# Latent Class Analysis (LCA) Part 1: Common Questions about LCA

## with Stephanie Lanza and Bethany Bray

## Stephanie Lanza and Bethany Bray

## **November 26, 2012**

### Methodology Center scientists [Stephanie Lanza](http://methodology.psu.edu/people/slanza) and [Bethany Bray](http://methodology.psu.edu/people/bbray) and host Aaron Wagner discuss common, practical issues that arise in latent class analysis (LCA). Issues include selecting indicator variables, selecting a model, determining the necessary sample size, finding LCA software, and getting started in LCA. This is the first in a two-part podcast; the next podcast will address some of our recent research on LCA.

###

**Podcast Timeline:**

00:00 - Introduction

01:05 - What is LCA?

02:50 - How is LCA different from factor analysis?

06:48 - Indicator variables

10:54 - Model selection

14:30 - Sample size

19:07 - Software

20:06 - Getting started

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

Aaron: Hello, and welcome to Methodology Minutes. We’re sitting down on a grey November morning to talk about latent class analysis, and here to talk about it with me I have Bethany Bray, an assistant professor of psychology at Virginia Tech and visiting faculty this year at the Methodology Center, and also Stephanie Lanza, who is a research associate professor in the College of Health and Human Development at Penn State and the scientific director of the Methodology Center.

 Good morning, Bethany.

Bethany: Good morning, Aaron.

Aaron: Good morning, Stephanie.

Stephanie: Hi, Aaron.

Aaron: Thanks, both, for being here.

Stephanie: We’re happy to be here, and I just want to say I’m even happier that Bethany is with us as a faculty member this year. It’s been delightful.

Aaron: Agreed. Just to start out with as an introduction, Stephanie, can you tell us what latent class analysis is?

Stephanie: Sure. Latent class analysis is one type of latent variable model. Really generally, latent variable models are statistical models for measuring a variable that you can’t observe directly. You can only measure it with error, that is with items that you can ask people or observe about people but they don’t … none of them perfectly maps on to that latent variable, but they all have a little bit of error in them.

 For example, if you just wanted to know about adolescent alcohol abuse, you might just ask a group of adolescents, “How much did you drink last weekend?” They would give you an answer and you might assume that that’s measured without error, and that’s your variable.

If, on the other hand, you’re interested in understanding what are the complex substance use behaviors that adolescents are engaging in, then you might ask them about different dimensions of their alcohol use, how frequently they drink, how often they binge. You might ask them about their smoking behaviors, and marijuana use, and other drug use. Together, those indicators would allow you to identify subgroups of adolescents who have different patterns of substance use. In latent class analysis, an individual’s subgroup membership is actually unknown. It’s not observable, and so we don’t really know it. It’s a probabilistic model.

Latent class analysis has been increasingly popular among social scientists, behavioral scientists. Its use is really ramping up, and the broad use of LCA includes lots of different examples. Substance use is one big one, but also dieting patterns, eating behaviors, temperament types and risk exposure.

Aaron: Okay, great. Thank you. Another latent variable model that’s very common is factor analysis. Some of our listeners might be familiar with factor analysis. How is LCA different from factor analysis?

Bethany: Factor analysis and LCA are both latent variable models, but they differ somewhat in their primary goal or the primary reason why you might choose to use one technique versus the other. Factor analysis we would consider to be a variable-centered approach or a variable-centered technique, because it strives to identify sets of variables that are similar to each other. So, sets of variables that work together to identify some underlying latent construct.

 In comparison, we consider LCA to be a person-centered technique or a person-centered approach because we use it to identify subgroups of individuals who are similar to each other in terms of their patterns of responses to indicators. As Stephanie said, subgroups of individuals who are similar in the way that they answer questions about their drug use.

 Practically speaking, factor analysis is used to identify a latent variable that’s continuous, so it lies along a single continuum and typically those factors are measured using continuous indicators. In comparison, LCA identifies a categorical latent variable where those categories are the subtypes of individuals. Typically, we use categorical indicators to identify those types.

 If we were thinking about a specific example, let’s say that we asked individuals eight questions about their depressive feelings. We ask them whether they’re feeling the blues, whether they’re feeling depressed, whether they’re feeling sad, whether they’re feeling their life is like a failure and so on.

If I was interested in a factor analysis type of question, I might be interested in using all eight of those questions to measure a single unidimensional construct of depression. I might determine whether or not all eight questions assess depression, and if they do I could sum the scores or use a factor score to place individuals along that continuum, determining how depressed or how not depressed they might be.

In comparison, for LCA I’m more interested in whether or not there are qualitatively different types of depression that can be identified using those same eight questions. For example, I might identify a group of individuals who are sad, meaning they answer that they’re feeling lonely, and sad, and depressed, but they’re not feeling disliked and they’re not feeling like a failure.

In comparison, there might be another group of individuals who are feeling disliked only. They might be feeling disliked only, but not sad and not like their life is a failure. There might be another group of individuals who are feeling all of those things, and we might term them as being highly depressed.

Stephanie: Another way to think about the distinction between factor analysis and latent class analysis, with factor analysis typically you have a bunch of items and you’re looking for what are the different constructs we’re tapping into that happen to be theoretically continuous dimensions.

So, what are the different things that we’re tapping into with these items? In latent class analysis you also have a bunch of items on a bunch of people, but what you’re tapping into is not the underlying dimensions but organizing people into subgroups where their actual subgroup, the inherent nature of those subgroups expresses the intersection of those dimensions for those people.

Aaron: Okay. Thank you very much. One of the most common questions we get as soon as somebody starts to try and do a latent class analysis is how they select these indicator variables. You mentioned the different indicators that go into it. People want to know, how do they select their indicators and how many indicators are too many?

Stephanie: That’s a great question, and you’re right, we get it all the time. The first thing I tell people when they’re just embarking on a new latent class in a new area for them is when you think about choosing the indicators for your latent class variable the single most important thing I would argue is to choose indicator variables that tap into all of the dimensions of the underlying construct you’re trying to measure; all the dimensions, all the aspects of that behavior for example, but nothing else.

For example, if you wanted to measure smoking behavior in adults and you wanted to look at smoking behavior broadly defined, then you might want to have indicators of how frequently they smoke, how much they smoke throughout the course of the day, how soon after waking they smoke their first cigarette, so some sort of measure of dependence. Together, you would want to make sure that you are set of items taps into all the dimensions of smoking behavior that you think are important and nothing beyond that.

Of course, if you’re looking at a complex construct, a complex behavior maybe, you want to have more indicators in there. But how many are too many? The problem is, as we continue to add indicators the size of our underlying data table, our contingency table that we’re analyzing grows exponentially.

Aaron: How does it grow exponentially?

Stephanie: Let’s say you have three binary indicators that tap into some sort of smoking behavior. An individual might say yes or no to item one, yes or no to item two and yes or no to item three. That results in a contingency table, which is what’s being analyzed in an LCA contrast. Correlation matrix is being analyzed in a factor analysis; here it’s a contingency table.

Individuals could respond yes or no to each of those three items, so that results in two, that’s two response options, raised to the power of three for three items: Two to the third, which equals eight possible patterns of responses to those three items. Let’s say you throw in a fourth dimension, a fourth aspect of smoking behavior. Now you have another binary item. Now you have two to the fourth, or two times two times two, which is 16 possible patterns of response to those items.

You’ll see, as you add an item, the dimensions of that underlying table or the number of possible response patterns increases exponentially. The downfall of that, of course as that number of possible response patterns increases then you’re going to start to introduce sparseness into your data. That means you might not have enough people showing all those different response patterns in high frequency. You’re going to have sparseness in that data table, and that can lead, sometimes quickly, to problems with computation and problems with identifiability.

Bethany: Stephanie, would you say that practical advice would suggest that the fewer the indicators you have, so the fewer items you want to include in your model, you can probably have more response options, but the more indicators that you have, it’s typically prudent to have fewer response options in the indicators that go into the model?

Stephanie: Definitely. I think for a given sample size, if you have items that have more than two response options, more than just yes and no, maybe you have low/medium/high or maybe you have four qualitatively different responses to that item, if you have a lot of items like that then you have to reduce the number of items. If your items are all binary, that’s the simplest case for indicators and I would say the most common situation people are in, if all of your items are binary you can get away with introducing more into the model.

Aaron: Great. Thank you very much.

Stephanie: Sure.

Aaron: We have the indicators now. Bethany, what’s the best way to select an LCA model?

Bethany: When we select an LCA model there are often two things that we want to consider. The first is whether or not our model fits the data absolutely well, so that the model is an accurate reflection of the variability in our data. In comparison, we probably also want to consider whether or not our model fits relatively better than other models. The reason that we want to consider both absolute model fit and relative model fit is because there are challenges to doing both.

 Again, absolute model fit tries to determine whether or not our model fits the data in an absolute sense. In practice, this is difficult to do because we need to have some kind of test statistic that tests whether or not our data fit that model.

Typically we rely on the G2 likelihood ratio fit statistic, but as Stephanie mentioned, these contingency tables upon which LCA is based get large very quickly and as they get large they also get sparse. It’s known that when our contingency table is too sparse, our G2 test statistic does not follow any known reference distribution. Meaning, we can’t actually test whether or not our model fits well.

 When we’re in such a situation, which happens nearly all of the time, we need to rely on relative model fit and we need to try to assess that. In that case, what we’re trying to determine is whether or not, for example, a model with three latent classes fits just as well as a model with four latent classes, or whether or not we actually need an additional latent class to do a good job explaining the variability in our data.

 When we rely on relative model fit we use several different criteria to try to help us determine which model fits optimally or which model fits best. Two very popular ones are the AIC and the BIC, so the Akaike Information Criterion and the Bayesian Information Criterion. There are a variety of others that people might be interested in and a variety of others that get printed out in software, but the goal of any penalized fit criterion is to balance fit with parsimony.

Every time we add to latent classes we’re also adding parameters that need to be estimated. Every time we add a class our model fits better, but we don’t want to over fit. We don’t want to extract too many latent classes, for example. These penalized criteria help us determine how to balance that improved fit with the decreased parsimony.

Stephanie: Right. You want to have a model that on the one hand fits your data well, your data are plausible given that model, and on the other hand that model is not tailored too much to that particular data set and its idiosyncrasies, so it’d be more generalizable.

Bethany: That’s right, and there are other ways to assess relative model fit that are not based on these penalized criteria. For example, we can also compute a bootstrap likelihood ratio test to compare two models. This is becoming more common, and that’s referred to as the BLRT.

Aaron: Another common question when somebody wants to do an LCA is about sample size.

Bethany: The bigger the better.

Stephanie: Done.

Aaron: If you have any other thoughts you might care to share on the nature of how big should someone’s sample be?

Stephanie: It’s a great question, again, Aaron, because of course we get this question all the time. I would say increasingly so, LCA is being adopted by lots of researchers now, lots of researchers who are applying for funds at NIH.

As I’m sure our listeners know, when you’re applying for funds at NIH they like to see a power analysis included in your grant. Of course, people are writing to us and posting on blogs and things, “How many individuals do I need to do this study?” As with power analysis, you have to make a whole set of assumptions. Under those assumptions it might be possible in a lot of scenarios to figure out what is the minimum number of people that I need to have in my study to find those effects.

 The tricky thing in LCA is what is it that people want to size their studies to do? They want us to know the statistical power of what, exactly? It’s a little bit elusive, so we’ve tried to tackle this problem in a very practical way. We are trying to develop, derive the first ever – that I know of – power curves in LCA. So, for certain situations that we’ve posited that are based on our experience and wisdom as it were, we’ve developed a number of power curves. These power curves and this methodological paper we expect to be accepted for publication very soon, so you can watch for a paper on that.

What we found is that, not surprisingly, statistical power is better. You can get away with a smaller sample size when you have fewer latent classes. That’s a simple model, when you have a simpler measurement model, fewer indicators in your data set. Also, and the key factor – and this is for anything that you’re looking at in LCA, what we’ve found historically – the most important factor ever is what we’ll call the strength of our measurement model, and that is that you have good items tapping that underlying construct.

The better you’re observed items are, the more tightly they map onto that latent construct, then the smaller sample size you can get away with. That’s the key here. I will say that we’ve seen a lot of empirical studies published with samples smaller than 100 participants, you can definitely do it, but of course we don’t really know what the power was for those data sets before they did the empirical study.

Bethany: Stephanie, this issue about the strength of the measurement model I think is really key for listeners because this is the case with any latent variable model, right? I mean with recent factor analysis work there are all of these rules of thumb about how many participants you need to have per parameter estimated. But if you follow the literature, those rules of thumb, they’re not that helpful. It really depends. The stronger your factor analysis measurement model, the fewer participants you can have to identify underlying factors.

The idea is exactly the same here. It’s that the better your items map onto your latent construct, whether it’s continuous or categorical as the case is in LCA, the better job you’re going to be able to do identifying your model and having confidence that those classes are useful.

Stephanie: That’s right. When you’re thinking about sizing a study for LCA, the most common things I think people actually really want to know their statistical power for, one is, “Once I identify latent classes can I predict them, and so can I identify those associations?” That’s a question we have not dealt with yet. We’ve considered a more basic one that I think is where we need to start with identifying power in LCA and sample size requirements for LCA.

It is, very simply, if in reality there are four latent classes out there … we never know that, but let’s say that there actually are in reality four latent classes. What sample size do I need to have in my data set to identify all four of those classes? Or, what sample size do I need to achieve that with the power of .8? Those are the power curves that I’m referring to that we’ll be sharing with the public soon.

Aaron: If somebody wants to perform LCA, what software is available for them?

Bethany: Software is very near and dear to my heart, so I am happy to talk about options. There are two options right now available from the methodology center, one that’s been extensively developed that we’ve made available for quite some time now, and that is PROC LCA, which is a plug-in for SAS Version 9 and above. That provides the capability of running many different LCA models in SAS. We’ve recently been testing Stata plug-in, also for LCA, that has similar capabilities to what we have available in the PROC.

There are other commercially available software products also. For example we use Latent GOLD. We also use Mplus. Both of those software packages will run any of the models that we’ve been talking about here or in our papers.

Aaron: Great, thanks. If a listener is interested in getting started with latent class analysis, where should they begin?

Stephanie: Great question. One thing I would say is the methodology website in general has a really outstanding webpage right now on LCA, introducing the concept, what is LCA.

Aaron: If we do say so ourselves.

Stephanie: If we do say so ourselves. Thank you, Aaron, for righting that. If one wants to delve into reading about it more deeply, what Bethany and I tend to recommend is a Handbook of Psychology book chapter that we wrote along with Linda Collins, our center director. This chapter has actually been around for over 10 years, but we were recently invited to update it for the second edition of the Handbook of Psychology.

That is currently in the publication pipeline right now, and you can email mc@psu.edu for an advance copy of the new version, the updated version of that chapter. The example that Bethany mentioned earlier about depression, latent classes of depression in adolescents, that is described in detail in that book chapter.

Once a user gets really involved in fitting latent class models and starts to face some of the challenges that we’ve alluded to in this podcast, I would probably suggest turning to the book that I recently published with Linda Collins in 2010 that is a comprehensive book on the latent class and latent transition analysis, a closely related model.

In addition to those written materials, I would say probably … Bethany and I have been going around giving workshops on LCA for years now, and I think one of the most useful things for people to really get using LCA and using it properly is to attend one of those workshops.

We often do two-day hands-on workshops where people can bring their computers, install the free SAS plug-in, bring their data even, and learn how to use it and answer questions in their research. I would recommend that. The next one we have planned right now is going to be in June 2013 here in State College, so bring your data and join us for those two days. We’ll be advertising that summer institute soon.

Aaron: Thank you both very much. I think this is going to be really useful to our listeners. They may be interested to know that we have another podcast on latent class analysis coming up. What are we going to be talking about then?

Bethany: I’m very excited about that podcast in particular because Stephanie and I are going to have a little discussion or debate back and forth about the current research that we’re working on in this area. These are going to be a selection of very hot topics that we’re currently researching. Stephanie, what are you going to be talking about?

Stephanie: Obviously one topic that we both are working on closely, that we have just released an updated software package for, is LCA with a distal outcome. So, what are the implications of belonging to a particular class? What are the implications of it? Linking that to a distal outcome. I know we’ll talk about that.

Bethany: We’ll talk about a couple different approaches to that, so a macro and some recent issues that have come up with classifying analyze approaches in LCA. Then also, very exciting, we’re going to be talking about our causal inference work in LCA that we’re working on together.

Stephanie: Yes, with Donna Coffman. That has been really fun, and so it’ll be a nice opportunity in the next podcast to really talk about what you can do with these latent classes once you identify them.

Bethany: Then of course some longitudinal extensions I’m sure will be thrown in there too.

Stephanie: Absolutely, yeah.

Aaron: Stephanie, Bethany, thank you both very much.

Bethany: Thanks.

Stephanie: Thank you.

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

**References:**

Collins, L. M., & Lanza, S. T. (2010). Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences. New York: Wiley.

Lanza, S. T., Bray, B. C., & Collins, L. M. (2013). An introduction to latent class and latent transition analysis. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.),Handbook of psychology: Research methods in psychology (2nd Edition, Vol. 2, pp. 691-716). Hoboken, NJ: Wiley.

**NOTE: When we recorded this podcast, the Handbook was in press. It has since been published.**