What is Structural Equation Modeling?

What is structural equation modelling? Well I think one of the first useful things to understand about SEM, as I'll refer to it, is it isn't a single technique as such. We wouldn't want to compare it’s to say learning ordinary least squares regression or logistic regression, log linear modeling which, although these techniques have a number of different aspects, we can think of them as, if you like, single approaches to address research questions. I think SEM is much better thought of as a general modeling framework that integrates a number of different multivariate techniques into this overall framework. It is a framework which draws on a number of different disciplines it brings together measurement theory from psychology, factor analysis also from psychology and statistics, path analysis from epidemiology and biology, regression modeling from statistics and simultaneous equations from econometrics, and all these different techniques come together to form structural equation modeling as a general modeling environment. And it's also an environment which is somewhat dynamic. It is not set in stone at this point in time it is actually often integrating new ways of fitting models as the technique develop over time.

What sort of research questions would SEM be particularly suitable for addressing? Well I think it’s being a general model-fitting environment, it can address many different kinds of research questions. But I think it’s particularly suitable in situations where the key constructs, the key concepts that researchers are interested in are complex and multifaceted, often relating to psychological social psychological concepts. So these kinds of concept can be quite difficult to measure and are often measured with error and one of the useful aspects of SEM, as we'll see, is its ability to make corrections for errors of measurement. Other kinds of research questions that SEM is well-suited to are ones which specify systems of relationships rather than, as we may be used to if we're fitting regression models where we have a single dependent variable and a set of predictors or independent variables, structural equation models may have numerous different outcomes or dependent variables, each of which is affecting other dependent variables in a more complex system. So if a researcher is interested in modeling a causal system then structural equation models are particularly suitable. Another kind of research question that structural equation models are often used to address is where the researcher is interested in indirect, or mediated, effects. So in many research questions we're interested in the effect of variable X on variable Y. That would be thought of as the direct effect of X on Y, but in many research contexts we're interested in more complex kinds of relationships where the first variable X perhaps influences a second variable Z which then has a second effect on Y. That would be seen as an indirect effects and SEMs are very well suited to addressing those kinds of mediated research questions.

Now SEMs are known by a number of different names in the existing literature and this can be somewhat confusing. Sometimes they are referred to as covariance structure analysis models. This relates to the fact that with SEMs we're actually analyzing covariance matrices, not variables directly. We will come on to that in later films. They're also known as analysis of moment structures. This is what gives the software the SEM software Amos its name because this is in recognition of the fact that more modern SEMs analyze not just covariances but also means, so higher order moments. It's also know sometimes that LISREL model which again takes its name from possibly the most well-known software certainly the first software for fitting SEMs which is LISREL. More controversially, SEMs have been referred to as causal modeling and they're often, or certainly have historically, been associated with analysis which get a cause and effects but I think that is probably more controversial name to give to any modeling technique because the claims for causal inference will come from the research design rather than the statistical model that we apply to analyze the data.

There are many different software packages that are available for fitting SEMs and this is a list that's changing and growing all the time. As I mentioned the probably best known is LISREL which was developed by Karl Joreskog and Sorbom, one of the first available packages. Now there are many more software packages available Mplus, EQS, AMOS. R is a free package, Stata and many of these packages have more limited versions that are available for free for students to download and try to see which one is most suitable. I wouldn't want to make a recommendation for any particular software package each one has its own particular advantages and disadvantages.

So what is structural equation modeling? Well there are many possible answers to that question. The one that I'm going to propose in this film is that SEM can be thought of as path analysis using latent variables. Now this definition may not be very helpful to you if you are not very familiar with either path analysis or latent variables so for the remainder of the module I'm going to run through what path analysis is and what latent variables are.

So what are latent variables? Well, most of the concepts that we're interested in in social science are not directly observable: things like intelligence, social capital, trust. It's impossible to go and put some kind of meter into people and get a direct reading of their level of social capital or trust, so this makes these concepts hypothetical or latent, as we refer to them. We believe that they are latent within people at some level and they drive attitudes and behaviour, but we can't actually directly observe them. So we're in a bit of a difficult position if we can't measure these concepts that were interested in but fortunately we can use approaches which measure these latent variables using observable indicators, using variables that we can measure directly that we believed to be caused by the underlying latent constructs. So if we think of a questionnaire item, a question in a questionnaire that has been administered to a sample of people, this would be a good example of an observable indicator of a latent construct. So let's imagine that this question asked people how happy they are with their lives on a scale 1 to 10. Now some people will give higher answers or lower answers: there will be variability, variance in this variable across the individuals in the sample. Now we don't think that all of that variability is only to do with people's level of happiness. Some of it will be: so some of the variability will be caused by variability in the true level of happiness across people but there will be other factors that also caused variability, possibly to do with the questionnaire design, the temperature in the room, whether the question is it administered by an interviewer or completed on a computer. These are all other factors that we're not really interested in in what we're trying to measure which is happiness. So some of the variability will be to do with happiness, the latent construct, but some of the variability will be due to other factors: error and unique variance.

So we can summarize these ideas quite simply in this formula, the true score equation, where $X = t + e$. So here the measured variable, the observed indicator, is X and, as I said, the X the variability in X, is comprised of both true score and of error. So the true score is simply where the individual is on a true happiness dimension, their true underlying level of happiness. The error comprises two components: the first is what we could think of a systematic error, this is a bias where perhaps the question is phrased in a way which makes people give higher happiness ratings than their actual level of happiness, maybe it's because it's a question administered by an interviewer and they don't want to seem unhappy because that is socially undesirable. This would be a systematic error. A random error would be one where you're just as likely to overrate as to underrate your happiness. So we can think of the systematic error as being one where the mean of the individual errors doesn't cancel out, it doesn't equal 0, whereas a random error you are as likely to give a higher as a lower score so the expectation would be that the means, the mean of the errors, would cancel out and be zero. So this is all by way of saying that when we measure a variable, when we measure X, ideally what we will be able to isolate would be the t part of the variance, the true score, and to remove the error variance when we're trying to either predict t or use t as a predictor in a model.

So we can now translate this true score equation into a very simple path diagram which is key to representing structural equation models. So here we can see that the X reads over to being the observed item in the rectangle. The t reads over to being that latent variable, the true score in the ellipse, and the e reads over to being the circle at the top of the diagram, the error, and the arrows indicate that the observed score is caused by both the true score, the latent variable, and by other factors, the error, so we can encapsulate those ideas in this simple path diagram.

It would be nice if we could implement this as a statistical model. Unfortunately, when we only have one indicator of the latent variable, this is happiness, then this equation is what we would call unidentified. We have more unknown pieces of, quantities that we're trying to estimate, the t and the e we don't know what they are we would like to estimate them, than we have known pieces of information the X. We've measured X in our sample we have two unknowns and one known so we can't solve that equation uniquely the equation is unidentified, so we can't separate the true score from the error when we only have one measure of the underlying concept. What this then tells us is that we need to have multiple indicators of our latent constructs. When we have multiple indicators then we can start to over identify the true score equation and estimate the quantities of t and e for each indicator. So we can apply many different kinds of latent variable models, we can use principal components analysis, factor analysis, latent class models, depending on the metrics of the observed indicators that we have in our dataset. But what these are all going to do is to provide us with a summary score, a reduced set of factors or components relative to the full set of indicators that we start out with, and in doing that they will correct for the error in each of the individual indicators and give us a better measure of the true score of the concept.

We can represent this simply here with a common factor model. Here we have four measured variables, let's think of these as questionnaire items – again they might be measuring happiness, different aspects of happiness are you happy at home, with your work, with your friends, and so on. So we’ve got four indicators of the same underlying latent variable, happiness. Now because they measure the same thing we would generally expect these variables to be correlated in our population and that's what these double-headed arrows indicate. The curved double-headed arrows indicate that the axes are all correlated with one another, that's one way of representing what's going on here. Another way would be to do away with these correlations and add in the underlying latent variable, someone's true level of happiness which we've here denoted as eta. In this model now we have happiness, latent variable, having a causal effect on each of the indicators and that causal effect is what we can think of as the true score, the t part in our $X = t + e$ equation. Now if that's the case then we also need to have error terms for each of these equations here and that's what we have shown in the diagram there.

So, with these multiple indicators we can apply a latent variable, in this case a factor model, and we can get empirical estimates of these key quantities and here now the lambda coefficients there in this model, we will refer to as factor loadings, and these are the correlation between the factor the eta and each of the X variables. Now if these are good, if these indicators are good indicators of happiness, we would expect these correlations to be high. We would expect the correlation between a good indicator of the latent construct and the latent construct to be close to approaching 1.

So, if we are able to measure our constructs with multiple indicators we can apply latent variable models and this brings a number of benefits. Well, firstly the kinds of things that we're interested in modeling in social science are generally complex and multifaceted. If we think of happiness, for example, it's difficult to come up with a single question which covers all aspects of a person's individual well-being so we probably need to have multiple indicators to get a good coverage of the concept. As I mentioned, it also enables us to remove, or least reduce, random error in the construct that we are measuring. This, I think we can convince ourselves, that removing error seems to be a good thing to do. But, more formally, we can demonstrate that if we have random error in the dependent variable, although it leaves the estimates in a model unbiased, these will be less precisely measured: there'll be a noisier measure with wider confidence intervals. More seriously perhaps, if we have random error in independent variables then regression coefficients that we estimate using those independent variables will be attenuated, they will be smaller than they are in the population, systematically smaller, tending toward zero. So we will underestimate effect sizes and we will falsely fail to reject the null hypothesis

So what is path analysis? Well, again, there are many ways that we can answer this question but I think a key feature of path analysis, and one that makes it very appealing as part of structural equation modeling for social scientists, is that the model that you're wanting to fit to the data is represented diagrammatically rather than in the form of equations. Of course we can represent the structural equation model as a system of equations, but we can also represent it as a diagram and this visual aspect again is very appealing for social scientists who are perhaps less comfortable and less intuitive in their reading of equations. So the standardized notation of path analysis is a very important feature. The path analysis presents regression equations between our measured variables, so we're interested again in kind of systems of relationships between multiple observed variables. Now that's important that I'm saying observed variables there, because in a standard path analysis we would not be using latent variables, but variables which are directly observed – again perhaps single questionnaire items, other kinds of measures. A third key feature of path analysis is its focus not just on direct effects but also, as I was talking about earlier, on indirect effects and total effects. So for research questions where we don't have a simple linear model, where we 're estimating the effects of some set of predictor variables on an outcome, a dependent or a criterion, but we're interested in the pathways between multiple independent variables and possibly multiple dependent variables.

So in this slide I'm presenting some of the standardized notation, the way that we represent different parts of the model using diagrammatic notation. We can see at the top a measured latent variable, so a latent variable will be presented as an ellipse. An observed or manifest variable such as a questionnaire items, that we might use as an indicator of a measured latent variable, would be a rectangle and error variance or disturbance term is a small circle and there's a similarity with the measured latent variable: they are both circular shaped because an error variance is also a latent variable – it's is just that we are not specifying it as measuring anything in particular. It is the, what's left over, the residual or disturbance term. A covariance path, where we're specifying that two variables in the model are related or correlated with one another, would be represented as a curved double-headed arrow. This is a non-directional association i.e. we're not specifying there is any causal link from one variable to another but we want to indicate that they are correlated. And finally the single headed straight arrow represents a directional path or what we would generally think of as implying causality in the model a regression path from one variable to another.

So here are some examples of some simple path diagrams that we could represent in equation form or using standardized path notation. In this simple diagram we can see that the variable X has a causal effect on Y and the D term there is the disturbance term, so the error term in this model. We could, this is essentially a bivariate regression model. We can also write this in that standard equation notation. This second path diagram is somewhat more complicated but really is just adding in a second independent variable X2, so again this is equivalent to a multiple linear regression with two independent variables, a dependent variable Y and an error term which in this path diagram is labelled D for the disturbance term.

Now, as I mentioned, one of the things that path diagrams, path analysis are particularly useful for is for studying not just direct effects but also indirect effects. We can see now that we've introduced a more complex relationship between these variables where X1 has a direct effect on X2 but X2 also has a direct effect on Y, so we now have an indirect effect of X1 on Y through X2. And we can use standard formulae to decompose these regression coefficients indicated by beta1 to beta3 into the direct indirect and total components. So here beta1 represents the direct effect of X1 on Y, beta2 is the direct effect of X1 on X2 . beta3 now is the direct effect of X2 on y and beta2 times beta3 will give us the indirect effect of X1 on Y. And we can also compute from this path diagram the total effect which is the sum of the indirect and the direct effects between one variable and other. So if we take the sum of beta1 and the product of beta2 and beta3 this will give us the total effect of X1 on Y.

So that's given a very brief overview of both latent variables and path analysis and what I'm encouraging you to think about, to understand what we're doing with structural equation models, is that when we have a path diagram that includes latent variables rather than just observed variables, as we can see in this diagram, then we're representing a structural equation model.