# Latent Transition Analysis: Video 2 - Measurement Models

## Transcript

Video: <https://youtu.be/KyCiykheghs>

Oliver Perra: This is the second presentation of my introduction to latent transition analysis. In the first presentation, I highlighted that latent transition analysis is a person-centred approach applied to repeated measures and longitudinal data, so it allows to identify the categories of individuals that make up a sample at different timepoints and how those categories can explain the patterns of behaviours we observe at different timepoints. But also it allows to investigate the structural relationships between those latent categories or latent classes, so how individuals may change or not across development and across time and how they may transition from one category of behaviour to another.

In this second presentation, I will provide more details on how to conduct latent transition analysis and I will follow this outline, which I have taken from a chapter I published in the paper linked here. I will cover the first three stages of this outline, focusing particularly on issues concerning the selection and specification of measurement models for each age or each timepoint of data collection. So I will start with a similar example of fictional data that I provided before where we asked adolescents to report a frequent use of these substances when they were aged 14 years and again when they were aged 15 years of age. I will emphasise here that in these and other examples, I refer to categorical indicators, in this case dichotomous variables. However, latent class analysis and latent transition analysis can be easily applied to any type of indicators from nominal to count data to continuous variables.

So we have collected this data over two timepoints, so we will observe associations between the indicators and the associations will involve associations between the indicators at each timepoint and longitudinal associations between the same indicators at different ages. The first stage of latent transition analysis is to find person-centred measurement models at each timepoint, that is latent classes that can explain the variability of behaviour patterns at each timepoint. So for this reason, we start by looking for the best latent class models at each timepoint separately, and to indicate this separation, here I have added a grey line that divides the two timepoints.

So latent class analysis allows to select a number of latent classes and the line subgroups that can explain the variability in behaviours we observe in the sample. And if you want to learn more about the interclass analysis, please see my resources for NCRM. Note here that when I introduced the latent classes, the associations between indicators disappeared. This is to represent a key assumption of latent class analysis, that of conditional independence. This assumes that the indicators are independent conditionally on latent class membership, that is what explains the associations between observed behaviour is the underlying latent class.

So let's assume in this example that I identified two classes - one of substance use experimenters and one of substance use abstainers. And I just want to remind that latent class models will provide these parameters. Firstly, the item response conditional probability, that is the probability indicated by P to report a type of behaviour in one indicator, for example alcohol use, conditional on latent class membership. For example, the probability of reporting use of alcohol for people in the experimental class, and so on. And in the outputs, these probabilities will be reported in tables like this one where we see the probabilities of answering different ways to the different items for people in different classes. Note that by inspecting these probabilities, we can interpret the meaning of the latent classes. For example, we see that individuals in latent class one have higher probabilities of reporting alcohol and cannabis use, and those in latent class two have low probabilities of reporting use of these substances.

The other parameters provided by latent class analysis are the prevalence of the latent classes, so how many individuals are likely to be in latent class one and in latent class two. Note that these are probabilistic. In fact, latent class analysis also provides the posterior probabilities of latent class membership, that is for each individual in the sample, latent class models estimate the probability that that individual belongs to latent class one or latent class two. So an individual is assigned to a latent class with uncertainty. For example, in the table here, you can see that individual 103 is assigned with latent class two, but this assignment is very uncertain compared to others. And latent class models vary in the degree of uncertainty. Some models may provide classifications that are more certain and precise, but not perfect. And the statistic entropy provides a measure of the classification precision. Entropy varies from zero to one where one indicates a certain classification.

[0:06:23]

So spotlight on age 15 now. And I assume that, at this stage, there is more variability in behavioural patterns we observe. Therefore, we might identify an additional class of substance users, that of abusers. And how should we decide the number of latent classes underlying behaviour patterns at each timepoint? And there are a number of statistics we can use, and I talk about this in my resources on latent class analysis. The main point is that, similar to structural equation modelling, there isn't a single statistic we can use, so we need to consider different statistics and different indicators. For example, statistics that compare the actual distribution with the distribution associated with the latent class model, we can use information criteria to decide which is the optimal latent class model. We can consider entropy. And it's also important that we consider theory and substantive knowledge to check the plausibility of the solutions, latent class solutions we identify.

But one important point concerning latent transition analysis is that, since there are different statistics to consider and often they don't agree on the best model, it is always advisable at this stage of latent transition analysis to consider a pool of plausible models to be investigated in other stages of latent transition analysis. So the selection of latent class models at each timepoint may be informed by how these models perform in other stages of the analysis. So when we are applying measurement models to different timepoints, we may also consider tests of measurement invariance of these models.

So let's consider a simplified example here, and I assume that there are two underlying classes that explain the observed pattern of behaviour at age 14 and age 15 years. These classes also appear similar - they seem to have the same meaning - and I've called them a class of experimenters and a class of abstainers. We can regress the latent classes at age 15 on those at age 14 using multinomial regressions. However, since the two classes appear to be similar, can we plausibly assume that they have the same measurement parameters at both timepoints? And the measurement parameters in latent class analysis are the item response conditional probabilities. In this example, we can see that they are not the same, but both two classes at each timepoint represent individuals that in latent class one report higher probabilities to use substances and in latent class two report low probabilities of not using substances.

So the question is can we assume that these differences may be trivial and therefore assume that the associations between latent classes and the indicators are the same across timepoints? For example, can we constrain the conditional probability of reporting alcohol use for someone in latent class one to be the same at age 14 and age 15? And, similarly, can we constrain the conditional probability of reporting alcohol use for those in latent class two to be the same at age 14 and age 15 years? So we can impose these equality constraints and ask whether those are plausible, whether those assumptions of those constraints are plausible. So once I have imposed these constraints, the condition item response probabilities are the same across time.

And this model with equality constraints is attractive because, first, we need to estimate fewer parameters once we constrain measurement parameters to be the same across ages. And, second, this also facilitates interpretation of the model because the latent classes have the same meaning. But is a similar model with equality constraints plausible? Since the model with equality constraints is nested within the model where measurement parameters are freely estimated, we can compare the two models using a likelihood ratio test which will provide a formal test of the null hypothesis that the two models provide the same fit. And in the exercise and the additional material to these resources, I provide more examples of how to run this test.

[0:11:43]

So here I'm providing a more complex example where, at age 15, I identified an additional latent class. But still we might consider plausible tests of measurement invariance if we identify classes that are similar across time. For example, I may identify a class of abstainers and, interesting, testing whether the measurement parameters of these abstainers are the same at age 14 and age 15. So it's possible to test models where there is partial measurement invariance across time or at least in measurement in various four classes that appear to be similar across time. But because there are many different models of partial measurement invariance, it's important that those tests and those models, the definition of these models are also guided by theory and substantive knowledge.

The first stage is to extract latent classes parameters, and the first stage is, basically, taking the measurement parameters from the latent class models selected for each timepoint and then investigate the association between the latent classes and change across time. So the question is then why do we need to do this, why cannot we jump to investigating the associations between latent classes while, at the same time, estimating the measurement models. But the reason why it might not be advisable to do so can be glimpsed by considering what happens if we also added a distal outcome to the model, for example here educational attainment at age 16 years. If we estimate the latent class models at the same time as the structural association between latent classes and the regression of the distal outcome on latent class at age 15, the latent class model at age 15 is represented not just the covariance of the full substance in the use indicators but also the covariance of the distal outcome with the age 15 indicators.

So this means that, while we may intend the latent class model to explain associations between the substance use indicators, the model is actually explaining the associations between indicators and educational attainment or, in other words, by including a distal outcome, the model changes because the latent class at age 15 is also taking into account other covariances. And this poses theoretical problems in interpreting the latent classes and practical problems when we add a distal outcome, the latent class measurement model at age 15, we change its parameters. And the same problem arises when we include covariates in the latent class and the data transition model.

So it may not be advisable to estimate the measurement model in parallel with the structural relationships, because problems arise as soon as other variables are included. And this problem includes substantive problems in interpreting what we're actually measuring and practical models that become apparent as soon as we have more timepoints, the models are more complex and the models become quite time-consuming to estimate because of all the computational costs in estimating those.

So there is a solution, and the solution is to use a three-step approach that has been recently devised, and I will discuss this in more detail. The advantage of this approach is basically that the measurement models and the structural models are not estimated at the same time, but rather at different steps. The measurement models are estimated and selected in the first step and individuals are assigned to the most likely class based on the posterior membership probability. And after that we can assess measurement error in latent classes, that is the uncertainty in latent class membership. And since this latent class membership is estimated with error and is uncertain to some degree, it is important to control for this uncertainty to avoid biased results. In the first step, we take individuals' latent class membership accounting for measurement errors estimated in the previous steps and we can investigate the associations between latent classes as well as covariates and distal outcomes. But because we are not re-estimating the measurement model at the same time, the first step only concerns associations between latent classes and other variables and not the indicators.

So I will provide a practical example. So let's assume that at each timepoint we have separately estimated these models where there are two latent classes at age 14 and three latent classes at age 15. The issue may be complicated by measurement invariance, and in the exercises I provide more guidance on this. Having selected the best measurement models at each timepoint, we can assign individuals to an underlying latent class based on their posterior probabilities, as you can see in this example with individuals assigned to different classes and the probability of being in that class.

[0:17:52]

So let's focus on age 14. Here we see three individuals have been assigned to latent class experimenters and five are assigned to latent class abstainers, and you can also see the probabilities to be in those respective classes. If we take the three people assigned to experimenters, we can calculate the average probability of being in this class if they are experimental members. So that would be 0.79 plus 0.76 plus 0.86 divided by three. And since these probabilities add to one, conversely the average probability of being an abstainer when people are considered as experimental members is 0.21 plus 0.33 plus 0.23 divided by three. That is 0.26. You can report these probabilities in a table like this one I report here.

And when we consider the number of individuals assigned to each class, we can calculate the classification probabilities of being an experimenter if someone is considered an experimental member. And this is basically the product of the average probability of being an experimenter if someone is an experimental member by the number of individuals assigned to the experimental class. And this is divided by the same product, plus the product of the average probability of being an experimenter if someone is considered an abstainer member by the number of people assigned to the abstainers class. So once we have these classification probabilities, we can use this to calculate the odds of being an experimenter if someone is considered an experimenter rather than an abstainer. And the logarithm of this all represents in a single number the uncertainty in assigning participants to the class experimenter. So how much can we be certain that those assigned to experimenters class are experimenters?

So once we have calculated these odds, we can use the latent class membership as a nominal variable and use it to estimate latent class membership, fixing the uncertainty in this classification to the value of the (inaudible 0:20:31) of being in one class if someone belongs to that class rather than the other class. In other words, we are feeding to the model the information about latent class membership together with the information about the uncertainty in this classification, which is represented by the (inaudible 0:20:49) I calculated before. In this way, the measurement model here is fixed, it is not estimated anew, and we can investigate the associations between latent classes at different timepoints considering the measurement models we have selected as optimal.

So to summarise, I have illustrated the first stages of latent transition analysis. Firstly, we have to investigate the number of classes that adequately explain heterogeneity of behaviour at different timepoints. We can then test measurement invariance between classes when these classes appear to be similar. And once we have selected one or a few optimal measurement models, we can extract the measurement parameters and use the three-step approach to investigate the associations between latent classes across time and also between those classes and other variables, as I will discuss in more details in the third presentation. I have highlighted that the advantages of the three-step approach lie in the reduction of computation in testing latent transition models. And there are substantial advantages in working with a given measurement model that has a given interpretation and is not re-estimated anew with every analysis. So thank you very much.

National Centre for Research Methods (NCRM)  
Social Sciences  
Murray Building (Bldg 58)  
University of Southampton  
Southampton SO17 1BJ  
United Kingdom

**Web** www.ncrm.ac.uk   
**Email** info@ncrm.ac.uk  
**Tel** +44 23 8059 4539  
**Twitter** @NCRMUK