In the first two videos on structural equation models I have covered some of the sort of conceptual background the history some of the key ideas.

In this video we moved to understanding some of the applications of the actual model fitting that go on in structural equation modeling and this focuses particularly on confirmatory factor analysis.

So in this video I'm going to talk about the general idea of how we measure concepts using latent variables and I'm going to contrast two approaches to using latent variables to measure concepts. The first is the more conventional historically the main way of doing this using exploratory factor analysis and I'm going to contrast this then with the more modern approach of confirmatory factor analysis. I will then move on to talking about some of the ways that we go about actually fitting in and estimating confirmatory factor models and some of the important procedures that we have to do. And I'm going to finish off by talking about some of the kind of extensions that we can take CFA into notably when we are modeling the means of latent variables as well as their relationships their associations. I'll talk about the difference between formative and reflective indicators in CFA a procedure called item parcelling and also the situation which we may sometimes be interested in of a fitting a factor model to variables which are themselves latent variables rather than to observed variables which is the usual case and that would be called a higher-order factor model .

So in the first video I gave a sort of pithy definition of structural equation modeling as being path analysis with latent variables. We can also think of this is as really being a distinction between two stages or two parts of the modeling process the first is where we want to get good measures of our concepts or constructs and in the second part is looking at the relationship between those measured constructs so there is a, if you like, the emphasis firstly on measurement and measurement accuracy and adequacy and then secondly moving on to look at the structural relationships between the constructs that we've measured.

So again we saw in the first video that anytime we want to measure something in science and particularly in social science is that the measurements contain various kinds of error that error can be random and/or systematic so what we want to do in our statistical approach to the data is to isolate the true score in a variable and remove the error and this is really what we're trying to do using latent variables for measurement. So we want to decompose our X variables X is what we have actually measured and we can decompose that into the t and the e components the t is the true score and the e is the error and we need some kind of model to enable us to split the X into these t and e components.

Now one quite straightforward and useful way of doing this is simply to add the score across a number of different X variables if we have say four variables which are all measuring the same underlying concept then we could just add those up and take a summed score and this has some benefits because the random error in each of those measurements will cancel out as we add items together but it's rather unsophisticated approach and in particular it gives equal weight to each item in the construction of the true score and that's often something that we don't want to do so in other approach is to actually estimate some kind of a latent variable model. Now in understanding the ways that we do this in SEM it's useful to sort of go back in history if you like and think about an earlier approach to estimating latent variables and this isn't to say that exploratory factor analysis is no longer used of course it is but the more modern procedure of confirmatory factor analysis has some attractive property shall we say compared to EFA. So the exploratory factor model is also referred to as the unrestricted

factor model or an unrestricted factor analysis because as we'll see when we get to to looking at CFA, CFA does place restrictions on the variance covariance matrix whereas EFA doesn't do this. EFA or principal components analysis is a similar technique that finds the factor loadings which best reproduce the correlations that are observed between the observed variables in our model. So let's say that we have six questionnaire items that all measure more or less the same thing they're intended to measure some concept that we're interested in and EFA I will simply kind of reorder the data in a way which has best accounted for the observed correlations between those variables now it does this in a way of producing a number of factors which are in EFA equal to the number of observed variables that we have so this is really just a reordering of the observed data. We end up with the same number of factors as we have observed variables so at this point in just this reordering EFA hasn't done very much in way of summarizing or simplifying which is often what we're trying to do with latent variable model. So we have the same number of factors as we have observed variables and all the variables in our model the observed variables are allowed to be correlated with all of the factors.

Now we need to get from this point of having the same number of factors as observed variables to retaining a smaller number so that we're doing some job of summarizing rather than just transforming the observed relationships and there are different rules for doing this. One kind of heuristic judgment would be to keep or retain a number of factors which is less than the number of observed variables that explains some satisfactory amount of the observed variance so we might say will retain as many factors as are needed to explain seventy percent of the variability of the correlations between the observed variables. Something else that we have to do in in addition to summarizing is to understand what the factors that are produced by the factor analysis what they mean what are they measuring now we do this by looking at the pattern of factor loadings between the factor and the observed variables so we do this in a sort of inductive way we work out what the factors are by looking at how they are related to the observed variable.

And another thing about exploratory factor analysis is that there is no unique solution where we have more than one factor and so we can rotate the axis of our solution in ways that can help us to see what the underlying structure is. So rotation of axis in exploratory factor analysis is quite common.

To give an example of what I mean by some of those previous points here is some made up data and we have nine observed items and these are you like knowledge quiz items that have been administers to a sample of children and what we're measuring is some construct like intelligence or cognitive ability. Now if we were to apply an EFA or principal components analysis to this data then we would initially have nine components or factors which is the same number as the observed items so the first thing that we would need to do would be to implement some judgment about how many factors to retain now in this case you can see that three factors have been retained in this model and that may have been based on one of these heuristic guides around amount of variance explained or some kind of plot like a scree plot. So once we've done that we want to know what each of these three factors is actually measuring and we do that by looking at the pattern of correlations and that's what are in the rows and columns of this table between each factor and each and the set of items. So if we look first at factor 1 we can see that the factor loadings or the correlations are high between factor 1 and the observed items which are measuring mathematical ability so this is saying that if you have a high score the higher your score on factor 1 the more likely you are to get the item math one correct there's a high correlation between your score and the factor and

your score on the item. The factor 2 there are high loadings on the visuospatial items and low loadings on the other items and for factor 3 we see this other pattern where is the verbal items that have a high score and low scores on the other. So we did this inductive process of figuring out what the factors are measuring by looking at the correlations between the factors and the observed variables once we've retained a smaller number that we think is in some ways satisfactory. So this is a very useful procedure and has been widely used in social science for many decades but it does have some limitations firstly EFA is an inductive is rather a theoretical procedure and that is something which in general we are less happy with in terms of the way that we build theory in quantitative social science so we've got a situation where the data is telling us what our theory should be when generally we would prefer to do that the other way round we would have a theory and tested against the data. Another unattractive property of EFA in similar techniques is that it relies on a subjective judgment and heuristic rules about what's a large amount of variability to explain and so on so there is a lot of room for subjectivity in determining what our model should be and of course when we are analyzing data of this nature where we have indicators of underlying concepts it's rarely the case that we have no theories at all about which concepts the different indicators are actually measuring we've usually written the questionnaire indeed with a specific intention of measuring particular concept. So actually the more realistic and accurate assessment of what's going on here is that we're starting with a theory and then we're assessing it against the data that we've collected. So the idea that we are going from the data to the theory is not generally an accurate representation of how this procedure actually works. So given that that is the case given that we do have a theory about how the indicators are related to the concepts it's better to be explicit about that from the outset and then use statistical tests of those theories of measurement against the sample data that we've collected.

So we can compare this approach of exploratory factor analysis with a confirmatory approach so confirmatory factor analysis is also referred to as the restricted factor model because unlike EFA it places restrictions on the parameters of the model it can't be therefore rotated. You cannot rotate the solution there is only one unique solution for the CFA. And the key difference now with CFA to EFA is that we specify our measurement model before we have looked at our data and this is sometimes referred to as the "no peeking" rule. If we have a theory about how the indicators are related to our concepts then we should set that down a priori as our theory and then test it against the data rather than tweaking our theory as a function of the particular sample data that we happen to have. So when we do things in this way in a confirmatory way the key kinds of questions that we have to answer are which indicators measure or caused by which factors which indicators measure or have caused by which factors and importantly and this is the real distinction with EFA is which indicators are unrelated to which factors. Remember in an EFA we say that every variable is related in some ways is allowed to correlate with every factor. In CFA that isn't the case we will say that the correlations of the covariances between some of the indicators and some of the factors are zero. We will make that as a parameter restriction. And we will also need to answer questions about the correlations between the factors rather than leaving that as a default assumption in the model.

Here we have six observed variables x1 to x6. Now the first part of the model will have produced six factors or components so at this stage we've already retained just the two factors that we think explained enough of the variability between our observed variables but what you also see here still is that there is a single headed arrow running from each of the two latent variables e1 and e2 to all six of the observed variables. So we are estimating a correlation between each factor and each of the observed variables. Now what we would be

looking for in this kind of situation is that some of those loadings would be large and some of them would be close to 0 so if we look at e1 for example we might in an EFA context hope or expect that the loadings between e1 and x1 to x3 would be high say points 7 or above in standardized form or and the loadings that run from e1 to x4 to x6 would be close to zero and the opposite would apply for e2. So what we're doing there is a say estimating all of those relationships and expecting some pattern of high and low loadings between them. By way of contrast the same variables and the same two factors now in the form of a confirmatory factor model now rather than having estimates for all of those relationships between e1and x1 to x6 and e2 and x1 to x6, we say that there is no relationship between e1and x4 to x6 there's no arrow pointing from e1 to any of those observed variables and the same for e2 there is no arrows pointing at x1 to x3. So the fact that there isn't a arrow there means that in our model we are constraining those to 0 we're not just estimating them in saying are they nearly 0 we are specifying our model a priori to say that those paths are indeed 0.

So those are the kinds of parameter constraints and parameter restrictions that I was referring to in talking about in video 2 that it's quite unusual in other branches of statistics that we use in social science to make these constraints and fix parameters to particular values but that's why we call the confirmatory model the restricted factor model because we place restrictions on the loadings. So sometimes as I just gave an example of we will fix particular parameters to 0 for indicators that do not measure or do not influence a measured variable and the important thing to understand is that our theory of the measurement of our concepts how we think the concepts are related to the indicators that we've selected and written if there are questionnaire item that that theory is expressed in the constraints that we place on the model. So we're not just estimating everything but we are placing restrictions on what the parameters the values that the parameters can take and those restrictions those fixing of parameters they over identify the model so we are placing restrictions which give us more degrees of freedom in our model which enable us in turn to test the fit of our model compared to the matrix that we've actually observed S the sample variance covariance matrix.

Another way that we apply restrictions to the parameters in a confirmatory factor model is to give the latent variables a metric. Now what I mean by that is that if we have a measured variable we will have specified some kind of scale for respondents to answer on so maybe it would be strongly agree is the value 1 and strongly disagree is the value 5 so the scale is 1 to 5 for that measured variable. For latent variable we don't have any metric it is an unobserved variables it's a hypothetical variable so it doesn't have a metric on its own we have to give it one. And there are two ways that this can be done the first is to essentially producer standardized solution so that all variables are measured in standard deviation units this can be done by constraining the variance of the latent variable to 1. And this has some benefits but the downside of course is that we no longer have an unstandardized solution if we require all latent variables to be measured in standard deviation units than they don't have any retention of the unstandardized metrics that they could be given.

So the second approach is to constrain one of the factor loadings to take the value 1 and by doing this we take the scale from that particular item which will call the reference item so if we fix the factor loading of a particular item to 1 then that will be the reference item and the latent variable will have the same scale as that item so if it's measured again on a 1 to 5 scale of strongly agree to strongly disagree then the latent variable will be on a scale of 1to 5 if it's a 1 to 10 scale latent variable will be on that same scale . Now this is generally preferred to the first approach of having a fully standardized solution because we can also get standardized solution using the second approach of fixing one loading to the value 1 and we

also get the standardized solution in that approach as well.

So in confirmatory factor analysis we are interested in making a good measures of our key constructs concepts in our theories and we are then in this the next stage usually going to move on and look at the relationships between the measured concepts and so conventional SEM is focused on that the structural model the relationship between concepts. So we are not so interested in the means of the observed or the latent variables and as i said the conventional way of doing SEM that isn't the focus the focus is on covariances and correlations relationships between the variables but there are occasions within SEM context where we would be interested in the means of latent variables. There are two main areas where we would want to estimate latent means. The first is where we want to see whether there are differences between groups on a latent variable and secondly if we're interested in change over time perhaps if we've got a longitudinal dataset we would want to estimate the mean of the latent variable to see whether that is changing over time.

So when we introduce means in to to our CFA then we do this by adding a constant to the model actually when you fit models in modern SEM software this isn't a choice that the analyst has to make it is if you like done underneath the hood but this is the process that is is actually implemented is to add a constant which has the value the same value 1 for all cases in the model. Now the regression of a variable on a predictor and a constant will give us the mean of that variable in the unstandardized beta of that regression and the mean of an observed variable is the total effect of a constant on that variable so the total effect as we saw in video 1is the sum of the indirect and the direct effects. So if we now introduce a constant which, in path diagrammatic notation, is represented as a triangle and here we have the number 1 inside the triangle to indicate that the constant is 1, then we in this path diagram have again a Y variable and X variable. We have a direct effect from the constant to Y which has the coefficient a , we have a direct effects from the constant to X which is b and a direct effect from x to y which is c so the indirect effect of the constant on y is the product of b and c. So by adding in this constant we can estimate the mean of x which is simply the co-efficient b and we can estimate the mean of y by taking the sum of a and the products of b and c that's the total effect the sum of the direct and indirect effects. So that's how we introduce means into our model now if we've added a mean structure in then we will require some additional identification restrictions because we are now trying to estimate more unknown parameters that's the latent means. So there is a question then about how we estimate and compare one mean to another and the way we do this is by having multiple groups so where we have more than one group in our sample then we can fix the means of a latent variable in in one of those groups to be 0 and then the means of the remaining groups on that latent variable are estimated at differences from the reference group so with mean models and CFA one of the groups always has to have a restriction to their mean value is 0 then the other groups are interpreted in terms of differences from that reference group. When we've looked at path diagrams and thought about the relationship between concepts and indicators between latent variables and observed variables the arrow will be pointing from the latent variable to the observed indicator. So what this is saying in theoretical terms is that the latent variable causes the indicators that's why the arrow points in that direction so we can think of that as meaning if we're trying to measure let's say someone's social capital and we've asked lots of questions in a questionnaire what's actually causing their answers to those questions in the questionnaire is their underlying level of social capital so the causal arrow points from the latent variable to the observed indicators. Now for many concepts that direction of causality makes sense. In other contexts the idea that the causality flows from the latent variable to the indicator doesn't really make sense. So let's think of an example where

we want to measure socioeconomic status and we're going to use indicators of someone's level of education, what kind of occupation they have , their earnings and so on we want to combine these somehow into a latent variable that measures their socioeconomic status . Now what's problematic about this in the reflective indicators context is that it doesn't really make sense to say that I have some underlying socioeconomic status and that if that were to change then my educational level would change or my earnings will change or my occupation will change because actually causalities flowing in the other direction if there is any causality going on here at all. So someone's level of education influences their socioeconomic status as to their earnings so now we're in a situation where the causality makes more sense to flow from the indicator to the latent variable. So the key point here is whether manipulating if we could somehow change someone's score on the latent variable it will make sense to change the score on the observed indicator. Now for some concepts that makes sense for others it doesn't and in the case where it doesn't make sense we 'll essentially turn the arrows round and make the arrows point from the indicators to the latent variable. And in this context we've now got what we call for formative indicators rather than reflective indicators. Now that said it's a different sort of latent variable now that we're dealing with it's essentially a weighted index of the observed indicators and it doesn't have a disturbance term there is no error in it so it's not the same kind of a variable as we would have with a reflective indicator. The key thing is that in the path diagram the arrows point from the indicator to the latent variable rather than the other way round there are off course some quite different procedures for estimating this kind of a model but for now the concern is to understand the conceptual difference and the fact that we have the indicators related differently to the latent variables .

Another common procedure in confirmatory factor analysis is when a researcher may have a very large number of indicators for a latent constructs or for a number of latent constructs this is quite often the case in psychology where there are quite complex latent variables and each one may be has 10 12 or more indicators. One of the problems that researchers run into with this kind of data is that the model can become extremely complex very quickly and there is lots of difficulties that people can run in to with estimation and interpretation and so on simply because there are so many relationships in the observed data because there are such a large number of indicators and latent variables. And this is often combined with sometimes quite small sample sizes which can add to this problem. So when in this situation researchers will sometimes use an approach called item parcelling which is a first stage of taking some score adding up the score for those large numbers of items or for subsets of those subgroups of those items and then those subgroups of parcelled items of some scales then act as the observed indicators for the latent variables so this is a sort of a parsimonious way of treating rather complex data. It does rely on some assumptions about the unit dimensionality all the items in that parcel but it is an approach that researchers who are in that context of having lots of indicators for their latent variables and large numbers of latent variables can pursue.

Last I'm going to talk about a kind of confirmatory factor model where the latent variables are not measured by observed indicators but are themselves measured by latent variables. So we have a sort of hierarchical structure where a first set of latent variables are measured using observed indicators we have to have observed indicators at some point in the mode. But once that first set of latent variables measured then a higher-order factor can be added which is a function of the first stage latent variables. Now this is an approach which is often useful when our theories are not so much about the relationship between variables but are in the dimensional structure of the data for example in psychology there are debates about the number of personality dimensions and often belief systems and so on it's important to understand how many different dimensions there are in addition to how those dimensions

might be related to other variables. So intelligence personality and so on higher-order factor models can be useful. They can also be applied in a longitudinal context. So here's what a path diagram for a confirmatory factor model with a higher order structure would look like. We have the bottom of the diagram now the observed variables in rectangles, there are nine of those and each set of 3 measuring a latent variable. And then the highest level variable e1 is then measured as a function of those three latent variables.

So in this third video I've looked at some of the important issues in confirmatory factor analysis started off by looking at the general idea of using latent variables to measure concepts in our theories. I've contrasted the historical approach the conventional approach of exploratory factor analysis, the unrestricted factor model to more modern confirmatory factor model the unrestricted factor model.

We've looked at how we can give a metric a scale to latent variables by fixing one of the indicators to take the value 1 and therefore take the scale from that reference item. We've thought about how we can analyze means within a confirmatory factor model usually we're mainly focused on associations correlations but we can also estimate means. We've looked at some special cases where we have formative indicators rather than reflexive indicators where we have a first stage of item parceling when there are many indicators and a large number of latent variables and we finished by the special case of a higher-order factor where a latent variable is measured not by observed items but by lower-level latent variables.