# New Book on Latent Class and Latent Transition Analysis

### with Linda Collins and Stephanie Lanza

### Stephanie LanzaLinda Collins

### ****May 5, 2010****

Host [Michael Cleveland](http://methodology.psu.edu/people/mcleveland) interviews [Linda Collins](http://methodology.psu.edu/people/lcollins) and [Stephanie Lanza](http://methodology.psu.edu/people/slanza), authors of the new book, *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences.*

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.

Michael: I'm here today to talk with Linda Collins and Stephanie Lanza about their new book called "Latent Class and Latent Transition Analysis with Applications in the Social, Behavioral and Health Sciences". Welcome to you both, Linda Collins and Stephanie Lanza.

Linda Collins: Thank you.

Stephanie Lanza: Thank you, Michael.

Michael: First I'll just give a little bit of background about the two of you reading from the back of the book. Linda Collins is director of The Methodology Center and professor of Human Development and Family Studies at the Pennsylvania State University. Stephanie Lanza is scientific director and senior research associate at The Methodology Center at Penn State. The book is part of the Wiley Series in Probability and Statistics and was published by Wiley in December of 2009 but has a copyright date of 2010.

 I want to start by just reading a little bit from the back of the book to give the listeners a sense of what the book is about and then we can go from there. The back of the book states "On a daily basis, researchers in the social, behavioral and health sciences collect information and fit statistical models to the gathered empirical data with the goal of making significant advances in these fields. In many cases it can be useful to identify latent or unobserved subgroups in a population. This book, Latent Class and Latent Transition Analysis, provides a comprehensive and unified introduction for this topic."

 Can we start by just talking about latent class analysis in general? Can you perhaps use one of the examples in the book to help those listeners who are a little bit unfamiliar with LCA?

Linda Collins: Maybe we can start by drawing a parallel with an approach that some of our listeners may already be familiar with, factor analysis. Factor analysis is a latent variable model where there are observed variables that are intended to measure an unseen, sometimes called unmeasured, latent variable. In factor analysis, the latent variable has a continuous distribution typically a normal distribution. In latent class analysis, that latent variable is categorical.

Stephanie Lanza: It might help if we draw our attention to an empirical example that is presented in the book in Chapter 5. We downloaded data from the Monitoring The Future study from the year 2004 which is a data set about the health and behavior in high school adolescents. We looked at the sample of high school seniors and wanted to identify groups of individuals based on a whole range of health behaviors. We looked at whether or not they reported eating breakfast regularly, eating green vegetables or some fruit, whether they exercise vigorously and whether they reported getting at least 7 hours of sleep each night.

 Based on their responses to these 5 different health behaviors we were able to identify 5 underlying types of high school seniors in the United States. One group was labeled "the unhealthy adolescents" who were unlikely to report any of these 5 behaviors. The second latent class we labeled "sleep-deprived". This group was characterized by pretty typical behaviors of adolescents but were unlikely to report getting enough sleep. This class was quite a bit more prevalent among females.

 A typical latent class emerged which was a large portion of the male population and a smaller portion of the female population. The fourth group was "the healthy eaters" latent class. Finally we have a "healthy" latent class which of course we were happy to find in this sample which is more likely among males. In addition to having the healthy eating behaviors, they also were likely to report eating breakfast everyday and exercising vigorously and getting at least 7 hours of sleep.

Michael: Can you just maybe give us a more general sense of what LCA is doing to get those groups?

Linda Collins: Well, it's based on the likelihood of a particular group member responding in a particular way to one of the questions. For example the healthy eaters group were people who were very likely to say that they ate breakfast nearly every day or every day and were pretty unlikely to say that they never ate breakfast. You can look at that pattern of the likelihood of responses across the various questionnaire items. That's what distinguishes one group from another.

Michael: A follow-up question to this example here, are there any other interesting analysis that you did with this particular example that the listeners might want to be aware of that you could do with a typical LCA approach?

Stephanie Lanza: One interesting result that we found was that maternal education was highly related to membership in these different health behavior groups. However that association varied across gender. What we found was among female seniors, higher maternal education corresponded to an increased odds of membership in both the healthy eaters and the overall healthy latent classes but that relationship didn't really exist for males. The association we expected we only found among females.

Michael: To go back to my original question which was to use an example to introduce the listener to LCA, I noticed when thumbing through the book that there are quite a few other empirical examples used. I'd like to spend a few minutes talking about those examples and wanted to know first of all how did you come up with so many examples and what were you looking for when you were looking for examples to use in this book?

Linda Collins: Well one thing we wanted was examples that came from publicly available data sets. We didn't use any examples from for example work with collaborators where the data belonged to a project and aren't on a website that can be downloaded. That was one criterion. We were interested in examples that told a story, that where something interesting emerged from the data.

Michael: I also noticed that a lot of the examples are what can be termed let's say messy a little bit.

Linda Collins: Yeah. A lot of them are messy. We wanted to use examples that are like the kinds of messy examples that we come across in normal data analysis because we figured that that's the kinds of analyses that our readers will be doing. We didn't want to convey a false sense of what it's like to be doing latent class and latent transition analysis on empirical data.

Stephanie Lanza: We are hoping that many of the readers of our book are actually social, behavioral and health scientists and actively conducting a research in their areas. Their data is messy I guarantee you. We wanted to provide a treatment of latent class and latent transition analysis that was accessible to these scientists so that they can see how we would work through issues that might arise in their research.

 In The Methodology Center we develop software for conducting latent class and latent transition analysis. The main reason we develop software and distribute it on the internet is that we wanted to make user-friendly software that can handle estimation issues that can arise in real messy analysis.

Michael: Right. Once again the idea that this is a realistic portrayal of what a listener who potentially might be using the LCA is going to encounter.

Stephanie Lanza: That's right.

Michael: Very good. Can I ask why now? Why did you write this book and what made you decide to write the book now?

Linda Collins: Well probably like most people we talked about it for a while before we finally decided to write it. We both felt that there was a need for first of all a book that was comprehensive, that went from a very simple and standard latent class models all the way through to a much more complex latent class models. We also felt that there was a need for a book that was technical in the sense that it didn't back away from some of the difficult technical aspects of latent class analysis but also explained things in a conceptual way.

Michael: What audience is this book aimed? I know you mentioned a comprehensive book that covers both the technical and the conceptual details of LCA and LTA. Is this book appropriate for users or listeners who have never used LCA? Is it appropriate for people who have been using LCA or LTA for several years?

Stephanie Lanza: We really tried the cover the range in this book. If you were to sit down and read chapter 1, the general introduction, that chapter would be accessible to someone who has absolutely no familiarity with LCA or related statistical methods. We think this book would be ideal for graduate students, for post-doctoral researchers, for scientists out there, professors who are conducting research on messy data who believe that there may be some underlying subgroups in their data.

Linda Collins: Even someone who is pretty familiar with latent class analysis may not be familiar with latent transition analysis. That would potentially be another reason for taking a look at the book.

Michael: Very good, very good. Speaking along those lines of different types of users and different levels of experience, how technical is the book?

Linda Collins: In writing the book we tried not to back off from technical details. They're in there but we also tried to use the examples to explain things conceptually as well.

Stephanie Lanza: I would actually recommend to readers to read at the level that they're comfortable with. If you come from an applied background, read the general introduction and read the empirical examples throughout and you will get a very good understanding of what these methods are, what the parameters are and how they can use this in their research. That said if you come from a more statistical background and can understand the formulas for many readers that would be the place to start and go from there.

Michael: I want to spend the next few minutes talking about the factual content of the book and get into the details of each chapter or at least maybe some of the major points in each chapter. You've divided the book into 3 major sections. The first part is titled fundamentals. It includes the first 4 chapters. It begins with a general introduction to latent class analysis that we've already covered a little bit but I like to go into a little bit more of the details in these first few chapters. Can you walk us through some of the highlights of this first section?

Linda Collins: Chapter 1 provides a very basic general and very conceptual introduction to both latent class analysis which is for cross-sectional data and latent transition analysis which is for longitudinal data. In that first chapter we don't really get into anything technical at all. We're just trying to get things across in a very conceptual way as a foundation for the rest of the book.

 In Chapter 2 we get into the mathematical model, notation and assumptions and we present some additional empirical analysis. In Chapter 3 we talk about the relation between the latent variable and the observed variables. There are some material in Chapter 3 that I've never seen anywhere else. We talk in very conceptual terms about how to interpret the overall pattern of results in a latent class analysis.

Stephanie Lanza: In Chapter 4 we go into the more technical details of latent class analysis. Here I'm talking about the mathematical model that's underlying a latent class model and what those parameters are but we also got into those fun and messy issues of model selection, model fit and model identification. If anyone has attended a latent class model in the past they're familiar with these issues.

Linda Collins: It's the messiest part.

Michael: Once again that idea of the realistic portrayal of data.

Stephanie Lanza: Absolutely. For example in our 5-class model of positive health behaviors among the high school seniors, we started with a data set of only 5 variables but each one had 3 response categories whether the adolescents never or seldom engage in that particular behavior or sometimes they did it or they did it basically every day. In this relatively simple example with only 5 variables we had a contingency table of 243 cells. You can imagine that not all the cells were well represented in our data set.

 Section 2 was about advanced features of LCA. We cover in Chapter 5 incorporating a grouping variable into latent class models. In Chapter 6 we go into incorporating co-variants into the model so that you can predict who may be likely to belong in the different latent classes. This is where I think a lot of the scientific questions really get addressed.

Michael: Before we get to the final part 3, I wanted to back up a little bit and talk about part 1. You didn't say much about model selection but I know that's a big topic for lot of listeners I'm sure out there that have ran these models and estimated LCA before. What can you say about model selection?

Stephanie Lanza: What's the bottom line?

Michael: What's the bottom line? Is there a bottom line?

Stephanie Lanza: One point that I hope we made in the book is there's not one true model. There's no one latent class model that represents the underlying groups that actually exist in your population.

Linda Collins: I think it was G.E.P. Box, the famous statistician who said "All models are wrong, some models are useful." I think that's kind of our perspective on this. We believe that models are not so much an expression of the truth in some absolute sense as they are a lens by which you can look at your data to reveal interesting things.

Michael: That's a very nice quote.

Stephanie Lanza: I like that too.

Linda Collins: I hope it was box who said it. I better go look it up but I think it was.

Stephanie Lanza: That's right. We would never claim that there are 5 types of adolescents in the United States.

Linda Collins: Right, only 5.

Stephanie Lanza: They are either unhealthy or they're sleep deprived or typical or a healthy eater or a healthy adolescent. That's not at all the case. However LCA was a great tool for us to take a lens to these data and say "Together these 5 subgroups can really characterize the multifaceted health behaviors in high school seniors today."

Michael: It's very useful to start at least with this.

Linda Collins: Well, yes. These 5 latent classes is still the essence of 243 cells in a contingency table.

Michael: Good point.

Linda Collins: It's much more meaningful to think of these 5 latent classes and then try to use co-variates and grouping variables to explain the 5 latent classes than to stay with the 243 cells and try to make sense out of that.

Michael: Right. No one can do that. That's impossible.

Linda Collins: No, no.

Stephanie Lanza: Michael getting back to your question about model selection ...

Michael: Yes. That's great.

Stephanie Lanza: We didn't mean to dodge it.

Michael: Of course. Thwarting the issue.

Stephanie Lanza: Yes, but I guess the point I wanted to make is there's not one right answer and every other answer is wrong. It is a bit of an art and I think it's important to keep that in mind so that the different fit indicators that you can get in LCA and LTA will help you and guide your choice but that at the end of the day it has to be an interpretable solution.

Michael: Just a reminder to our listeners LCA stands for latent class analysis, LTA is latent transition analysis which is covered in part 3 of the book Latent Class Models for Longitudinal Data. Can you just give us a few ideas of what's covering those last 3 chapters of the book?

Linda Collins: Right. Latent transition analysis is a unique feature of this book. Latent transition analysis as we mentioned before is an extension of latent class analysis to longitudinal panel data. With LTA you can model movement between latent classes. There's an array of latent classes at each time but people may not be in the same latent class, in fact probably won't be in the same latent class across the various times. You estimate what's called a transition probability matrix that expresses the probability of changing into a latent class, a time to say conditional on latent class membership at time 1.

Michael: Stephanie, can you maybe give an example of application of LTA?

Stephanie Lanza: Sure. In Chapter 7 Linda and I present an example of sexual risk behavior among late adolescents and early adults in the United States. In this model we were considering different aspects of sexual behavior including the number of dating partners, of whether or not the individual reported having sex, the number of sexual partners that they had and also a measure of possible exposure to STIs. We felt that this would encompass a broad range of sexual risk behaviors going from the very early dating behaviors to the more high risk multi-partner unsafe sex.

 In this example we found 5 latent classes of individuals. We labeled them non-daters - as you can imagine they were not even dating yet and we're not having sex, a latent class of daters, a monogamous latent class that was likely to be having intercourse with just one partner and then we had two multi-partner latent classes - one that was engaging in safe sex every time and the other that wasn't. We labeled them the multi-partner exposed latent class.

 The most interesting part of this study was that we looked at this behavior longitudinally. Over the course of 3 years we found that one group of these adolescents was at high risk for transitioning into the multi-partner exposed latent class. That was the monogamous class.

Linda Collins: Well I think it's interesting that people didn't transition so much from multi-partner safe to multi-partner exposed.

Stephanie Lanza: That's right, not at all. Part of what's going on there is that these individuals in the monogamous latent class, they were quite likely to not be using condoms regularly. Again they were with just one partner. It doesn't seem to be as much of a concern but then when they transition into having multiple sexual partners, in that setting they're also not using condoms. I think that that's pretty interesting.

Michael: Continuing with this example going back to maybe some of the other chapters in part 3, can you also do some of the more advanced techniques that you described with latent class analysis adding multi-group analysis or co-variates to the LTA model?

Linda Collins: Yes. You can use both the multiple groups approach and you can introduce co-variates with LTA. Both of those approaches are explained in chapter 8.

Stephanie Lanza: We demonstrate both of those approaches using the sexual risk behavior example. We found that the prevalence of 3 of the latent classes was equal across genders. That's the non-daters, the daters and the high-risk class, the multi-partner exposed.

Michael: That's interesting.

Stephanie Lanza: It is. It gets even more interesting when you look at the monogamous and the multi-partner safe class. The females are more than twice as likely to report being a monogamous latent class whereas the males are twice as likely to report being in a multi-partner safe latent class. Now when I present this to groups they often ask "Do the females think they're in monogamous relationship?" The males think not and report truthfully.

 We also took a look at how substance use behaviors are associated with high risk sexual behavior. We found that while cigarette use was mildly related, it was drunkenness and marijuana use that were both really strongly associated with high risk behavior.

Michael: Stephanie, at the very beginning of the podcast you mentioned a little bit about PROC LCA and the software that's distributed thought The Methodology Center website. Did you use PROC LCA to run all these analysis, estimate these models in the book?

Stephanie Lanza: We did. In The Methodology Center about 2-1/2 years ago we began releasing a software for latent class and latent transition analysis as SAS procedures.

Michael: Of course they can fit these models using other methods and other software as well?

Stephanie Lanza: Yes. Available software includes Latent Gold and Mplus. There are a few others out there as well.

Michael: For the readers who want to try to replicate these examples and whether they're using PROC LCA or Mplus, Latent Gold whatever. How can they go about going through the book and try to replicate example by example the results that you found?

Linda Collins: Well we set up a website with some supplemental material on it. On that website is information about how to obtain the data that we used from the publicly available data sets on the web. Most of the examples are based on subsets of data. We explained how to get the subsets you would need. Then we also include the PROC LCA or PROC LTA code so it would be possible for readers to try to replicate the examples in the book.

Michael: Great. The website that we've been referring to is methodology.psu.edu/latentclassbook and Latent Class Book is altogether there with no spaces. How do listeners actually order a copy of the book?

Linda Collins: Listeners can order a copy of the book from the Wiley website or from Amazon.com or other online book vendors.

Michael: Well Linda and Stephanie, it's been very fascinating talking about the book today with you and getting into the details of LCA and LTA. I'd like to end this with just a little bit of a maybe a more personal question or two. What was it like to write this book? What was the writing process like and your collaboration together?

Linda Collins: Fortunately Stephanie and I have been collaborating for a long time so there weren't a lot of surprises. The way we divided up the work was Stephanie amazingly found all the great examples in the book and did the model fitting for those. It was usually my job to write a first draft of every chapter with a couple of exemptions. Then Stephanie's job was to give me comments that were about as long as the first draft of the chapter in most cases.

Stephanie Lanza: Then we followed that up by intensive rating retreat meetings where we would come together and polish it out.

Michael: How long did the process take from beginning to end?

Linda Collins: About a year and a half.

Stephanie Lanza: There was a maternity leave too.

Linda Collins: That's right. There was that, yes.

Stephanie Lanza: About a year and a half.

Linda Collins: Yeah yeah, roughly. We also wrote the book in LaTex. I don't know how many of our listeners know what LaTex is. It's a freeware word processing software program that is simultaneously extremely versatile and extremely maddening. It was an immersion program for me in LaTex but the nice thing about writing a book in LaTex is that you type, send it as you go but no, we never slowed each other.

Michael: You're still here together talking.

Linda Collins: That's right. We're still speaking.

Stephanie Lanza: We still work together.

Linda Collins: We still work together. No, it was really a great experience writing the book. I've worked in the area of latent class and latent transition analysis for a long time. I feel like I learned a lot by working on the book. When you're writing a book you find yourself questioning things. You write something and you say "Do I really mean that?" and then you end up going off and learning a lot about it. It was some really good experience in a lot of ways.

Stephanie Lanza: It really was. When you have to put something down on paper that you think you know, it's amazing how much you realize you don't know and we had to learn but so many times, I can't count how many times both Linda and I said "I could never have written this book alone."

Linda Collins: Absolutely, no.

Stephanie Lanza: I don't believe it.

Linda Collins: I don't think either one of us could have written it alone. It was a real team effort.

Michael: Great. Thank you both for being with me today.

Linda Collins: It's been a pleasure.

Stephanie Lanza: Thank you for having us, Michael.

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.