Cohort 1 MM4DBER Parking Lot

This google doc is a space for you to add your questions. Add your questions to the top of this document. We will check this google doc regularly and respond to questions. If you want an immediate response, please send us an email directly (education-mm4dbers@ucsb.edu)

I know we talked about this a bit ago, but I've got a small question about the 'here' function. Will the 'here' function search for the named folder anywhere on my computer or does it only search for folders that are associated with an Rproj file? Our early discussion on folder organization is seeming even more important now that I'm beginning to understand how I'll be using R projects and the 'here' function.

> Great question! The goal of R-projects and the here() function tools are to have all project related files housed within a single project folder (i.e., your "topmost project folder"). The here() function defaults to the starting file path being your topmost project folder (wherever this is located on your computer; try just running the code "here()" in your script to print this location). Within your project folder we want to designate where separate files should be organized. For example data should go in your "data" subfolder, figures in your "figures" subfolder and so on. For a complex analysis like LTA I would separate Mplus model files further using the following folder structure: "LCA_T1" (time 1 LCA enumeration files), "LCA_T2", and "LTA" (e.g., invariance & non-invariance LTA model files).

I'm anticipating using course grade as one of my distal outcomes. For a couple of reasons, I'd like to stay away from/unable to use a numerical value to represent grade and I had planned on using a categorical variable. What does that look like? Adam mentioned that I could go over this in an office hour (and I plan to do that), but I just wanted to drop this question here in case anyone else had similar plans.

> The short answer is that you can treat grades as either a continuous variable or categorical. The most common practice is to calculate the GPA which is generally how grades are modeled when you are averaging grades for each student across multiple semesters &/or classes. If you are using it as the distal

outcome grades from only a single class & time point then you may consider modeling grades as categorical (with more classes/ time points things get complicated quickly, multilevel modeling would be required). This comes with some challenges or decisions. Such as how many levels of grades will you model: Will grades be modeled using a 5-point scale? I would need to learn more about your data to understand what makes the most sense in your specific research context.

I'm trying to reconcile my previous modeling experience with what we learned in Training Day 3. Is CAIC different from AICC (sample size-adjusted AIC)? If so, how?

Yes, the CAIC, the AIC, and the aBIC are all information criteria which differ slightly in the way the penalty term is calculated. We typically do not recommend reporting the sample size adjusted AIC but do report the sample size adjusted BIC (aBIC). For details on each of these equations you could email us for more specific detail or the equations can also be readily found online as the ICs are widely used in statistics (e.g., search; "equation BIC").

Are the results in the following tables affected by the number of categories within each indicator? We have been discussing dichotomous type indicators, and I wonder if I use indicators with three options (0, 1, 2) if the calculations for the statistics below should change.

Model Fit Summary Table[1]

Classes	Par	*LL*	BIC	aBIC	CAIC	AWE	BLRT	VLMR	BF	*cmPk*
1-Class	6	−5,443.41	10,932.50	10,913.44	10,938.50	10,996.19	–	–	0.00	<.001
2-Class	13	−5,194.14	10,487.26	10,445.96	10,500.26	10,625.24	**<.001**	**<.001**	0.00	<.001
3-Class	20	−5,122.48	**10,397.24**	**10,333.70**	**10,417.24**	**10,609.53**	**<.001**	**<.001**	**>100**	**1.00**
4-Class	27	−5,111.76	10,429.10	10,343.32	10,456.10	10,715.69	0.03	0.01	**>100**	<.001
5-Class	34	−5,105.59	10,470.07	10,362.04	10,504.06	10,830.95	0.67	0.18	**>100**	<.001
6-Class	41	−5,100.83	10,513.84	10,383.58	10,554.84	10,949.03	0.60	0.56	–	<.001

[1] *Note.* Par = Parameters; *LL* = model log likelihood; BIC = Bayesian information criterion; aBIC = sample size adjusted BIC; CAIC = consistent Akaike information criterion; AWE = approximate weight of evidence criterion; BLRT = bootstrapped likelihood ratio test p-value; VLMR = Vuong-Lo-Mendell-Rubin adjusted likelihood ratio test p-value; *cmPk* = approximate correct model probability.

Model Classification Diagnostics for the 3-Class Solution						
k-Class	*k*-Class Proportions	95% CI	$McaP_k$	$AvePP_k$	OCC_k	Entropy
Class 1	0.249	[0.166, 0.329]	0.282	0.675	6.264	0.635
Class 2	0.106	[0.083, 0.136]	0.095	0.904	79.420	
Class 3	0.644	[0.561, 0.731]	0.623	0.893	4.614	

Note. $McaP_k$ = Modal class assignment proportion; $AvePP_k$ = Average posterior class probabilities; OCC_k = Odds of correct classification;

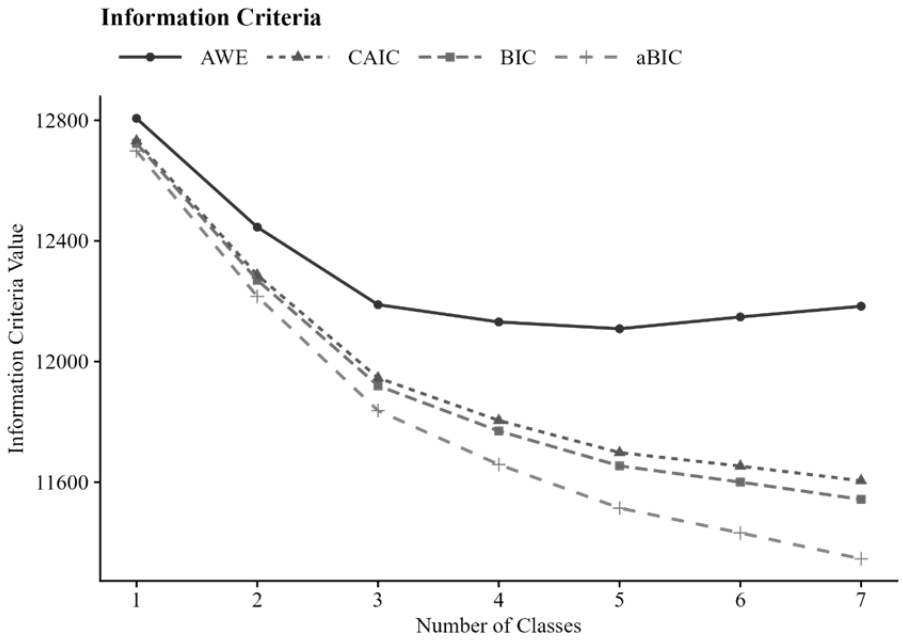
> Great question! The equation for calculating the fit indices will remain the same but the resulting values for each indices would be different given that the data has changed. So we would use the same approach but the numbers found in the table will depend on the model and data provided in your analysis.

I have been testing the enumeration procedure with datasets. I am testing to see if I can run the code on my data and understand the fit interpretation. For one dataset, I got the following table and figure. Based on these results (assuming the model estimation ended normally for the runs), would you suggest increasing the number of classes, or should I consider one of the classes within the table?

Model Fit Summary Table[1]

Classes	Par	*LL*	BIC	aBIC	CAIC	AWE	BLRT	VLMR	BF	*cmPk*
1-Class	8	−6,332.50	12,723.64	12,698.23	12,731.64	12,806.27	–	–	0.00	<.001
2-Class	17	−6,072.77	12,270.13	12,216.12	12,287.13	12,445.72	<.001	<.001	0.00	<.001
3-Class	26	−5,864.72	11,920.00	11,837.40	11,946.00	12,188.56	<.001	<.001	0.00	<.001
4-Class	35	−5,756.67	11,769.85	11,658.67	11,804.85	12,131.37	<.001	<.001	0.00	<.001
5-Class	44	−5,665.90	11,654.28	11,514.50	11,698.28	**12,108.76**	<.001	<.001	0.00	<.001
6-Class	53	−5,606.02	11,600.47	11,432.11	11,653.47	12,147.92	<.001	<.001	0.00	<.001
7-Class	62	−5,544.33	**11,543.07**	**11,346.11**	**11,605.07**	12,183.47	<.001	<.001	–	**1.00**

[1] *Note.* Par = Parameters; *LL* = model log likelihood; BIC = Bayesian information criterion; aBIC = sample size adjusted BIC; CAIC = consistent Akaike information criterion; AWE = approximate weight of evidence criterion; BLRT = bootstrapped likelihood ratio test p-value; VLMR = Vuong-Lo-Mendell-Rubin adjusted likelihood ratio test p-value; *cmPk* = approximate correct model probability.



Suggest you make an appointment to come to office hours to discuss your models more closely. Might need to check some diagnostics before we go into the interpretation.

Based on the table below, do you think there is enough evidence to select the 3-class model instead of the other models? I know the theoretical framework plays a role, but I wonder what your decision will be based on the results here.

Model Fit Summary Table[1]

Classes	Par	*LL*	BIC	aBIC	CAIC	AWE	BLRT	VLMR	BF	*cmPk*
1-Class	8	−1,031.96	2,108.04	2,082.68	2,116.04	2,176.14	–	–	0.00	<.001
2-Class	17	−947.90	1,989.52	1,935.63	2,006.52	**2,134.25**	<.001	<.001	0.00	0.00
3-Class	26	−917.05	**1,977.46**	1,895.04	**2,003.46**	2,198.81	<.001	<.001	4.76	**0.82**
4-Class	35	−893.80	1,980.58	1,869.63	2,015.58	2,278.55	<.001	<.001	**>100**	0.17
5-Class	44	−879.22	2,001.02	**1,861.54**	2,045.02	2,375.61	**<.001**	<.001	**>100**	<.001
6-Class	53	−871.63	2,035.48	1,867.47	2,088.48	2,486.69	1.00	0.02	–	<.001

[1] *Note.* Par = Parameters; *LL* = model log likelihood; BIC = Bayesian information criterion; aBIC = sample size adjusted BIC; CAIC = consistent Akaike information criterion; AWE = approximate weight of evidence criterion; BLRT = bootstrapped likelihood ratio test p-value; VLMR = Vuong-Lo-Mendell-Rubin adjusted likelihood ratio test p-value; *cmPk* = approximate correct model probability.

Suggest you make an appointment to come to office hours to help think beyond this summary table. The short-answer is that the enumeration decisions should be based both on substantive and empirical considerations.

I am interested in using items with more than two categories. Do you happen to have code for R to create the probability plots or related tables for models using items with more than two categories?

> Yes, I have a code example of a conditional probability plot for items with greater than two categories. I can provide you with this example but it does add a layer of complexity so it might also be helpful to come to office hours to discuss (email us and we can discuss further; Adam).

Is the Training Day 1 video link correct? I'm having trouble accessing the video from github?

> Could you please check again? We are able to view it from different computers on our end. Here is the link.
>
> - Got it, thanks! Not sure why I was having difficulty, but I've got the file saved.

In the article by Campbell et al. "From Comprehensive to Singular: A Latent Class Analysis of College Teaching Practices" from a few classes ago, they used a qualitative analysis (observation protocol of class styles) to do LCA. Can LCA be used with qualitative data? I have a qual data set where students each identified 1-2 constructs and I'm curious to see if certain demographics of students (e.g., women, Mexican-Americans) identified more strongly with certain constructs.

> Yes– they quantified qualitative observations. You certainly can use those in an LCA. Here are some other examples: here and here (coding student diagrams).

What is a good *n* for LCA?

> Million dollar question. There is not a one-size-fits-all recommendation but generally, 200 or more is recommended. There certainly have been published LCAs with n< 100 or so. However, with small samples, there is limited power to detect more than 2-3 classes if there really are some. So generally speaking, 200-300 would be the ideal lower bound of an ideal sample size.

Given that MPlus requires licensing, can you provide brief details about how that works? (price point, "how to", etc.) I know you set it up for this first round, but I'm thinking ahead to future years.

> We bought you Mplus for use this year. You will have a one-year active license. Once the year is over, your license will not be active. BUT– MPlus will continue to work. The only downside is that you won't be able to upgrade Mplus when

new versions come out. For a few years, this shouldn't be a problem. As time goes on, you may want to update your license and to do so you can go to the Mplus website and update your license (we have the "combo" package which allows you to do mixture modeling)

What drives the decision about which class is the reference class? Does it matter? Does this influence the results/interpretation of the multinomial regression?

It does not matter statistically. If you change the reference class the model fit is exactly the same. In an LCA context within Mplus, the reference class is the last class. Since the latent class variable is an unordered categorical variable, it is arbitrary which one is the reference class. HOWEVER, we may want to have one specific reference class when we interpret results. So we will show you how you can change the reference class.

For example, the 3rd class in a 4 class solution is the "high performance" class and you may want to compare all other classes to that one class. Since the default would be to have the fourth class as the reference class, not the 3rd, we want to change the reference class. The overall model fit would be exactly the same, we just reshuffle the classes. We will show you where you can do that– Mplus does it for you in most contexts. It presents the output for each class as the reference class.

I've never done any of this modeling before, and I'm still lost in the jargon. Can someone write general definitions to the terms below as they are being used in these conversations about logistic regression / LCA. I need a place that I can return to when I need to translate sentences that are said into non-jargon for the naive learner (me). I know you're saying these definitions as you go along, but I need a reference to give myself a chance to reflect and make these conversations mean something more to me.

- *Outcome Variable:* Something being predicted by something. In the context of mixture modeling, this can either be a distal outcome or the set of indicators that measure the latent variable. It's a multivariate outcome vector (e.g., set of indicators/observed variables)
- *Covariate: ex*ogenous variable, predictor
- *Predictor:* exogenous variable, covariate
- *Logit / Logit Coefficients: Coefficients from the Logistic Regression Models*
- *Nominal Variable:* Variables with no order to them (e.g., variable with ice cream flavors, chocolate, strawberry, mint chip. Or, instructional methods like 1) traditional, 2) hybrid, 3) mixed. Order doesn't matter.

- *Ordinal Variable:* Ordered Categorical Variables (likert type variables, 1 = strongly disagree, 3= disagree, 5= agree, 7 = strongly agree. Where there is n order to the categories, in this example higher values means more agreeable.
- *Indicator:* observed variable (e.g., item on the survey)
- *Distal Outcome:* outcome variable that is traditionally measured later– things that are often predicted by something. For example, if you did an intervention in freshman year your distal outcome may be if they graduate in a STEM major or not.

The video from Pre-Training Day 4 has a lot of choppy audio. Any chance we can get a cleaned-up version? (I realize that this may be asking for someone to re-record a walkthrough of the slides - sorry Adam - but I feel like I'm missing things from the recording.)

Sorry. We are not going to re-record the videos. Check out the links to the videos on logistic regression and make an appointment to talk with us at office hours about logistic regression if you need it.

I would like to understand why we need to take the difference of the OR and not just report the results of the OR. I was under the impression that you would present either results so long as you interpreted them properly. Is that not the case? Should you always be reporting the difference of the OR?

We only sometimes subtract it from 1 when OR is less than one. [We'll post some references soon]

I'm coming from a space where it was highly emphasized that you need multiple indicators to capture a complex construct (i.e., EFA/CFA) but now LCA seems to only use single indicators to capture different ideas (which feel like constructs). I've also gotten push back from reviewers when I use single indicators, they cite that APA has outlined how single indicators are not enough to capture complex ideas/concepts/etc. I need help reconciling this dilemma.

This is a great question– more soon.

Follow-up to the question below about transforming Likert scales to distinct categories, all of my data is collected on a Likert scale. I would like to find a way to use LCA with my data that will not create measurement issues.

Unfortunately measurement issues are everywhere. We can talk more about how to be aware of the issues and to build an argument for the choices that you decide on.

Can you provide literature that has advocated for transforming likert-type scales into distinct groupings? Ex. Likert-scale 0-strongly disagree to 6-strongly agree (3-neutral).

This is a great question that comes up a lot in any statistical modeling, as well as mixture modeling. If you dig in the literature, you actually can find "support" (in the form of a reference) for a range of options for how to treat this variable in statistical modeling. Some would say it's okay to treat them as continuous, some suggest treating them as ordinal, others creating a reduced binary item. My general response is always that given the research context, you made a decision to collect the data using a specific format, which likely was informed by either previous work or your own ideas on what you are measuring. I will note that a lot of times we as researchers model our work after other scholars, seemingly without question. So if possible, keeping with the scale you presented the respondents would be the ideal approach (e.g., if you create a binary variable, it's unclear that the respondents would respond in your categorization if they had, in fact, been given a binary response option). Additionally, reducing categories results in a loss of information (add references).

I will say, however, that when we do mixture modeling with nominal variables we often revert to focusing on one category for understanding the emergent classes. This is because modeling the probability of all possible categories (e.g., 7 categories in a 7-point scale) for each of the latent classes not only becomes computationally parameter heavy (lots of parameters!) it also then becomes quite difficult to make sense of the differences/similarities of the response patterns across a set of items/categories and classes. So there is sometimes a statistical necessity to reduce categories in addition to conceptual.

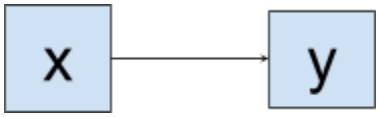
Adam: In the example of the 7-point Likert-type item mention above, I believe there is a strong argument for arguing that Likert scales including "agree", "neutral", and "disagree" response options have qualitative distinctions making categorization appropriate. We are currently looking for references which discuss this issue. Measurement issues (i.e., decisions about how to treat variables) will likely be a theme that comes up frequently in this training. The reference included below makes the argument for treating ordinal scales with relatively homogeneous response options as a continuous scale. Although this is a separate issue I believe it may be useful to the group.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*, 625-632.

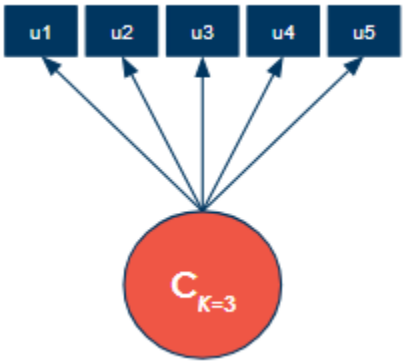
Carifio, J. & Perla, R. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences, 2*, 106-116. http://thescipub.com/PDF/jssp.2007.106.116.pdf

Why is U regressing on C but the arrows are pointed in the opposite direction (slide 13)?

Great question! This is commonly misunderstood. There are a few ways to think about it. In regression, we regress the outcome, y, on the predictor x. So if we drew a path diagram for regression it would be drawn in a path diagram like the one below. Here y is the endogenous variable or outcome variable that is being predicted by the exogenous variable, or predictor variable x.



In the SEM/LCA world (which I (Karen) would refer to as the latent variable framework) we have a multivariate outcome which, in the figure below, is the set of u's, u1-u5. They are the endogenous variables being predicted by the latent variable. In the path diagram, the latent variable, C, is the exogenous, or explanatory variable. Another way to think about it is that your class/group membership (C) is what is causing your responses on the U's, so the arrows go from C to U.



In the asynchronous videos, there's mention of a "repository" to download useful things to help us get acquainted with R (e.g. there's mention of an R project to use for the tutorial and a github containing data that doesn't look like the one we can access,

timestamp 4:50 and 13:25 in the Intro-to-R video). Where is this? Are we doing this later?

Yes- This video was made by another research assistant, Dina Arch, with support from a previous grant. The repository associated with this video is here (immerse-ucsb/intro_to_rstudio). On day-5 we will be going through a similar coding exercise in class that can be found here (MM4DBER/intro-to-rstudio) if you would like to start on this early. The handout for this tutorial is here (Handout_Intro-to-Rstudio).

What are the articles you referred to in the day 1 pre-training slides?

We included links to all of the articles in the slides and reposted a new day 1 pre-training PDF for you to download from the github space (https://mm4dber.github.io/). Here are references to the two articles that were not included in the original posting of the slides.

Suzuki, S., Morris, S. L., & Johnson, S. K. (2021). Using QuantCrit to advance an anti-racist developmental science: Applications to mixture modeling. *Journal of Adolescent Research*, *36*(5), 535–560. https://doi.org/10.1177/07435584211028229

Tabron, L. A., & Thomas, A. K. (2023). Deeper than wordplay: A systematic review of critical quantitative approaches in education research (2007–2021). *Review of Educational Research*. Advanced online publication. https://doi.org/10.3102/00346543221130017

Clarify terms like: "Classes", "Patterns", "Subgroups", "Indicators", etc.

What might be confusing is that some of these words can be used interchangeably. In a manuscript we suggest you pick one and consistently use it throughout so that the readers aren't confused. For example, you could refer to indicators (we will talk about this more on day 3 pre-training) that are used to identify classes or subgroups. The indicators are the "observed" items that tell us something about each individual. An example indicator is: I work hard at science. Imagine this being an item that you ask students on a survey. The response to this indicator could be yes/no. This survey item is just one indicator of student motivation in science. Imagine having several indicators of student motivation in science (another survey item might be: Science is useful for my future career plans) that taken together tell us something about student motivation in science. Students respond to these indicators and their pattern of

responses are used to explore how individuals within the classes or subgroups have similar patterns in terms of the indicators and different patterns from other classes or subgroups. Karen spoke to this a bit in the day 1 pre-training so if you want to hear more about it, check out the video. We will also use these terms throughout the training and will go into greater depth (and you'll see it written about in different ways), so keep checking on your understanding.

What are some existing databases that I might be able to use?

ICPSR is a repository of datasets. Some STEM-specific datasets is the Longitudinal Study of American Youth which was "designed to examine the development of: (1) student attitudes toward and achievement in science, (2) student attitudes toward and achievement in mathematics, and (3) student interest in and plans for a career in science, mathematics, or engineering, during middle school, high school, and the first four years post-high school. The relative influence parents, home, teachers, school, peers, media, and selected informal learning experiences had on these developmental patterns was considered as well." You can search through the ICPSR to see if there are other databases that may be relevant to you.

The National Center for Education Statistics (NCES) has a variety of publicly available databases that might be of interest to you. For example, one dataset focused on higher education is the Integrated Postsecondary Education Data System which includes 12 interrelated surveys that gathers annual data from every college, university, and technical and vocational institution that participates in the federal student financial aid programs. Here are some specific other database suggestions from NCES:

a. https://nces.ed.gov/surveys/nsopf/ - National Study of Postsecondary Faculty (NSOPF) Features
b. https://nces.ed.gov/surveys/npsas/ - National Postsecondary Student Aid Study (NPSAS) Features
c. https://nces.ed.gov/surveys/b&b/ - Baccalaureate and Beyond (B&B) Features
d. https://nces.ed.gov/surveys/bps/ - Beginning Postsecondary Students (BPS)

Trajectories into Early Career Research Data Set: An 8-Year Longitudinal Mixed Methods Data Set of Biological Sciences Ph.D. Students: The dataset contains 8 years of surveys (biweekly and annual), interviews, and performance-based data from a national cohort of 336 Ph.D. students who matriculated into U.S. biological sciences programs in Fall, 2014. This dataset includes data on students' research skills, career goals, to explore beliefs about postdoctoral

research opportunities, and data on how race and gender shape graduate and postdoctoral research experiences. This dataset lends itself well to analyzing the relationship between graduate and post doc research experiences and research skill development and career trajectories and for assessing issues of inequity in graduate and postdoctoral programs. These de-identified data will be publicly released on the Open Science Framework data repository in 2023.