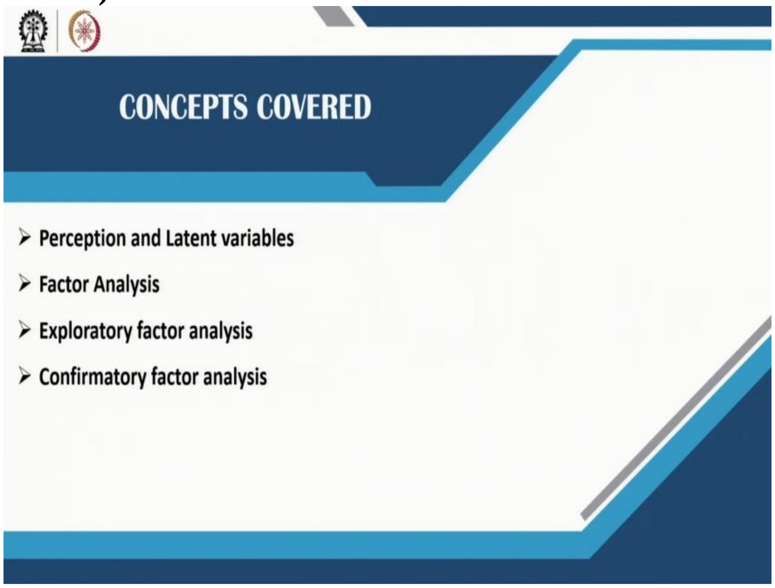
Urban Landuse and Transportation Planning
Prof. Debapratim Pandit
Department of Architecture and Regional Planning
Indian Institute of Technology, Kharagpur

Lecture - 38
Hybrid mode choice model 1(Factor Analysis)

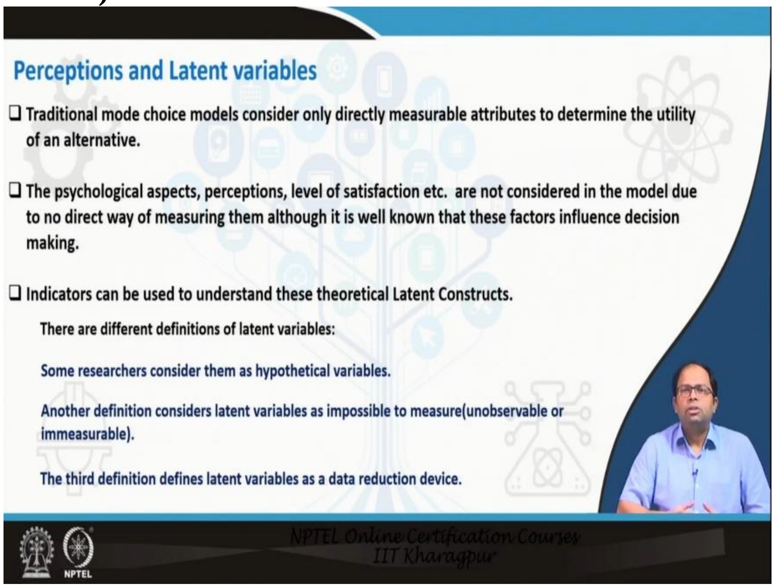
In lecture 38, the focus has particularly been given to factor analysis.

(Refer Slide Time: 00:34)



The different concepts have been covered are; perception and latent variables, factor analysis, exploratory factor analysis, and confirmatory factor analysis.

(Refer Slide Time: 00:48)



Introduction to Hybrid choice models:

Hybrid choice model bears its name by virtue of the incorporation of variables other than the types that are usually included in a choice model like time, cost, income, etc. There are many variables that are based on human perceptions, attitudes, or psychological aspects. These variables have been long ignored in the domain of transportation mode choice modelling, but now experts believe that these play a significant role when a person chooses a particular mode. Including these variables have been found to increase the prediction power of mode choice models. However, they can't simply be added to a model like other usual parameters. There are specialized techniques to handle such variables in order to use these in mode choice models. Factor analysis is the broader technique that serves as a tool to translate these idiosyncratic variables to entities that can be used directly in mode choice models.

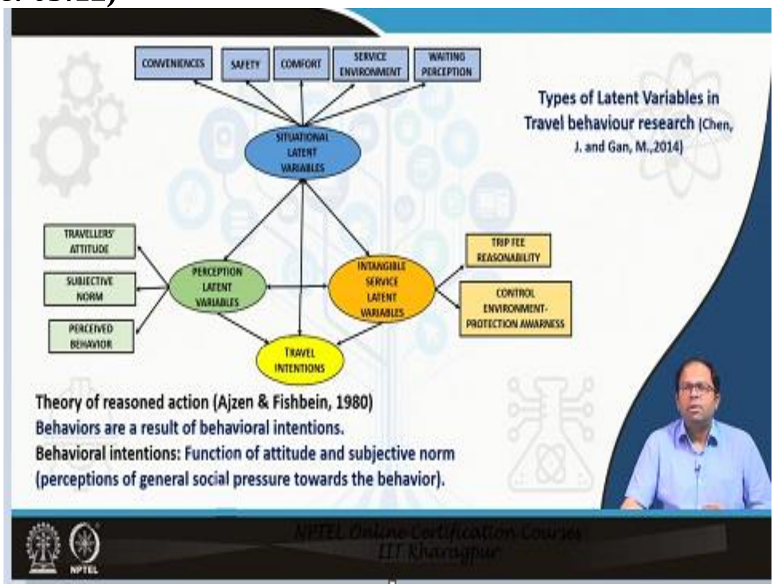
Factor analysis is of two types; exploratory factor analysis (EFA), and confirmatory factor analysis (CFA). The details on these methods have been covered briefly in this lecture. There are many standard statistical analysis texts available that cover this topic in depth and are highly recommended for further detailed investigations.

Perception and Latent variables:

Traditional mode choice models consider only directly measurable attributes to determine the utility of an alternative. Although factors related to psychological aspects, attitudes, etc. cannot be directly measured, as discussed above, experts believe; they also play a major role in mode choice. Perception is the process of interpretation of stimuli by an individual, given the intrinsic characteristics like psychological constructs, morals, values, etc. the person possesses. Although these cannot be measured like time and money, indicators can be used to infer about these indirectly. In the context of transportation mode choice, these idiosyncratic elements are broadly called latent variables. In literature there are many definition of latent variables like; hypothetical variables, unobservable variables, immeasurable variables, data reduction device, etc.

Although latent variables like psychological aspects, perceptions, level of satisfactions, are very important to understand the feeling of people about a particular mode, usually they are not included in a model in large numbers. In factor analysis, the underlying factors or theoretical latent constructs are identified through the indicators of these latent variables. These factors are then used to represent the latent variables in a mode choice model. For example, the satisfaction with bus service can be related to the satisfaction from the safety that the service offers from various elements; the quality of service can be dependent upon the quality of cleanliness, the on-time performance of the service, the real-time information disbursement, the grievance redressals, etc.

(Refer Slide Time: 05:12)



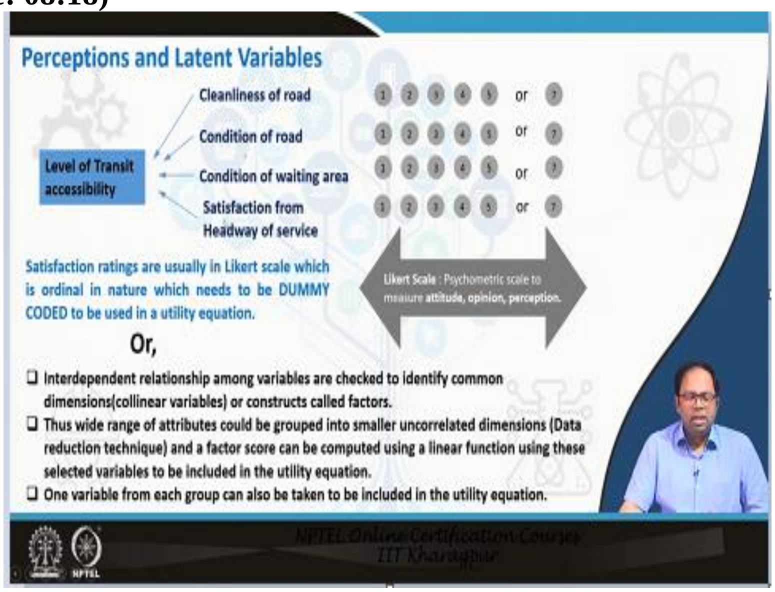
In transportation studies different forms latent variables are used like: **situational latent variables** which deals with the feelings of people at a certain point of time regarding the service, the alternatives, etc. These include variables like perceptions about safety, comfort, service environment, waiting time, how comfortable the journey was and conveniences about that particular mode.

There may be other kinds of latent variables that depend purely on **perceptions**. For example, travelers' attitude, the different subjective norms that he is subjected to and the perceived behavior. This follows the **theory of reasoned action**, which is a very popular theory in social science as well as in psychology in which Ajzen and Fishbein proposed that behaviour is a result of behavioral intentions. Behavioral intentions are a function of attitude and subjective norms. In the context of transportation mode choice, attitude refers to the attitude of a person towards a particular alternative. Subjective norm is perception of general social pressure towards that individual's behavior or in other words, what the society makes a person to think about that particular subject.

That means, a person’s behaviour towards an alternative is dependent on the attitude of the person, and the popular belief of the society regarding that particular alternative. So, perception related variables need to be included in a mode choice model. For example, determining if a person is a pro-environment individual is very difficult to measure, but it could be measured using several parameters or indicators like; usage of bicycle or transit over auto-rickshaw or personal car; buying costlier green products over the cheap polluting ones; etc.

There are also intangible **service latent variables** like, the reasonability of trip fee, which is very difficult to measure. Satisfaction ratings can be taken to record the response to such variables. All the different types of latent variables are together called **travel intentions**. So these are the things that influences mode choice in addition to the easily measurable variable such as travel cost, travel time, waiting time, etc.

(Refer Slide Time: 08:18)

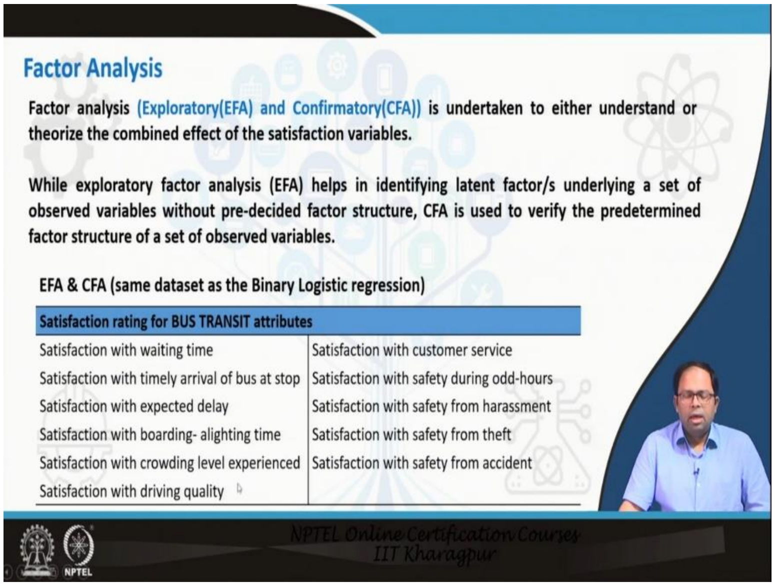


For a variable like **level of transit accessibility**, the perception or attitude of people towards it can be measured through many indicators. For example, cleanliness of the road leading to a particular stop; condition of the road; condition of waiting area in the bus stop; satisfaction from headway of service; satisfaction from the waiting time a person is subjected to in a particular bus stop. Indicators like this could be measured using a **Likert scale** which can be a rating scale of ranging from 1 to 5 or from 1 to 7 or even more. Likert scale is a psychometric scale to measure attitude, opinion, perceptions.

Satisfaction ratings which are ordinal in nature and recorded in Likert scale, can be used directly in the utility equation by dummy coding them as discussed for nominal variables like vehicle ownership, possession of driving license, etc. So, if an indicator has 5 levels of satisfaction, 4 different dummy variables are required to be introduced in the model. Thus 4 indicators/attributes will require 16 new dummy variables to be introduced in the model, which is a huge number. Estimation of such models often does not give good results. Another approach for introducing these indicators in the model is by identifying the latent constructs and loading multiple indicators into fewer entities. These common dimensions or collinear variables or construct are identified through factor analysis and are grouped together in a linear combination, to be represented by a fewer number of uncorrelated variables known as factors. For a given observation, factor score can be computed using the linear combination of the indicators that are loaded together. The linear function is also estimated in factor analysis. For example, if there are 4 indicators which are correlated and they contribute to make up 1 factor, they can be combined using a linear equation. Certain weights can be assigned to each of these indicators and then based on the values of the indicators for a particular observation, the score of the factor for that particular observation can be obtained. This factor score is then eligible to be used in a utility equation directly and hence represents all the indicators it is constructed with.

While doing the analysis, it might be observed that many indicators or variables, although correlated, have very less contribution to the combined factor. In such cases, these variables or indicators can be discarded. In many cases, only a single indicator or variable out of the many correlated variables can be selected to represent the underlying construct in the utility equation i.e. the indicator or variable in itself represents the all the indicators/ variables loaded in the factor, in an efficient way.

(Refer Slide Time: 12:32)



Factor Analysis:

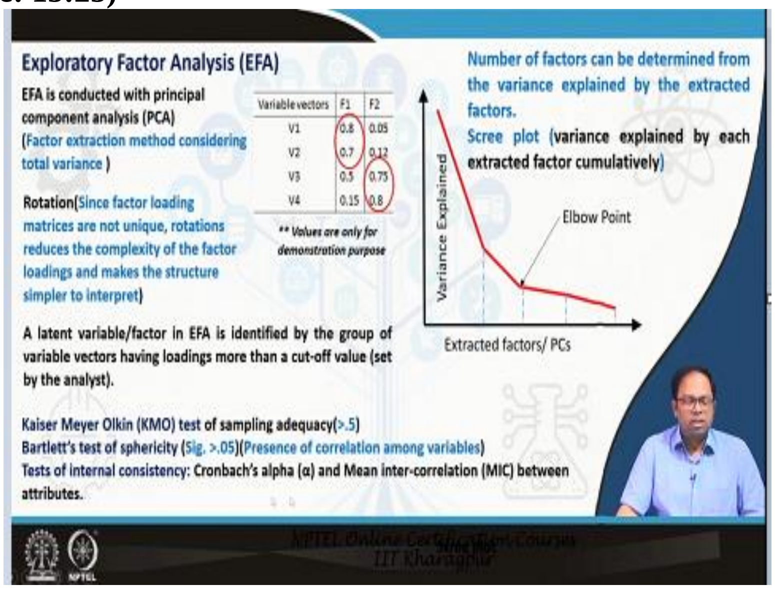
As already discussed briefly, factor analysis is a data reduction technique which is undertaken to understand the latent construct or the hypothetical constructs. The two different approaches to achieve the same are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). **Exploratory factor analysis** helps us to identify the latent factors, the factor loadings or the variables which are loaded to those factors. Statistical analysis is employed to find the variables that are likely to form a factor. The factors hence formed might not match with the apriori knowledge. For example, there are 4 indicators which could lead to a particular latent construct based on the EFA. It may be found that these four indicators are not cohesive theoretically and one of the indicators is related to some other factor, or multiple indicators are getting cross-loaded to some other factors. So, exploratory factor analysis is where the relationship between the supposedly collinear indicators are explored.

Confirmatory factor analysis is a tool where the hypothesized relationships are theorized or tested for correctness based on theory. In simpler words, EFA helps in identifying latent factors underlying a set of observable variables, without any pre-decided factor structure. Whereas CFA is used to verify a predetermined factor structure of a set of observed variables based on theoretical knowledge, or apriori knowledge, or previous research.

In the study discussed in binary logistic mode choice model, some indicators (satisfaction) were used to determine some of the underlying factors behind selection of bus over auto-rickshaw.

Indicators used were: satisfaction with waiting time; satisfaction with timely arrival of bus at stop; satisfaction with expected delay; satisfaction with boarding-alighting time; satisfaction with crowding level experienced; satisfaction with driving quality; satisfaction with customer service; satisfaction with safety from during odd-hours; satisfaction with safety from harassment; satisfaction with safety from theft; satisfaction with safety from accident. While, many variables can be included in a survey, factor analysis can be used to determine the factors.

(Refer Slide Time: 15:25)



Exploratory Factor Analysis:

Consider a list of four variables V1, V2, V3, V4. Let us assume 2 factors F1 and F2 are identified to explain the four variables. Each of the variables has some impact on both the factors, as shown by the factor loadings. Factor loading of V1 on F1 is 0.8, and that on F2 is 0.05; Factor loading of V2 on F1 is 0.7, and that on F2 is 0.12; Factor loading of V3 on F1 is 0.5, and that on F2 is 0.75; Factor loading of V4 on F1 is 0.15, and that on F2 is 0.8. So, it can be seen that V1 and V2 is primarily loaded on F1, and V3 and V4 are primarily loaded on F2. Apart from the primary loadings, the variables are also cross-loaded on the other factor.

To understand the source of these loadings, the theory behind EFA must be understood. EFA can be done in many ways, principal component analysis (PCA) being the most common one. In the study mentioned in the previous section, PCA was employed to extract factors, in which factors are extracted based on the total variance explained by indicators for each of the factors. PCA dictates that, much of the variance-covariance structure of a given dataset with ***p*** variables can be accounted by a smaller number of ***k*** variables. In this case, ***p*** represents the indicators or variables, and ***k*** represents the factors.

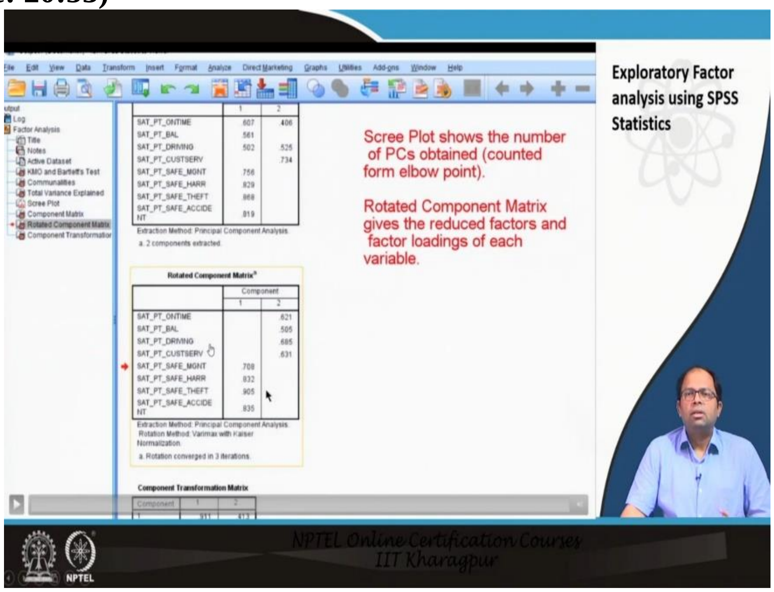
If the existing variables and their observations are assumed to be in a space, each point can be resolved or projected along as many axes or factors as the total number of variables, but this does not solve the purpose of data reduction. The axes or factors are rotated in such a way that, for each variable, the projections or loadings are most profound on any one of the many axes used for the resolution. Hence it can be observed that most of the variability in the data is being explained by a fewer number of factors only with all the original variables loading their respective impacts on them. This phenomenon of rotating the axes to capture most of the variability is called factor rotation. There are many types of factor rotation (orthogonal: varimax, quartimax, equamax; Oblique: oblimin, promax); **varimax rotation** being the one most commonly used in which the angle between each of the factors in 90^{o} (orthogonal). In other words, since factor loading matrices are not unique (considerable amount of cross-loadings exist), rotation reduces the complexity of the factor loadings and makes the structure simpler to interpret.

A latent variable or a factor in EFA is identified by the group of variable vectors or indicators having loadings more than a cut-off value set by the analyst. In the mentioned study, this cut-off value was taken as 0.5 which means, any factor loading below 0.5 was not considered. In cases where the variables were found to be loaded on multiple factors, the one on which it had higher loading, was considered to be related to that.

After the identification of the factors, the loadings by various variables, the number of factors that is to be considered adequate to describe the variability or variance-covariance structure of the whole dataset, needs to be determined. For this, something called a **scree plot** is used where variance explained by each extracted factor is arranged in a descending order and is cumulatively added and plotted, as shown in the figure. The variability decreases as the number of factors increase. After a point, there is not much decrease in the variability anymore. This point is called '**elbow point**'. The factors that are towards left of the elbow point (excluding the point itself), are considered to be the appropriate factors that can be used to represent the indicators. For example, in the figure shown, only 2 factors are more than enough to explain most of the variance.

Keiser Meyer Olkin (KMO) test of sampling adequacy is a test to check if the sample is adequate for performing a factor analysis. A value of KMO > 0.5 is acceptable for a given dataset. The, there is **Bartlett's test of sphericity** which checks for correlation between variables. If the significance of this test is > 0.05, then the dataset is said to have correlated variables and hence factor analysis can be performed. Test of internal consistency of the factors are done using **Cronbach's alpha** and **mean inter-correlation** (MIC) values.

(Refer Slide Time: 20:53)



EFA using SPSS:

A demonstration of the process to carryout EFA in SPSS has been shown in this section. The steps to be followed are given in the below. From the **analyze** option, select **dimension reduction**, and then '**factor**'. In the factor analysis tab, from the list of variables, select the variables which are to be analyzed for communalities and add them to the variables section. From **descriptives** tab, select the KMO-Bartlett's test of sphericity. From the **extraction** option, select the method of extraction. In this case 'principal components' is selected. Select **scree plot** to enable the plot in output. In the extract section, select the method to determine the number of factors. In this case, it is based on **eigen values**. From the **rotation** option, select the rotation method or the factors. Form **display** section, enable display of rotated solutions. From the **options** tab, select the method to handle missing values, and enable suppression of small coefficients. Click **OK** to run the estimation.

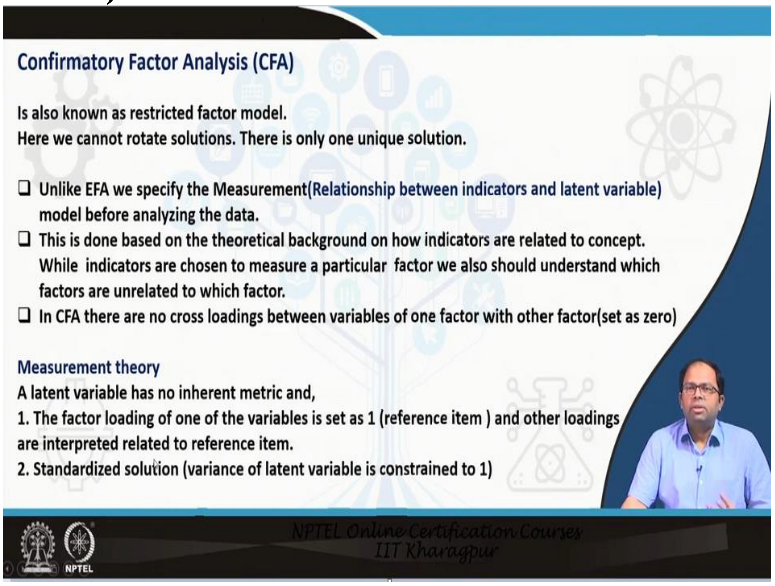
Analyze → Dimension reduction → Factor
→ Select the indicators or variables that are to be analysed for communalities
→ From ***descriptives option****, select KMO − Bartlett's test*
→ From ***extraction option****, select extraction method (principal component, maximum likelihood, etc.)*
→ Check ***scree plot*** *(to see the plot)*
→ In ***extract****, select the method to determine number of factors (based on eigen values,*
a fixed number of factors)
→ From ***rotation*** *option, select the factor rotation method (varimax, quartimax, promax, equamax, etc)*
→ In ***display*** *section,*
check rotated solutions → From ***save*** *option, select the factor scores to be saved (if required)*
→ From ***options****, select the method to handle missing values*
→ Check ***supresssmall coefficients*** *and set the cut − off value → Click* ***OK***

In the results, the scree plot shows that two factors are adequate to represent the variables. In the rotated component matrix, as shown in the following table, the factors and the factor loadings are shown.

Rotated Component Matrix			
Variables	**Variable names in database**	**Component**	
		1	**2**
Satisfaction from on-time performance.	SAT_PT_ONTIME		0.621
Satisfaction from boarding-alighting time	SAT_PT_BAL		0.505
Satisfaction from driving quality	SAT_PT_DRIVING		0.685
Satisfaction from customer service	SAT_PT_CUSTSERV		0.631
Satisfaction from safety during odd hours	SAT_PT_SAFE_MGNT	0.708	
Satisfaction from safety from harassment	SAT_PT_SAFE_HARR	0.832	
Satisfaction from safety from theft	SAT_PT_SAFE_THEFT	0.905	
Satisfaction from safety from accident	SAT_PT_SAFE_ACCIDENT	0.835	

The factor loadings are pretty unique, and if at all any cross loadings exist, it is less than the cut-off value of (0.5). Looking at the variables and the factors, it is pretty evident that one of the factor is related to the bus operation and service characteristics, and the other is related to the overall safety in bus. Hence the eight variables are reduced to two factors or latent variables.

(Refer Slide Time: 22:59)



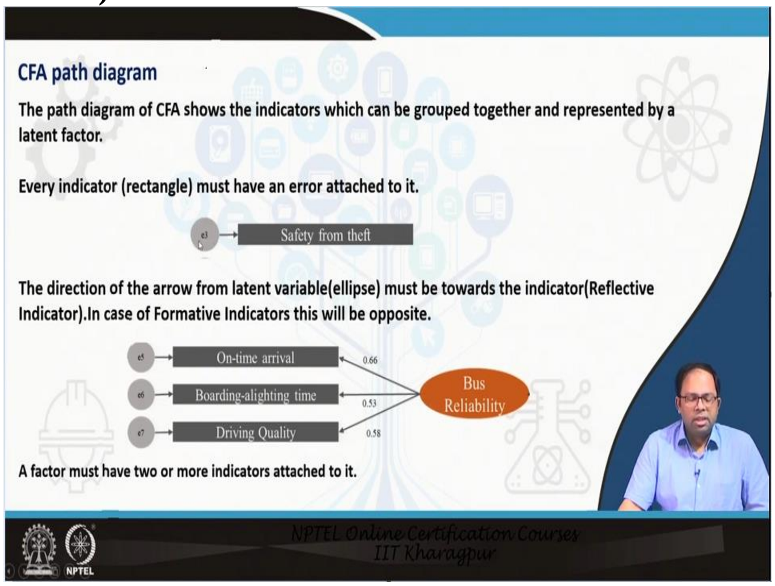
Confirmatory Factor Analysis (CFA):

EFA gives an idea about how the variables could be formed into factors, but the results may not be matching the apriori knowledge. So, in confirmatory factor analysis (CFA), using the results of EFA, a model is proposed by specifying the variables and linking them with the factors to represent loading. The model is then tested for theoretical correctness. CFA is also called restricted factor model as the factors cannot be rotated and no scope of one variable loading in multiple factors (cross-loadings) is allowed. As cross loading are assumed to be zero, there would be some covariance between the factors, which is natural.

Unlike EFA, the measurement model is proposed before analyzing the data with all the relationship specifications between indicators and latent variables, based on the theoretical knowledge. That means, fair idea about which variables or indicators impact which factors, and which variables don't impact a certain factor, needs to be there before specifying the measurement model.

In order to measure the loadings in CFA, the measurement theory specifies that a latent variable has no inherent metric. For each factor, the factor loading of one of the loaded variable is set as '1', and loading of other variables are estimated with reference to that. For a standardized solution, the variance of the latent variable is constrained to 1, and the loading are re-estimated accordingly.

(Refer Slide Time: 25:20)

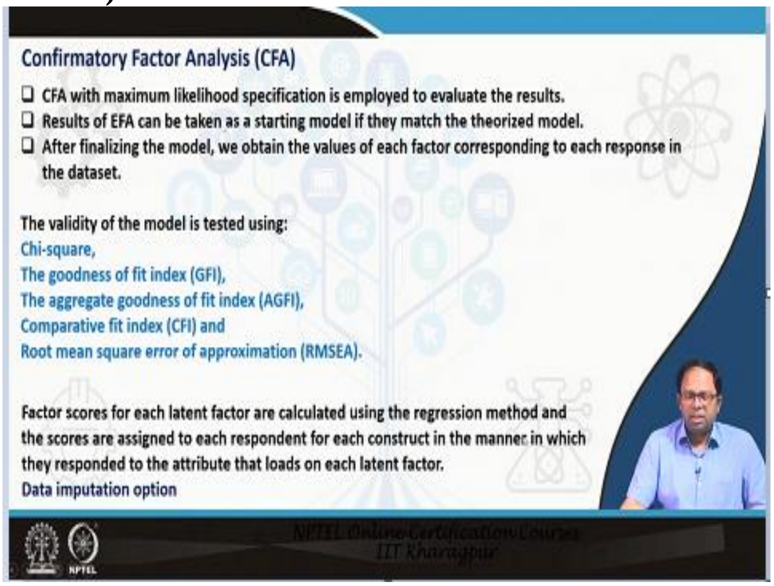


CFA path diagram:

The measurement model in CFA is constructed manually, and is called path diagram. The diagram represents the indicators that can be grouped together as they get linked to a common factor. So, in CFA we construct something called the path diagram. As shown in the figure, the factor 'bus reliability' is linked to the indicators; on-time arrival, boarding-alighting time, and driving quality. These kind of relationships are hypothesized in CFA.

There are certain norms that needs to be followed while constructing these relationships. A factor must have two or more indicators attached to it. If not, there is no point in constituting a factor as the variable in itself is a singular entity. The direction of arrows need to be paid attention to. In case of reflective indicators, the direction of arrows must be from the factor to the indicators. In case of formative indicators, the direction of arrow must be from the indicators to the factor. Each variable has a measured part, and an error part. Both of them are represented in the measurement model by the elliptical shaped error terms attached to the rectangular shaped indicators or variables in the diagram.

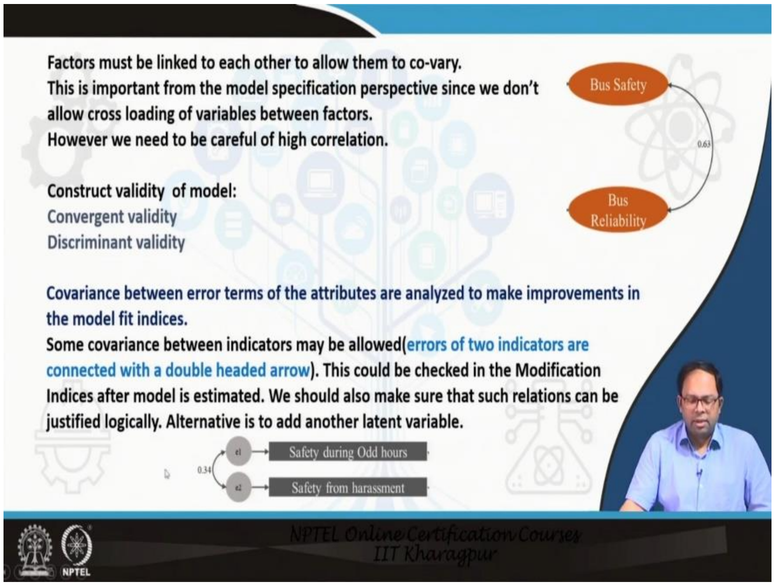
(Refer Slide Time: 26:54)



Confirmatory factor analysis is solved using maximum likelihood specification. The results of EFA can be taken as the starting model in CFA, and then gradually it is refined by removing or adding some variables. Alternatively, based on theory, a model can be developed afresh. After finalizing the model, the factor loadings for each of the indicators in the model are determined.. Based on the factor loadings assigned to each indicator or variable, factor scores for each latent variables are calculated using a linear regression. For example, If an individual has responded to four satisfaction ratings, and these four indicators have been found to be reduced to a single factor, the individual will be assigned a score based on the factor loadings of the four satisfaction ratings (indicators). In order to do that, the values reported for each of the indicators are multiplied by the respective factor loadings and they are added up to obtain a number, which is the factor score for that particular individual. Usually the software has an option to '**impute**' factor scores back to the database for each observation. After imputation, the factor scores can be directly used in mode choice model.

The validity of a model is tested using several indices like; chi-square, the goodness of fit index (GFI), the aggregate goodness of fit index (AGFI), comparative fit index (CFI), root mean square error of approximation (RMSEA), etc. These are like the R-square value in regression which are used to determine how good is the model.

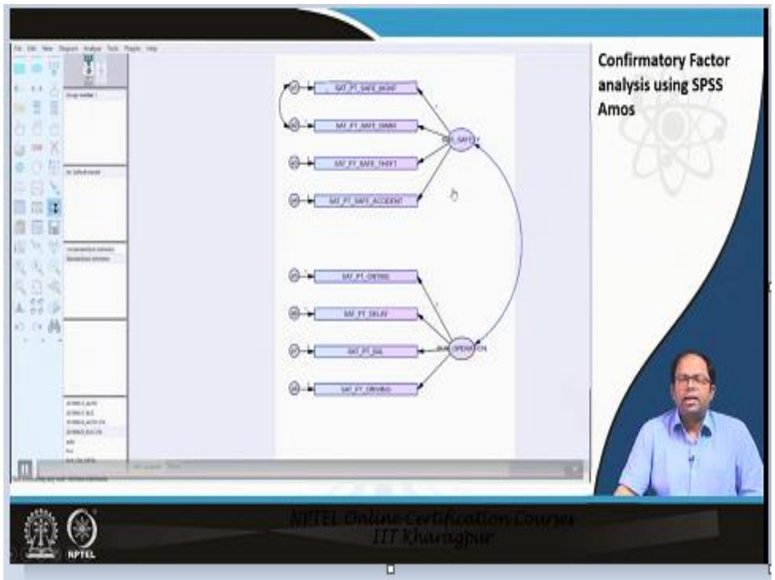
(Refer Slide Time: 29:20)



In a CFA model, factors need to linked to each other to allow them to co-vary as shown in the diagram, where bus safety, and bus reliability is are connected with each other. Although there are no cross-loadings, the scope for them to co-vary needs to be there, as it is important from the model specification perspective. But in case of the correlation being too high, the model might be required to be rejected. This can be checked using **construct validity**, which has two methods; convergent validity and discriminant validity.

Once the estimation of a model is complete, various statistics and scores of various indices are shown in the results. Often a model has scope to be improved by allowing covariance between error terms. These are indicated by the software as **modification indices**. Modification indices list out the various possible error covariance that can be introduced, with the corresponding improvement of the model due to the modification. It is up to the judgement of the analyst to allow all, some, or none of the error covariance presented by modification indices, as there needs to be proper justification for doing that. Allowing the covariance also improves the RMSEA value or other model fit statistics. However, if covariance is too much, then a new factor needs to added to the CFA model.

(Refer Slide Time: 31:43)



CFA using SPSS AMOS:

A demonstration of CFA has been shown in this section which has been done in SPSS AMOS. **Analysis of a Moment Structure** (AMOS) is an added SPSS module used for sequential equation modelling, path modelling, and confirmatory factor analysis. After doing the EFA in SPSS, the rotated component matrix is used as a guide for the first model for CFA. In AMOS, the same database as EFA is loaded (in ***.sav*** format) to perform CFA. The steps to be followed are given as follows:

→ *From the rightmost panel,* ***Select data file(s)*** *menu, load the data file (in sav format)*

→ *Copy and Paste the ROTATED COMPONENT MATRIX from EFA, in the* ***pattern matrix builder*** *plugin.*

→ *Ckick* ***OK*** *(Alternatively, the path model can be drawn from scratch using the various drawing elements in AMOS)*

→ *Make sure all the elements are connected as they are supposed to be*

→ *Name the factors by double − clicking on the Factor ellipse*

→ ***SAVE*** *the file (not doing so will generate error while estimation)*

→ *From* ***Analysis Properties****, set the preferences based on the appropriateness to the study*

→ *Click* ***Calculate estimates*** *to run the estimation*

→ *From the* ***View text*** *option, the output window can be accessed, and inferenes can be made*

→ *If required, make changes to the model, and keep estimating till adequate fitness is achieved*

→ *From the top part of the panel adjacent to model space the option to* ***View the output path diagram***

→ *From the bottom part of the same panel,* ***standardized estimates*** *needs to be selected.*

→ *From Analyze, select* ***Data imputation***

→ *Select the imputation method (Regression is default) and make sure* ***destination file path*** *is correct*

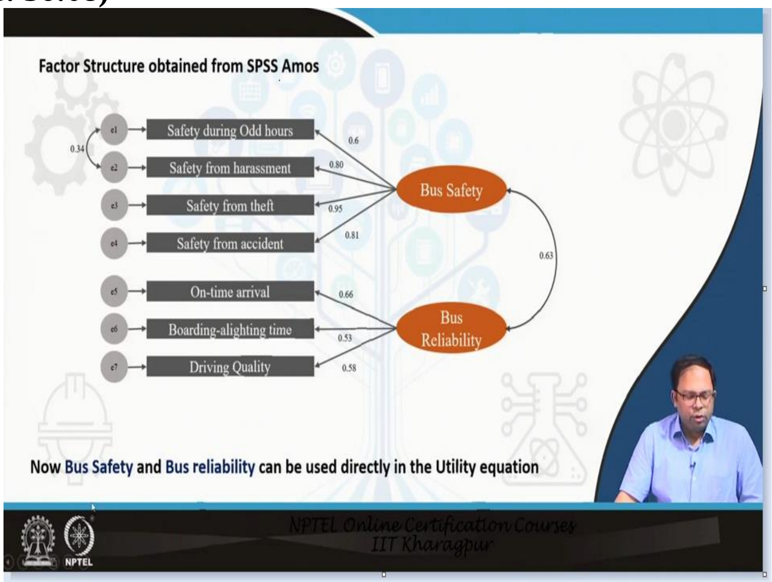
→ *Click* ***Impute*** *to automatically get the factor scores for each observation in the specified file.*

After the file is loaded, the rotated component matric from the EFA output can be directly copied and pasted in the plugin named '**pattern matrix builder**'. If the plugin is not available, the path model can be drawn from scratch using the various geometrical elements given for observed variables, unobserved variables, factors, and arrows. The tools provided in AMOS can be further exploited to enhance the readability of the model. The factors need to be named appropriately before starting the estimation. For example, in the demonstration file, the factors have been named '**Bus_safety**' and '**Bus_reliability**'. The standard nomenclature rules of SPSS as discussed previously, applies here also. This is done simply by double clicking the factor element in the model. It also needs to be made sure that, all the elements are connected properly. Before calculating the estimates, the file needs to be saved, otherwise error is generated while estimation. In order to estimate the results, '**Calculate estimates**' is clicked. The model estimates and the goodness of fit statistics can be seen from the '**View text**' option. Apart from the standard outputs, an analyst may choose other statistics and preferences from '**Analysis properties**' option, before estimating the model. Modification indices, standardized estimates, factor scores, correlation of estimates, etc. are usually options in the analysis properties. Modification indices lists the new correlations that can be introduced to enhance the model fit (if required). The standardized estimates can be accessed to view the final result. This can be done by clicking '**View the output path diagram**', from the top-most part of the panel adjacent to model space.

In the standardized model, a few variables can be found to have very less factor weights. This implies that those variables do not contribute to the factor significantly. Such factors are removed from the model, and the model is re-estimated. There are various fit indices like GFI, RMR, RMSEA, CFI, etc. In order to obtain a fit model, these goodness of fit indices needs to be satisfied. The cut-off values for each of them depends on widely accepted values from literature, and the analyst's judgement. For example, RMSEA and RMR should be $\leq$0.08; the value of GFI as nearer as possible to 1.0 is considered as a better fit. If the analyst feels that the estimated model, although showing good fit, is not theoretically appropriate, other factors can also be added. Based on the various fit indices, and the judgement of the analyst, a model can be considered to be the final path model for CFA.

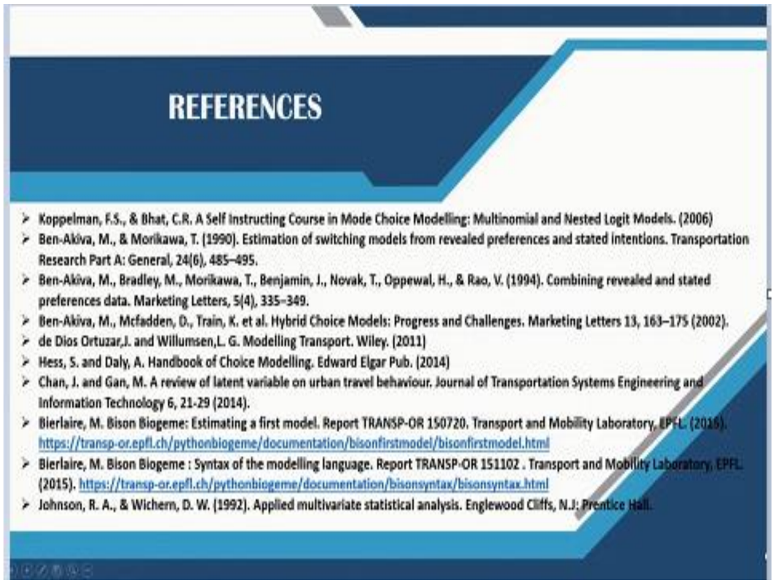
The factor scores can be directly obtained for each of the observations in the dataset by opting for data imputation from '**Analyze**' menu. The desired file path and the appropriate imputation method also needs to be specified. By default, '**regression**' is opted from the list of imputation methods. As '**Impute**' is clicked, a separate data file (in ***.sav** format) is created in the specified file location with all the old data in the database and newly added columns for factors, with the factor scores in them for each observation. So, this is how path diagram is built for CFA.

(Refer Slide Time: 36:08)



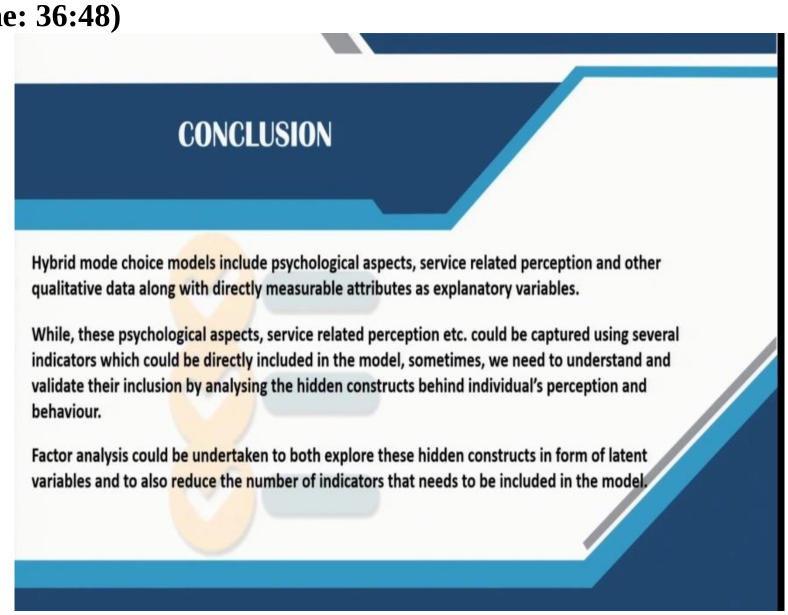
The final output path model or the factor structure, from the demonstration is shown in the figure. The different factor loadings for '**bus_safety**' and '**bus_reliability**' are shown for each of the indicators. These loadings, or regression weights are used to compute the factor scores using the inbuilt data imputation method in AMOS. So, these two factors can now be directly used in the utility equation.

(Refer Slide Time: 36:41)



These are some references for further reading.

(Refer Slide Time: 36:48)



In the conclusion, it can be said that, hybrid mode choice models include psychological aspects, service related perception and other qualitative data along with directly measurable attributes as explanatory variables. While these psychological aspects, service related perception etc. could be captured using several indicators, which could be directly included in the model, sometimes the hidden constructs behind individual perception and behaviour needs to be understood and validated before including them in the model. Factor analysis can be undertaken to both explore these hidden constructs in form of latent variables and to also reduce the number of indicators that needs to be included in the model.