# Latent Class Analysis (LCA) Part 2: Extensions of LCA

### with Stephanie Lanza and Bethany Bray

### Stephanie Lanza, Aaron Wagner, and Bethany Bray

### **January 22, 2013**

[Stephanie Lanza](http://methodology.psu.edu/people/slanza) and [Bethany Bray](http://methodology.psu.edu/people/bbray) discuss extensions of LCA with host Aaron Wagner. Topics include LCA with grouping variables and covariates, latent transition analysis, causal inference in LCA, and LCA with a distal outcome. The discussion assumes that users are familiar with LCA; part 1 provides introductory information.

**Podcast Timeline:**

00:00 - Introduction

01:00 - Adding grouping variables and covariates to an LCA

12:42 - Causal inference in LCA

17:40 - Predicting distal outcomes using latent class membership

26:26 - Upcoming LCA trainings

Speaker 1: Methodology Minutes is brought to you by the Methodology Center at Penn State, your source for cutting edge researched methodology in the social, behavioral, and health sciences.

Aaron: Hello again. We are back with Stephanie Lanza and Bethany Bray to talk again about latent class analysis. Bethany, Stephanie, thanks for being here.

Bethany: Thanks, Aaron.

Stephanie: Glad to be back.

Aaron: If people are looking for an introduction to latent class analysis they should really try the first podcast we recorded because we're going a little further down the rabbit hole today, is that fair to say?

Bethany: That is fair to say. We're going to talk a little bit about extensions to LCA and then talk about some fairly advanced topics that we've been conducting research on.

Stephanie: That's right. We like to be in the rabbit hole, so to speak.

Aaron: Let's jump right in. What are the most common and useful extensions to latent class analysis?

Bethany: The two things that people are most likely to start off with once they have a latent class model is to add either a grouping variable or a co-variate or perhaps both to that, what we would think of as a baseline latent class analysis or a baseline LCA. A grouping variable would be something like gender. Let's say that we started off with a latent class model of sexual risk behavior.

Maybe we asked participants questions about how many dating partner they had, how many sexual partners they had, and whether or not they used a condom 100% of the time, and we identified five latent classes of sexual risk behavior from those indicators. Let's say we have non-daters, daters, monogamous individuals, multi-partner safe individuals who have sex with multiple partners but use a condom 100% of the time, and then multi-partner exposed individuals who have sex with multiple partners but don't use a condom 100% of the time.

Once we have that baseline LCA model, with those five latent classes, perhaps we want to ask the question, "Do men and women have the same structure of those latent classes? Do multi-partner exposed females look the same in terms of their measurement structure as multi-partner exposed males?" That would be using a grouping variable to exam measurement in variance.

If it turns out that indeed our latent classes are measurement in variance, which is a good thing, typically that's what we want, what that means is that those five latent classes for women look the same as the five latent classes for men and we can move on to answer questions like, "Are the prevalence rates at those five classes the same for men and women? Are men more likely to be multi-partner safe compared to women who might be more likely to be monogamous?" for example. Stephanie has a paper, that maybe you could tell us about, that addresses this question, right?

Stephanie: Sure. I have a paper with Linda Collins that appeared in developmental psychology, so you can download that from our website. The paper investigated just what Bethany was talking about. One of the parts of it was to look at gender differences in the prevalence of membership in these sexual risk behavior latent classes.

We found that males and females, these are late adolescents, they were equally likely to be non-daters and daters, but we found that women were twice as likely to be monogamous compared to males, males were twice as likely to be multi-partner safe individuals compared to females. Both of those categories have some potential for risk, but not super risky. The multi-partnered exposed latent class, the high risk latent class, was equally prevalent for males and females.

Bethany: So once we're able to ask these questions about grouping variables maybe, or alternatively we're interested in asking something about co-variates, which allow us to predict membership in those latent classes. If we think about heavy episodic drinking or binge drinking, maybe we want to use that binge drinking variable to predict which latent class’s people go on to. For example are people who binge drink more likely to be in the high risk, multi-partner exposed latent class compared to another latent class relative to the people who did not binge drink at that time? I believe that's also investigated in that paper, is that right?

Stephanie: That's right. That’s exactly what we found, Bethany. To be clear the question we were answering with this analysis right here is "How does past year binge drinking predict membership in your sexual risk behavior latent class?" Past year drunkenness increased your odds of being in the multi-partner exposed, that high risk behavior class compared to being a non-dater more than eight fold for adolescents. So it's a big risk factor, sexually speaking here.

Bethany: As in the case in this paper, we talked a little bit about the results for males versus females and then the relationship between binge drinking and sexual risk behavior, but we might want to ask more sophisticated questions about moderation. Is that relationship between binge drinking and sexual risk behavior different for males and females? That's a fairly straightforward extension. All you need to do is include both a grouping variable and a co-variate into your model and that allows you to look at this moderation of the relationship between the co-variate and the latent class variable, which I think you also did in that paper.

Stephanie: We did. You're right. It's exactly that. It's essentially allowing the co-variant and the grouping variable to interact in the prediction model. In this case that role of heavy drinking, past year heavy drinking, it was really identical for the male and female adolescents, that behavior placed them at equal risk.

Aaron: Thank you very much. Another topic that Methodology Center researchers have done a lot of work on is latent transition analysis, another longitudinal extension of LCA. Could you talk to us about that?

Stephanie: Let me give a little bit of history here. Linda Collins is our Director of the Methodology Center. She came to Penn State in, I believe it was 1994. Shortly after her arrival she began directing the Methodology Center here. She brought with her her methodological innovations which were, basically, latent transition analysis as we know it today. She developed this technique for examining shifting latent class membership over time.

Bethany: The purpose of latent transition analysis, to put it simply, is that you're interested in examining development as movement between these discreet states. If we start with latent class analysis, LCA, at a single time and we identify these classes, maybe what we want to know is how do people move between classes over time. If we start in high school and we identify these five sexual risk behavior statuses, how do people move from a comparatively low risk status, maybe non-dating or dating, into a comparatively high risk status, something like multi-partner exposed?

How likely are participants to make those transitions? Those are the kinds of questions you can address with latent transition analysis or LTA. LTA is something that was also investigated in this Lanza and Collins 2008 paper in developmental psych. They were primarily interested in this question of how likely are participants to move between the statuses over time.

Stephanie: Right. We looked at adolescents across three subsequent years. At time one, these were throughout high school and into emerging adulthood. From time one to time two and also from time two to time three we estimated their shifts in behavior, their changes in behavior over time in terms of their sexual risk behavior.

One interesting finding that this paper shows is that the individuals, the adolescents who are most at risk for transitioning for high risk behavior at the next time point were not the multi-partner safe individuals, they were the monogamous individuals. Those are potential targets for intervention to prevent them from transitioning to that high risk status.

Bethany: Once we look at questions like that, "Who is most likely to make these risky transitions?", for example, maybe we want to go one step further and exam grouping variable and co-variates that can help us better understand who is making those transitions. This is conceptually similar to what we do in LCA; in the same way we can add grouping variables and co-variates to LTA.

In that case, just like in LCA, if we add a grouping variable, we can exam things like measurement and variance over time. "Do men and women have the same statuses across time?", once we establish whether or not that's true we can look at whether or not the prevalences are different over time, maybe several time points. We can also incorporate co-variates, just like in LCA, to predict latent status or latent class membership at time one. We were talking before about using binge drinking to predict sexual risk behavior at a single time.

Now we can still do that, perhaps in high school, maybe that's time one, but we're interested in transitions from high school to college let's say. We can now use those co-variates not only to predict the high school behavior, but also how those participants are transitioning from high school to college, "Is binge drinking associated with certain kinds of transitions? Are people who binge drink and are also non-daters the ones who are likely to move to multi-partner exposed or is it another kind of drinker or another kind of person with a different kind of sexual behavior?"

Stephanie: I like this model a lot. When you think about predicting transitions ... The way Bethany described it is one really, I think, intuitive way to think about it is "Does binge drinking predict transitions from this sexual behavior latent class at time one to another sexual risk behavior latent class at time two?" That can be very interesting.

Another way to word that questions, but it's the same approach is, "Does your baseline sexual risk behavior latent class membership moderate the effect of the co-variate on that later outcome?" Another way to interpret those effects is thinking about your latent class variable as a baseline latent class moderator, which I think is really cool. Michael Cleveland and I recently published a paper on this very topic.

Before we move on to the next segment, Aaron, I wanted to make sure that our listeners are aware that we developed software for latent class analysis and latent transition analysis. If you go to the website and download PROC LCA and PROC LTA, it's a single download. What that does is it installs on your computer a suite of two SAS PROCs and those become part of your local SAS installation that you can use to fit all of the models that Bethany and I have been talking about.

Bethany: That models that we've been talking about up until this point we've been heavily referencing the 2008 developmental psych paper by Lanza and Collins. That paper has fairly extensive appendix that walks you through the programming code for all of the models that we've been talking about here with grouping variables and co-variates.

Stephanie: In the context of latent transition analysis that paper really is about LTA. If you're interested in the nuts and bolts or more about LCA, I would probably refer people to, there's a 2007 paper we have, constructional equation modeling, called PROC LCA and a handbook of psychology book chapter that is in press.

Aaron: A couple more plugs I'll throw in, there is a reading list for latent class analysis and for latent transition analysis on our website, the latent class analysis research page. Also all of the functionality that's available in PROC LCA, we will soon be releasing a Stata plugin that incorporates all that functionality.

Stephanie: That's right. That's right.

Aaron: for non-SAS users. As interesting, as excited as you are in that work, I know that every researcher wants to talk about what they've been doing recently. I know both of you have been working on several extensions of LCA recently, would you like to talk about some of those?

Bethany: Yes we would.

Stephanie: Absolutely.

Bethany: We are very excited to tell you about what we've been working on. Before we get started, as a preview, we're going to talk about two or three different topics and all of this work is out now or are coming out very shortly so you can visit the website for the references for all of these papers. We'll try to make note of them so you can match them up.

Stephanie: Right. Many of our listeners might be aware that our center is funded by a NIDA P-50 Center Grant. One of the components in our center currently is being led by Donna Coffman who is working in the area of causal inference, in particular propensity score methods to causal inference problems. I am the principal investigator of a component on mixture models. We focus primarily on extensions of latent class analysis. Donna and I, along with Bethany who's interested in both areas, have been very excited to be collaborating on the intersection of these two areas.

The first area I wanted to talk about today is the intersection of causal inference and latent class analysis. The questions that people want to answer in latent is "Who's in these classes? What predicts membership in these latent classes?" We can take it one step further with new methods and talks about what are the actual determinants of latent class membership, so "What early exposures might determine what latent class you subsequently end up in, be it a problem behavior latent class, a sexual behavior latent class, dependence latent class, or anything?"

The question that we decided to address empirically in the paper where we lay out this method for the first time of "What is the causal effect of enrolling in college where binge drinking is so notorious and heavy? What is ultimately the causal effect of going to college on their subsequent problematic substance use? After they're out of college and they're entering early adulthood, what is the link that leads people from college enrollment into problematic behavior or less problematic behavior?" We don't know the answer to that question.

The golden standard to get at "What's the causal effect of going to college on their later drinking or problem behavior?" is a randomized trial. You can imagine intervention scientists might want to say, "Let's randomize all of the individuals in high schools, you guys go to college, you guys don't go to college, then we'll compare your adult drinking behaviors." That's never going to fly. We can't randomize people to go to college.

In fact there's huge selection effects related to going to college, paternal education, household income, expectations for your future, education attainment, all sorts of things. Propensity scores provide one avenue for answering these questions. What we did was we used two propensity score techniques in this paper. We used first propensity weights and we also investigated using matching. We did that to adjust for all sorts of confounders that might confound that estimated association between college enrollment and later substance abuse behaviors.

One of the things that propensity score techniques in general give us the ability to do is to make a very specific causal question that we want to answer. We thought that the question to answer here that made the most sense, that had the most practical implications, was fork individuals who actually went to college, "What did the act of going to college do for them?" because we know it doesn't make any sense to think about every individual in the population actually going to college because it's not possible in our society.

Among those who went to college, what did the effect of college do? Because we have the non-college comparison group we can use propensity scores to balance these two groups on all sorts of confounding characteristics and get our causal answered. At the end of the day, drum roll, there was no causal effect of college enrollment on adult substance use, latent class membership, meaning college was neither a protective for future behavior nor was it a detriment to the later performance. Anyway we're continuing to work in this area and we're very excited about it.

Aaron: That's not the only thing you guys have up your sleeve, what else?

Bethany: What Stephanie has been talking about is this causal prediction from a co-variate to latent class membership. In addition to our work in investigating causal inference methods on the prediction of latent class membership from the co-variate, we've also been investigated questions about how to take a latent class variable and use that to then predict outcomes in the future. So moving from the co-variate predicting latent class membership to now using that latent class membership as a predictor.

Doing this or wanting to do this has been around for a long time. It makes a lot of conceptual sense. For example we might want to take latent classes of depression and predict future alcohol use, but how to do this is not a straight forward methodological question. As Stephanie mentioned earlier, we've understood the model of latent class analysis with co-variate for some time, but the mathematical model for how to take an unknown latent class membership and use that to predict as distal outcome is somewhat more challenging,

The common approach to doing this would be something that we call a classify, analyze approach using either a maximum probability assignment or something called pseudo-class straws. In either approach what you do is you hard assign individuals to latent classes. We would take people's responses, all the participants in our study and based on their responses we would assign them to belong to one of five latent classes of depression, for example.

Then you would take that latent class membership and you would treat it as known in some larger analysis model. In this case, if we do indeed want to use latent class membership of depression to predict future alcohol use, what we would do is we would assign people to depression latent classes, then we would regress their alcohol use on that latent class membership, treat it as known.

For a long time it's been known that doing that is not ideal for a variety of reasons, one of those reasons is that we no longer make use of that latent class measurement model, which is an advantage to latent class analysis, but also that assigning people to latent classes and then relating that latent class membership to distal outcomes attenuates the association between depression and alcohol use in this case.

For example, perhaps the people who are most depressed are also the heaviest alcohol users. If we assign people to depression latent classes and then regress alcohol use on the depression latent class membership we're going to underestimate the alcohol use of people who are most depressed, that's problematic because the stronger the association between latent class membership and that distal outcome, the worse this approach performs.

Stephanie: Yep. It's ironic, but it's unfortunate

Bethany: It's ironic that the better the situation you're in, which is that you want your class membership to be highly related, the worse job you actually do. We've approached this problem from two different angles. One angle is to try to improve this classify, analyze technique. We've done some work in that area. We have a tech report that just recently came out by Bray, Lanza, and Tan from 2012. Its tech report number 12-18, if you're interested.

It looks at an alternative approach to this classify, analyze that makes use of a more inclusive model that reduces that attenuation. It's a highly general approach when classify, analyze is something that the scientist needs to do. It tries to improve that methodology. The other direction that we've gone which is very interested is this idea of a model based approach to the problem. What do I mean by that?

With latent class analysis with co-variates we have what we think of as a model based solution, that is we can directly incorporate the co-variate right into the latent class model and we don't need to hard assign people to classes. We wanted a similar solution to be able to do that with LCA, latent class analysis with a distal outcome.

There's a paper that's in press at structural equation modeling right now by Lanza, Tan, and Bray, it's coming out it the next month or so, that walks people through what we came up with as a model based solution. This is really cool because it doesn't require assigning people to classes in order to predict that distal outcome. How does it work? Let's say that we are interested in this question that I've been talking about, "How does depression latent class membership predict alcohol use?"

In order to answer that question what we really want to know, the mathematical question we're asking is "What is the distribution of alcohol use given depression latent class membership?" That question is hard to answer because we don't know people's depression latent class membership. By using base theorem, we can slightly alter that question to make it answerable.

Stephanie: Base theorem provides the key to this problem, right, Bethany? We want to know the distribution of alcohol use given depression latent class membership. You can express that association as a function of three other parameters, sets of parameters that we do know or can get, is that right?

Bethany: Yeah, that's right. We can get the overall distribution of depression by simply fitting a latent class model. We can also get the distribution of depression given alcohol use by using latent class analysis with co-variates because that's what it produces, latent class membership conditional on the co-variate of alcohol use in this case. Those two pieces of information are critical. The third piece of information that we would need to get the distribution that we want is the overall distribution of alcohol use in the population, that's the piece that's a little bit trickier.

If we have a binary variable or a count variable for example, we can take the direct empirical distribution in our sample to produce that, but if our distal outcome of interest is continuous, that becomes a little bit more difficult particularly if it's continuous but not normally distributed. To get that last piece of information to put all of this together using base theorem what we did was use a kernel density estimation approach.

All the kernel density estimation approach does is it estimates the empirical distribution of that variable from the data provided by our participants. By using these three pieces of information, the overall distribution of depression, the conditional distribution of depression given alcohol use, and the overall distribution of alcohol use, we can recombine that information using base theorem to produce what we want, which is the conditional distribution of alcohol use given depression, which answers our question of, "How does depression predict alcohol use?"

Stephanie: That's right. It makes some users uncomfortable, I think, to think about having to fit this model where you have to put your outcome in as a co-variate, and to get one piece of the puzzle that Bethany just described, the conditional distribution of the latent classes given the distal outcome.

It's a little bit weird, but that's just the piece of information that we need. It's a means to the end, which is to actually get parameter estimates that describe the association we want. Their simulation results show that that basically eliminates all of the bias that you would see with a classify, analyze approach, a standard classify, analyze approach.

Bethany: From the user's perspective it sounds a little bit complicated; you have to use these tools and you have to use kernel density estimation, but this is all automated in a macro that is distributed in conjunction with this paper. The macro is called the LCA Distal. It's downloadable from the Methodology Center's website.

All of this goes on behind the scenes and you just have to know what your distal outcome is and how to fit a latent class model using the PROC LCA in SAS. If you can do those things, you can use the macro in a very straight forward way to get the answer that you want, which is, depression predicting alcohol use.

Stephanie: The classify, analyze approach that Bethany alluded to earlier, that you can read about in the SAS report, it is even easier, I would say, to use.

Bethany: It's very straight forward.

Stephanie: It's very straight forward for the user. It's a really cool solution to the problem.

Aaron: Thank you both very much. Over these two podcasts we've talked about a lot of major topics from the most basic introduction to LCA to the latest cutting edge in LCA extensions. If a listener was interested in getting training from, I don't know, say the two of you, where would they go to get that training? What's coming up?

Stephanie: The most immediate opportunity is this June. In June 2013 Bethany and I, together will be teaching a two day, hands on, so bring your laptop, bring your data, we'll be doing an intensive two day training on latent class analysis. This is part of the summer institute series that we've been hosting for many years now. It is funded by NIDA so we will be primarily targeting drug abuse and HIV researchers who want to learn LCA to use in their research.

Aaron: Well, Bethany, Stephanie, thank you very much for your time.

Bethany: Thanks, Aaron.

Stephanie: Yeah, thanks for the chance to talk about it.

Speaker 1: You have been listening to Methodology Minutes brought to you by the Methodology Center at Penn State, your source for cutting edge researched methodology in the social, behavioral, and health sciences. This podcast is available on iTunes and at methodology.psu.edu.

**References:**

Lanza, S. T., & Collins, L. M. (2008). A new SAS procedure for latent transition analysis: Transitions in dating and sexual behavior. Developmental Psychology, 42(2), 446-456. PMCID: PMC2846549

Lanza, S. T., Coffman, D. L., & Xu, S. (in press). Causal inference in latent class analysis. Structural Equation Modeling: A Multidisciplinary Journal.

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. Structural Equation Modeling: A Multidisciplinary Journal. Advance online publication. doi: [10.1080/10705511.2013.742377](http://www.tandfonline.com/doi/full/10.1080/10705511.2013.742377) PMC Journal- In progress

Bray, B. C., Lanza, S. T., & Tan, X. (2012). An introduction to eliminating bias in classify-analyze approaches for latent class analysis (Technical Report No. 12-118). University Park, PA: The Methodology Center, The Pennsylvania State University.