	7.09E-04
4-class	4	-3504.30	23	100	50	50	100%	45	90%	189	14%	1.40E-04
5-class	5	-3498.43	29	100	50	48	96%	15	31%	136	10%	6.45E-04
5-class	5	-3498.43	29	500	100	100	100%	92	92%	136	10%	5.41E-04
6-class	6	N/A										

IMMERSE Project – IES funded Training Grant (R305B220021)

MplusAutomation Caution

- You need to study the output to know if the LL is replicated.
- MplusAutomation and the code we provide will not tell you when a likelihood has only been replicated once.

You must open the Mplus Output file!!

IMMERSE Project – IES funded Training Grant (R305B220021)

LAB 2: Model Estimation Tracking and Trouble Shooting

INSTRUCTIONS: Get materials from Lab 2 on the In-person Github website

Latent Class Enumeration

UC SANTA BARBARA

Now…the hard part…

- In the majority of applications of mixture modeling, the number of classes is not known.
- Even in “direct applications”, when one assumes a priori that the population is heterogeneous, you rarely have specific hypotheses regarding the exact number or nature of the subpopulations.
- Thus, in either case (direct or indirect), you must begin with the model building with an exploratory class enumeration step.

IMMERSE Project – IES funded Training Grant (R305B220021)

Now…the hard part…

- Deciding on the number of classes is often the most arduous phase of the mixture modeling process.
- It is labor intensive because it requires consideration (and, therefore, estimation) of a set of models with a varying numbers of classes
- It is complicated in that the selection of a “final” model from the set of models under consideration requires the examination of a host of fit indices along with substantive scrutiny and practical reflection, as there is no single method for comparing models with differing numbers of latent classes that is widely accepted as best.

Evaluating the Model

The **statistical tools** are divided into three categories:

- evaluations of absolute fit;
- evaluations of relative fit;
- evaluations of classification.

+

Model Usefulness

- Substantive meaningful and substantively distinct classes (face + content validity)
- Cross-validation in second sample (or split sample)
- Parsimony principle
- Nomological network

IMMERSE Project – IES funded Training Grant (R305B220021)

Class Enumeration for LCA

- Fit models for K=1, 2, 3,… increasing K until the models become not identified or empirically not well-identified.
- Collect fit information on each model using a combination of statistical tools
- Decide on 2-3 “plausible” models
- Reapply statistical tools to set of candidate models and evaluate the model usefulness.

IMMERSE Project – IES funded Training Grant (R305B220021)

Example Table of Model Fit

Table 2. Model Fit Indices for the LCA Model With 1 Through 5 Latent Classes.

Model (*k*-class)	*LL*	npar	CAIC	BIC	aBIC	AWE	*VLMR-LRT p value*	*BLRT p value*	*BF*	cmP_k
1-class	−1,294.0	5	2,623.55	2,618.55	2,602.68	**2,664.05**				0.0
2-class	−1,259.2	11	**2,596.58**	**2,585.58**	2,550.67	2,685.69	**<.001**	<.001	**>10**	**1.0**
3-class	−1,249.2	17	2,619.01	2,602.01	**2,548.06**	2,756.72	.125	**.004**	>10	0.0
4-class	−1,244.3	23	2,651.84	2,628.84	2,555.85	2,838.14	.012	.263	>10	0.0
5-class	−1,242.4	29	2,690.63	2,661.63	2,569.60	2,925.54	.352	1.000	>10	0.0

Note. Bold values indicate the model that the fit criteria endorse. *K* = number of classes; *LL* = log-likelihood; npar = Parameters, CAIC = Consistent Akaike Information Criterion; BIC = Bayesian Information Criterion; aBIC = adjusted BIC; AWE = Approximate Weight of Evidence Criterion; BLRT = bootstrapped likelihood ratio test; VLMR-LRT = Vuong–Lo–Mendell–Rubin adjusted likelihood ratio test; *p* = *p* value; *BF* = Bayes Factor; *cmP* = correct model probability. Bolded values indicate "best" fit for each respective statistic.

Tests of Model Fit

Absolute fit

- There is an overall likelihood ratio model chi-square goodness-of-fit for mixture measurement model with only categorical indicators (using similar formula to the goodness-of-fit chi-square for contingency table analyses and log linear models).

$$X^2_{LR} = 2\sum_{r=1}^{R}\left[f_r \log\left(\frac{f_r}{\hat{f}_r} \right) \right],$$

$$df_{X^2_{LR}} = R - d - 1$$

Caveats re: Chi-square Goodness-of-fit

- The sampling distribution of this statistic isn't chi-square when expected cells counts get small (<1.0 or < 5.0)
- The chi-square goodness-of-fit is also known to be sensitive to what would be considered negligible or inconsequential misfit in very large samples. In these cases, the null hypothesis may be rejected and the model determined to be statistically inadequate but, upon closer practical inspection, may be ruled to have a "close enough" fit.
- With increasing number of items, there is the problem of sparse cells.
- Thus– we do not rely on the chi-square for mixture models

Residuals

- "Inspection" = Look at standardized residuals evaluating difference between the observed response pattern frequencies the model-estimated frequencies (TECH10).

$$stdr\hat{e}s_r = \frac{f_r - \hat{f}_r}{\sqrt{\hat{f}_r\left(1 - \frac{\hat{f}_r}{n}\right)}}.$$

- Examine the overall standardized univariate and bivariate residuals (also part of TECH10) along with the class-specific univariate and bivariate residuals (produced by Output: Residual). Bivariate residuals can help you evaluate potential violations of the conditional independence assumption.

Residuals

- The values of the standardized residuals can be compared to a standard normal distribution, with large values **(e.g.>3 or <-3)** indicating response patterns that are more poorly fit, contributing the most to the and the rejection of the model.
- Since the number of possible response patterns can become large very quickly with increasing numbers of indicators and/or response categories per indicator, focus your attention on the residuals of these patterns where the bulk of data reside and complete data response patterns.
- Consider the overall proportion of response patterns with large standardized residuals. For a well-fitting model, one would still expect, by chance, to have some small percentage of the response patterns to have significant residual values.
  - We would likely only take proportions in notable excess of, say, **3-5%**, to be an indication of a poor-fitting model.

RESPONSE PATTERN FREQUENCIES AND CHI-SQUARE CONTRIBUTIONS

Response Pattern	Frequency Observed	Frequency Estimated	Standardized Residual (z-score)	Chi-square Contribution Pearson	Chi-square Contribution Loglikelihood	Deleted
1	4.00	2.15	1.27	1.60	4.98	
2	7.00	5.09	0.85	0.71	4.45	
3	889.00	888.61	0.02	0.00	0.78	
4	3.00	4.77	-0.81	0.66	-2.78	
5	271.00	271.47	-0.03	0.00	-0.94	
6	1.00	0.20	1.76	3.11	3.18	
7	5.00	4.01	0.49	0.24	2.20	
8	3.00	1.02	1.96	3.84	6.47	
9	4.00	2.75	0.76	0.57	3.00	
10	2.00	1.85	0.11	0.01	0.31	
11	6.00	5.74	0.11	0.01	0.54	
12	3.00	2.28	0.47	0.22	1.63	
13	55.00	54.69	0.04	0.00	0.62	
14	3.00	0.85	2.33	5.45	7.57	
15	1.00	0.70	0.36	0.13	0.72	
16	8.00	6.81	0.46	0.21	2.57	
17	9.00	9.59	-0.19	0.04	-1.15	
18	1.00	1.08	-0.08	0.01	-0.15	
19	12.00	12.70	-0.20	0.04	-1.35	
20	6.00	6.51	-0.20	0.04	-0.98	
21	4.00	3.02	0.56	0.32	2.24	
22	21.00	17.23	0.91	0.82	8.30	
23	3.00	2.75	0.15	0.02	0.51	
24	2.00	0.66	1.64	2.70	4.42	
25	1.00	2.36	-0.89	0.78	-1.72	
26	42.00	44.56	-0.39	0.15	-4.98	
27	3.00	4.02	-0.51	0.26	-1.75	
28	8.00	4.94	1.38	1.89	7.70	
29	19.00	18.93	0.02	0.00	0.13	
30	3.00	0.93	2.14	4.57	7.00	
31	6.00	2.75	1.96	3.85	9.37	
32	3.00	3.05	-0.03	0.00	-0.09	

IMMERSE Project – IES funded Training Grant (R305B220021)

UNIVARIATE MODEL FIT INFORMATION

Variable	Estimated Probabilities H1	H0	Standardized Residual (z-score)
PYDI1AB			
Category 1	0.357	0.357	0.000
Category 2	0.643	0.643	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
Number of Significant Standardized Residuals			0
PYDI2AB			
Category 1	0.071	0.071	0.000
Category 2	0.929	0.929	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
Number of Significant Standardized Residuals			0
PYDI3AB			
Category 1	0.130	0.130	0.000
Category 2	0.870	0.870	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
Number of Significant Standardized Residuals			0
PYDI4AB			
Category 1	0.067	0.067	0.000
Category 2	0.933	0.933	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
Number of Significant Standardized Residuals			0
PYDI5AB			
Category 1	0.134	0.134	0.000
Category 2	0.866	0.866	0.000
Univariate Pearson Chi-Square			0.000
Univariate Log-Likelihood Chi-Square			0.000
Number of Significant Standardized Residuals			0
PYDI6AB			
Category 1	0.050	0.050	0.000
Category 2	0.950	0.950	0.000
Univariate Pearson Chi-Square			0.000

Overall Univariate Pearson Chi-Square	0.000
Overall Univariate Log-Likelihood Chi-Square	0.000
Overall Number of Significant Standardized Residuals	0

```
BIVARIATE MODEL FIT INFORMATION

                                             Estimated Probabilities
                                                                   Standardized
Variable        Variable                      H1           H0         Residual
                                                                     (z-score)
PYDI1AB         PYDI2AB
  Category 1      Category 1                0.050        0.049          0.225
  Category 1      Category 2                0.307        0.308         -0.105
  Category 2      Category 1                0.021        0.023         -0.324
  Category 2      Category 2                0.622        0.621          0.100
  Bivariate Pearson Chi-Square                                          0.162
  Bivariate Log-Likelihood Chi-Square                                   0.164
  Number of Significant Standardized Residuals                          0
PYDI1AB         PYDI3AB
  Category 1      Category 1                0.087        0.084          0.341
  Category 1      Category 2                0.270        0.272         -0.212
  Category 2      Category 1                0.044        0.046         -0.452
  Category 2      Category 2                0.600        0.597          0.193
  Bivariate Pearson Chi-Square                                          0.349
  Bivariate Log-Likelihood Chi-Square                                   0.351
  Number of Significant Standardized Residuals                          0
PYDI1AB         PYDI4AB
  Category 1      Category 1                0.053        0.049          0.626
  Category 1      Category 2                0.304        0.307         -0.294
  Category 2      Category 1                0.014        0.017         -1.036
  Category 2      Category 2                0.629        0.626          0.280
  Bivariate Pearson Chi-Square                                          1.516
  Bivariate Log-Likelihood Chi-Square                                   1.583
  Number of Significant Standardized Residuals                          0
PYDI1AB         PYDI5AB
  Category 1      Category 1                0.103        0.102          0.074
  Category 1      Category 2                0.254        0.255         -0.051
  Category 2      Category 1                0.032        0.032         -0.127
  Category 2      Category 2                0.611        0.611          0.046
  Bivariate Pearson Chi-Square                                          0.023
  Bivariate Log-Likelihood Chi-Square                                   0.023
  Number of Significant Standardized Residuals                          0
PYDI1AB         PYDI6AB
  Category 1      Category 1                0.036        0.036          0.076

Overall Bivariate Pearson Chi-Square                                    6.599
Overall Bivariate Log-Likelihood Chi-Square                             6.692
Overall Number of Significant Standardized Residuals                    0
```

What assumption can we evaluate by looking at the bivariate residuals?

Relative Fit: Comparing models with (k-1) vs. k classes.

The most common ML-based inferential comparison is the likelihood ratio test (LRT) for nested models

Hypothesis testing using the likelihood ratio

H0: k-1 classes

H1: k classes

LRTS = -2 [log L(H0) - log L(H1)], ***df*** = npar(H1) – npar(H0)

When testing a (k-1)-class mixture model versus a k-class model, the LRTS does *not* have an asymptotic chi-squared distribution.

Why?

Regularity conditions are not met: Mixing proportions are on the boundary of the parameter space and the parameters under the null model are not identifiable.

Nylund, K., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: An Interdisciplinary Journal*, *14*(4), 535-569.

The Sampling Distribution for the LRTS

- Vuong-Lo-Mendel-Rubin Test (TECH11)
  - Analytically-derived distribution to test the difference between the *k*- and *k-1*- class models.
  - Provides *p*-value to indicate if there is statistically significant improvement in fit.

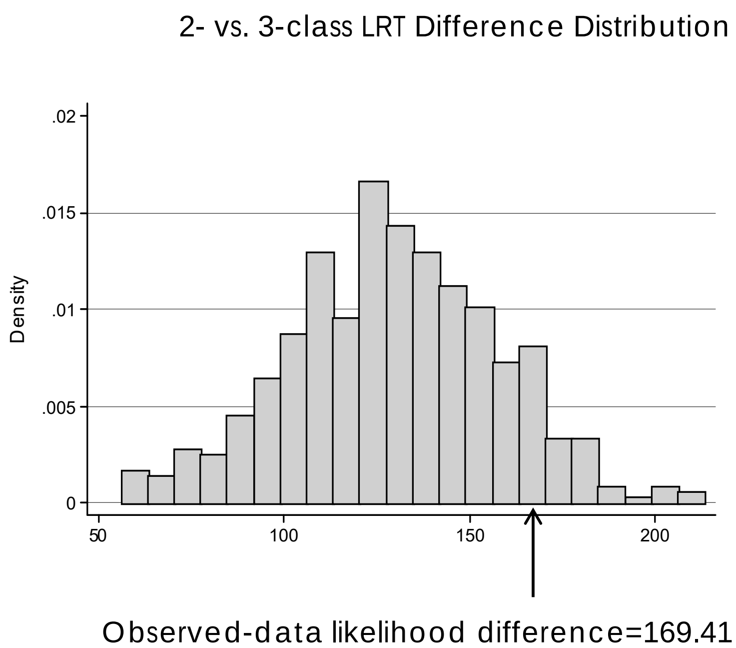
- Bootstrapping the Likelihood Ratio Difference (TECH14)
  - Empirically-derived distribution to test the difference between the *k*- and *k-1*- class models.
  - Provides a *p*-value to indicate if there is statistically significant improvement in fit.

NOTE: For both Tech11 and Tech14, Mplus computes the LRTS for your *K*-class model compared to a model with one less class (i.e., K-1 class model as the Null). The LRTS values and *df* should be the same for Tech11 and Tech14 and should match your own calculation outside of Mplus.

Nylund, K., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: An Interdisciplinary Journal*, *14*(4), 535-569.

Estimating the BLRT

- Calculate the likelihood difference between the *k*- and *k-1*-class models for the observed data.
- Randomly generate 500 replications with the *k-1* class parameters
- Estimate the *k*-and *k-1* class models and calculate the likelihood difference for each replication.
- Use the distribution to calculate the p-value of the calculated difference.



Tech14 Strategy

- TECH14 computations are time consuming b/c for each bootstrap draw random starts are need for the k-class model. Here's how to save time:
  1. Run models *without* Tech14.
  2. Run with Tech14 using the stable solution from Step 1 as start values (svalues option).
  3. Run again w/ LRTSTARTS = 0 0 100 20 to see if results are sensitive to the number of random starts for the k-class model.

LRT Difference in practice

You request the LMR and BLRT for models with *K*>1 classes.

- Ho: *k*-1 - model
- H1: *k*-class model

For a given run with k- classes, if you get a significant p-value, you can conclude that the addition of the additional class improves fit.

You can "stop" adding classes when you receive a nonsignificant *p*-value. The null is *k*-1, so you would conclude at that point the additional class does not significantly improve fit.

Classes	LMR	BLRT
1	--	--
2	0.0	0.0
3	0.0	0.0
4	0.0	0.0
5	0.0	0.0
6	0.0	0.0
7	0.0	0.0
8	0.2	0.0

In this example, the LMR indicates that you do not improve fit after 7 classes; the BLRT does not contribute information to fit.

Model Fit Indices

Information-Heuristic Criteria

- Fit indices that weigh the fit of the model (log likelihood value) in consideration of the model complexity.
- These information criteria can be expressed in the following form:

$$-2LL + penalty$$

- Traditional penalty is a function of *n* and *d*
  - *n*= sample size
  - *d*= number of parameters

Information Criteria

- Bayesian Information Criterion $$BIC = -2LL + d\log(n)$$
- Sample-size adjusted BIC $$SABIC = -2LL + d\log\left(\frac{n+2}{24}\right)$$
- Consistent Akaike's Information Criterion $$CAIC = -2LL + d\left[\log(n) + 1\right]$$
- Approximate Weight of Evidence Criterion $$AWE = -2LL + 2d\left[\log(n) + 1.5\right]$$

Plotting ICs

- For these ICs, **lower values** indicate a better model, relatively-speaking. Based on simulation studies[1,2] these are our current recommendations:
  - BIC seems to work best across range of conditions.
  - ABIC is consistent too, not as much as BIC.
  - AWE has show to sometimes under-extract (so may provide lower-limit on # of classes).
  - AIC tends to over-extract (so may provide upper-limit on # of classes)
- Sometimes, a minimum value is not reached and scree/"elbow" plots are utilized (scree may be easier to see plotting from k=2 instead of k=1).
- Look for the point of diminishing returns for each additional class.
  - E.g., by adding an additional class, we are seeing minimal increase in the BIC or AWE

If we fit k = 1-6 class models (k=7 is not identified) and k=6 has the lowest BIC, is k=6 the minimum?

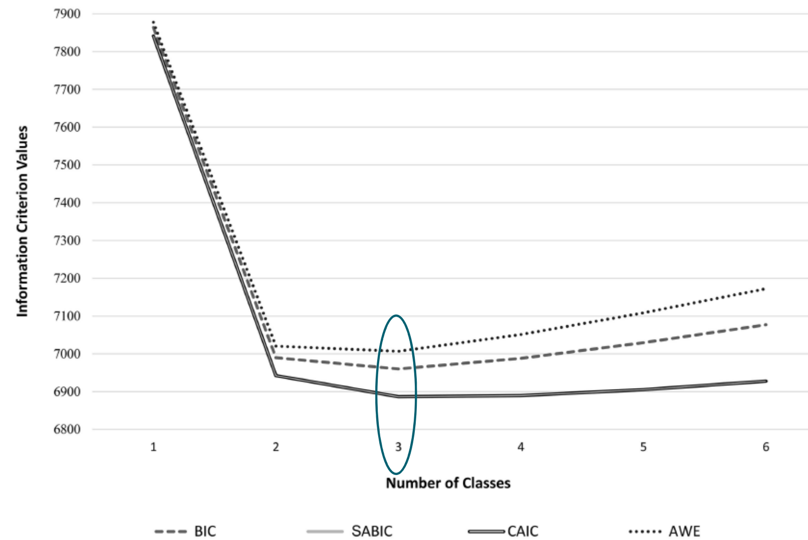


Figure 2. Plot of information criterion values: Positive Youth Development – Contribution Subscale latent class analysis models. The SABIC and CAIC lines are overlapping.

1: Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal*, *14*(4), 535-569.
2: Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 782-797.
3: Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, *4*(4), 440.

Relative Improvement

$$RI_{K,K+1} = \frac{LL_{K+1} - LL_K}{LL_2 - LL_1}$$

- Yields a number between 0 and 1, where a small value favors the *K*-class model and a larger value favors the (*K*+1)-class model.
- Plot the *RI* values to look for an "elbow" or scree.

IMMERSE Project – IES funded Training Grant (R305B220021)

Bayes Factor (Maysn, 2013)

How much lower does an IC value have to be to mean the model is really better?

- Bayes Factor: Which model, A or B, is more likely to be the true model if one of the two is the true model?

$$\hat{BF}_{A,B} = \exp\left[SIC_A - SIC_B\right] \approx \frac{\Pr(\text{Model A correct})}{\Pr(\text{Model B correct})}$$

$$SIC = -0.5BIC$$

Evidence for Model A: weak (1-3); moderate (3-10); strong (>10)

Correct Model Probability
(Maysn, 2013)

- The approximate correct model probability (cmP) for a Model *j* is an approximation of the actual probability of Model *j* being the correct model relative to a set of *J* models under consideration .

$$cm\hat{P}_A = \frac{\exp\left(SIC_A - SIC_{\max}\right)}{\sum_j \exp\left(SIC_j - SIC_{\max}\right)},$$

- Pick the model with the **largest cmP** value.

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, *4*(4), 440.

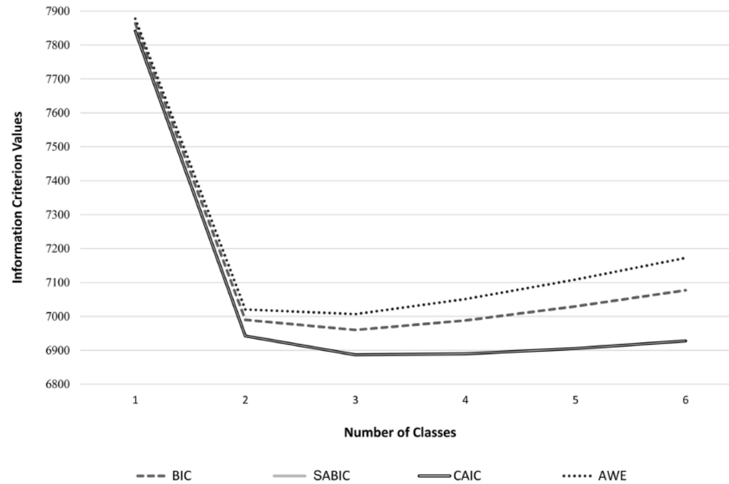


Figure 2. Plot of information criterion values: Positive Youth Development – Contribution Subscale latent class analysis models. The SABIC and CAIC lines are overlapping.

Table 3
Fit Statistics and Classification Coefficients: Positive Youth Development Inventory – Contribution Subscale Latent Class Analysis Models

K	LL	BIC	SABIC	CAIC	AWE	BLRT *p*	VLMR-LRT *p*	Entropy	*BF*	*cmP*
1	−3905.892	7863.555	7841.317	7841.267	7877.751	—	—	—	.000	.000
2	−3439.483	6989.902	6942.250	6942.145	7020.324	<.001	<.001	.790	.000	.000
3	−3394.950	**6960.001**	**6886.934**	**6886.774**	**7006.648**	<.001	<.001	.718	**>15.000**	**1.000**
4	−3379.436	6988.140	6889.658	6889.442	7051.011	**<.001**	**.002**	.753	**>15.000**	.000
5	−3370.534	7029.501	6905.604	6905.333	7108.598	.200	.429	.782	**>15.000**	.000
6	**−3364.745**	7077.089	6927.778	6927.450	7172.411	.500	.472	**.803**	—	.000

Note. K = number of classes; LL = log-likelihood; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC; CAIC = consistent Akaike information criterion; AWE = approximate weight of evidence criterion; BLRT = bootstrapped likelihood ratio test; VLMR-LRT = Vuong-Lo-Mendell-Rubin adjusted likelihood ratio test; *p* = *p* value; *BF* = Bayes factor; *cmP* = correct model probability. Bolded values indicate "best" fit for each respective statistic. Entropy is included in the table for brevity but should not be used as a model selection statistic (Masyn, 2013).

Classification Precision

Classification quality/Class separation

- A "good" (read: useful) mixture model in a direct application should yield empirically, highly-differentiated, well-separated latent classes whose members have a high degree of homogeneity in their responses on the class indicators.
- Most all classification diagnostics are based on estimated posterior probabilities.
  - Using Bayes Theorem, we can estimate their posterior probabilities, then assign them to class based on where their largest posterior probability values is. This is referred to as "modal class assignment".

Modal Class Assignment

	Posterior class probabilities					Modal class
Student	C1	C2	C3	C4	C5	C
1	0.99	0.1	0	0	0	1
2	0.49	0.25	0.01	0.25	0	1
3	0.35	.333	0.33	0	0	1
4	0.26	0.24	0.25	0.25	0	1

IMMERSE Project – IES funded Training Grant (R305B220021)

Posterior Class Probabilities

- The LCA model provides information about how well people are classified into each of the latent classes.
  - Relative Entropy is a single measure of this
  - Modal class assignments.

Average class prob for C1 | assigned to C1

Average class prob for C1 | assigned to C2

	1	2	3	4	5
1	0.842	0.001	0.039	0.064	0.054
2	0.001	0.934	0.061	0.005	0.001
3	0.029	0.068	0.822	0.081	0.001
4	0.051	0.006	0.106	0.817	0.02
5	0.067	0.001	0.001	0.02	0.913

  - Values down the diagonal are the average posterior class probabilistic (AvePP; Masyn, 2013)
  - Values greater than .8 are considered good (Rost, 2006)

Classification table

LSAY Example: Model Classification Diagnostics for the 5-class Unconditional LCA ($E_5 = .77$)

for Subsample A (n_A=1338)

Class k	$\hat{\pi}_k$	95% C.I.*	$mcaP_k$	$AvePP_k$	OCC_k
Class 1	.392	(.326, .470)	.400	.905	14.78
Class 2	.130	(.082, .194)	.125	.874	46.42
Class 3	.182	(.098, .255)	.176	.791	17.01
Class 4	.190	(.139, .248)	.189	.833	21.26
Class 5	.105	(.080, .136)	.109	.874	59.13

*Bias-corrected bootstrap 95% confidence intervals

Code to get CIs for class proportions

```
Analysis:
  estimator = ml;
  type=mixture;
  starts=100 50;
  processors = 4;
  bootstrap=5000;

%OVERALL%

      [ c#1*0.21591 ] (c1i);
      [ c#2*1.32069 ] (c2i);
      [ c#3*0.59664 ] (c3i);
      [ c#4*0.55474 ] (c4i);

Model Constraint:
New (cp1 cp2 cp3 cp4 cp5);
cp1 = exp(c1i)/(exp(c1i)+exp(c2i)+exp(c3i)+exp(c4i)+1);
cp2 = exp(c2i)/(exp(c1i)+exp(c2i)+exp(c3i)+exp(c4i)+1);
cp3 = exp(c3i)/(exp(c1i)+exp(c2i)+exp(c3i)+exp(c4i)+1);
cp4 = exp(c4i)/(exp(c1i)+exp(c2i)+exp(c3i)+exp(c4i)+1);
cp5 = 1/(exp(c1i)+exp(c2i)+exp(c3i)+exp(c4i)+1);

output: cinterval(bcbootstrap);
```

mcaP

- Modal class assignment proportion (mcaP) is the proportion of individuals in the sample modally-assigned to Class k.
- If individuals were assigned to Class k with perfect certainty, then mcaP(k) would be equal to the model-estimated Pr(c=k). Larger discrepancies are indicative of larger latent class assignment errors.
- To gauge the discrepancy, each mcaP can be compared to the to 95% confidence interval for the corresponding model-estimated Pr(c=k).

IMMERSE Project – IES funded Training Grant (R305B220021)

Odds of Correct Classification (OCC)

$$OCC_k = \frac{AvePP_k \big/ \left(1 - AvePP_k\right)}{\hat{\pi}_k \big/ \left(1 - \hat{\pi}_k\right)},$$

- The denominator of the odds of correct classification (OCC) is the odds of correct classification based on random assignment using the model-estimated marginal class proportions.
- The numerator is the odds of correct classification based on the maximum posterior class probability assignment rule (i.e., modal class assignment).
- When the modal class assignment for Class k is no better than chance, then OCC(k)=0.
- As AvePP(k) gets close to one, OCC(k) gets large.
- Nagin suggests OCC(k)>5 indicate adequate separation and classification precision.

IMMERSE Project – IES funded Training Grant (R305B220021)

LSAY Example: Model Classification Diagnostics for the 5-class Unconditional LCA ($E_5 = .77$)

for Subsample A (n_A=1338)

Class k	$\hat{\pi}_k$	95% C.I.*	$mcaP_k$	$AvePP_k$	OCC_k
Class 1	.392	(.326, .470)	.400	.905	14.78
Class 2	.130	(.082, .194)	.125	.874	46.42
Class 3	.182	(.098, .255)	.176	.791	17.01
Class 4	.190	(.139, .248)	.189	.833	21.26
Class 5	.105	(.080, .136)	.109	.874	59.13

*Bias-corrected bootstrap 95% confidence intervals

95% confidence interval of the

Large discrepancy between this value and the π indicate classification error. Ideally, mcaP will fall within the 95% CI of π

Value of 1 would indicate perfect assignment. Values >.7 considered "well separated" (Nagin, 2005)

Values >5 indicate that chances of being in this class are not just chance. Odds >5 indicate good class separation and classification accuracy

A few thoughts…

- We recommend reporting relative Entropy when describing the chosen model, not in model fit table (because Entropy is *not* a measure of model fit!). [see next slide]
- Paper excerpt:

  However, in this sample, the average latent class posterior probabilities for those who were modally assigned to the class, called the average posterior class probability (AvePP; Masyn, 2013) were high, all above the recommended value of .7 for good classification (Nagin, 2005). For example, for seventh grade, the AvePP for the four classes was 0.84, 0.79, 0.85, and 0.90 for classes 1-4 respectively.
- Not recommended to use model class assignment for subsequent analyses
  - There is a paper that suggests it might be OK to do if entropy is larger than .8 (Clark & Muthen, 2009); however, a high entropy value can mask a small class with poor classification precision.
- Other methods for comparing classification to other variables (e.g., covariates or distal outcomes)

Ram, N., Grimm, K.J. (2009). Growth mixture modeling: A method for identifying difference in longitudinal change among observed groups. *International Journal of Behavioral Development, 33(6)*: 565-576.

Clark, S., & Muthen, B. (2009). *Relating Latent Class Analysis results to variables not included in the analysis.* Retrieved from http://www.statmodel.com/download/relatinglca.pdf

IMMERSE Project – IES funded Training Grant (R305B220021)

Model Usefulness

Substantive model checking

“Some things have to be believed to be seen.”

- Ralph Hodgson

UC SANTA BARBARA

Model Usefulness

Working from 2-3 “best” models based on all the statistical criteria above, evaluate the model usefulness and construct validity of the latent class variable.

- Substantively meaningful and substantively distinct classes (face + content validity)
- Cross-validation in second sample (or split sample)
- Parsimony principle
- Nomological network (predictors, outcomes, etc.)

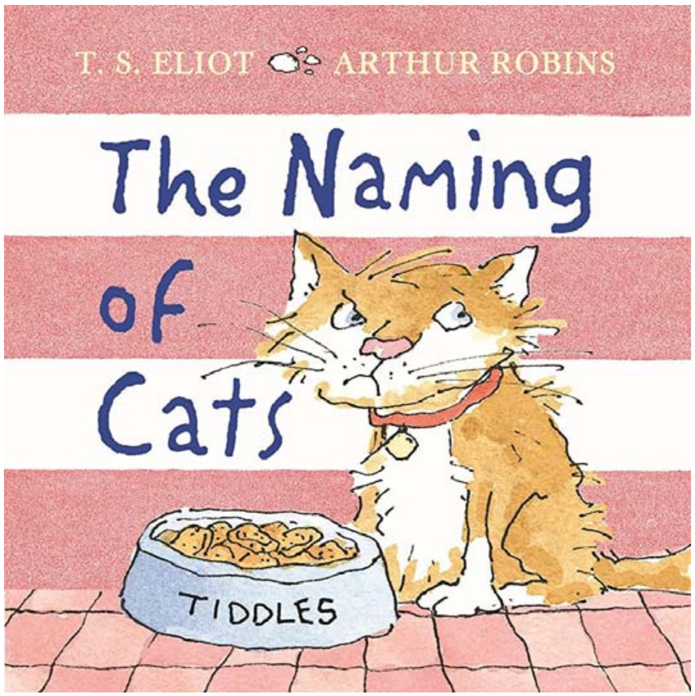
IMMERSE Project – IES funded Training Grant (R305B220021)

Substantive Meaning

- Substantive meaning and interpretability of resultant classes
  - Think about which items characterize each of the classes (class homogeneity)
  - Think about which items distinguish between which classes (class separation)
  - Think about the class proportions
  - Visualize the classes (e.g., profile plots)

IMMERSE Project – IES funded Training Grant (R305B220021)

Proceed with Caution…



- It is important to avoid two logical errors concerning names of the latent classes:
  - Just because a class is named something (e.g., "frequent peer victimization") does *not* mean that the underlying hypothetical construct is understood or even correctly labeled—to believe otherwise is the ***naming fallacy***
  - A hypothetical construct (e.g., "borderline personality") may not correspond to a real thing—to believe otherwise is the ***reification fallacy***
- A class name should be seen as a convenience, not as a substitute for critical thinking
  - **"Jingle-jangle" fallacy**
    - Jingle: Just because two variables are called the same thing doesn't mean they are
    - Jangle: Just because two variables are called something different doesn't mean they are

IMMERSE Project – IES funded Training Grant (R305B220021)

Substantive Model Comparisons

- Ask yourself whether the resultant latent classes of one model help you understand the phenomenon of interest better than another.
- Weigh the simplicity and clarity of each of the candidate models.
- Evaluate the utility of additional classes for the less parsimonious of the candidate models.
- Compare the modal class assignments of individuals across the candidate models (being mindful of label-switching)

IMMERSE Project – IES funded Training Grant (R305B220021)

Cross-validation

Within sample, e.g., split sample, *k*-fold, etc. or out-of-sample

- Do the number *and* nature of the resultant latent classes replicate across samples?
  - What criteria can we use for latent class replication?

Note: Keep an eye out for new developments in resampling for mixture model testing and validation.

Nomological Network

- In order to provide evidence of the construct validity of your latent class variable, it should relate to other observable and latent variables in a lawful way based on your theoretical model.
  - This means including your latent class variables in a larger variable framework with predictors, distal outcomes, etc.
  - That's up next!

LAB: Class Enumeration Fit and Classification Tables

INSTRUCTIONS:

- Stretch break
- Files are on GitHub