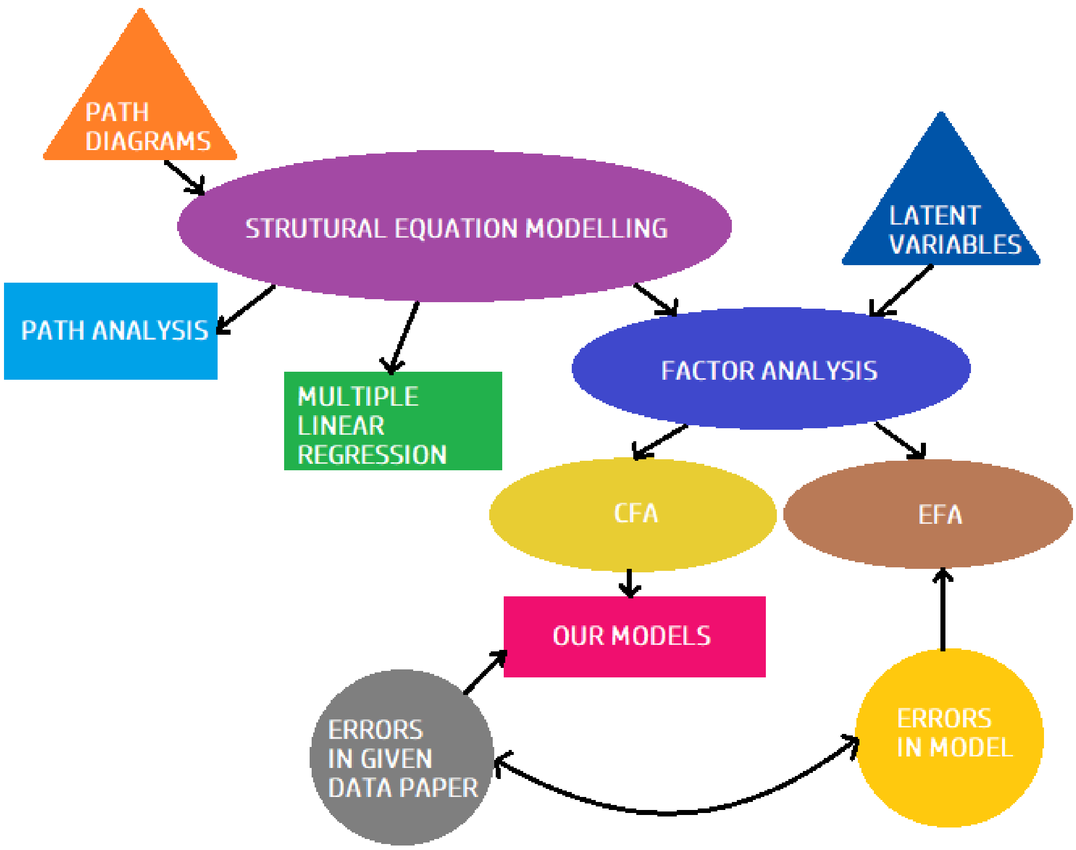
Indian Statistical Institute Bangalore

Bachelor of Mathematics (Hons.) First Year

Introduction to Statistics and Data Computation
Group Project

Survey data on the impact of COVID-19 on parental engagement

Malav Ruchira Disha Ashwinie



21st April, 2023

ABSTRACT

In this project we attempt to present and analyse the data provided in the data paper titled **'Survey Data on the Impact of Covid-19 on the parental engagement in 23 countries'** by

- Eliana Maria Osorio-Saez, Nurullah Eryilmaz, Kalyan Kumar Kameshwara, Dan Zhao and Andres Sandoval-Hernandez from University of Bath, United Kingdom
- Yui-yip Lau and YM Tang from The Hong Kong Polytchnic University, Hong Kong
- Elma Barahona, Leví Astul Castro Ordóñez, Esther Fonseca Aguilar, Ricardo Morales Ulloa and Carla Leticia Paz from Universidad Pedagógica Nacional Francisco Morazán, Honduras
- Adil Anwar Bhatti from University of Karachi, Pakistan
- Godfried Caesar Ofoe from Ghana Education Service, Ghana
- Artemio Arturo Cortez Ochoa and Carolina Valladares Celis from University of Bristol, United Kingdom
- Rafael Ángel Espinoza Pizarro from Universidad Nacional de Costa Rica, Costa Rica
- Maria Magdalena Isac from KV Leuven, Belgium
- K.V. Dhanapala from University of Colombo, Sri Lanka
- Ysrael Alberto Martínez Contreras from Pontificia Universidad Católica del Perú, Peru
- Geberew Tulu Mekonnen from University of Tasmania, Australia
- José Fernando Mejía and Lina Maria Saldarriaga from Programa Aulas en Paz - Universidad de los Andes, Colombia
- Catalina Miranda, Ernesto Treviño and Cristóbal Villalobos from Pontificia Universidad Católica de Chile, Chile
- Shehe Abdalla Moh'd from State University of Zanzibar, Tanzania
- K Kayon Morgan from University of Hartford, The United States
- Thomas Lee Morgan from Sacred Heart University, The United States

- Sara Mori, Silvia Panzavolta and Alessia Rosa, from Università Telematica degli Studi (IUL), Italy
- Forti Ebenezah Nde from University of Yaounde, Cameroon
- Lluís Parcerisa from Universitat Autònoma de Barcelona, Spain
- Oscar Picardo from Arizona State University, The United States
- Carolina Piñeros from Corporación Colombiana de Padres y Madres - Red PaPaz, Colombia
- Pablo Rivera-Vargas from Universidad de Barcelona, Spain
- Adrián Silveira Aberastury from Universidad de la República, Uruguay
- Kyoko Taniguchi from Hiroshima University, Japan
- Allison Zionts from Goldsmiths, University of London, United Kingdom

We also develop a firm understanding of the various methods of statistical inference like Confirmatory Factor Analysis and Structural Equation Modelling utilised to obtain meaningful and reliable deductions from the provided data. We conclude our project with the presentation of independently collected data and apply the analytical techniques that we have learnt on this data.

Keywords: Parental Engagement, **Confirmatory Factor Analysis (CFA)** , Covid-19, Acceptance, Confidence, Socioeconomic Status, Multi Group- Confirmatory Factor Analysis (MGCFA)

CONTENTS

CHAPTER 1

INTRODUCTION

"Every calamity is a statistical opportunity."

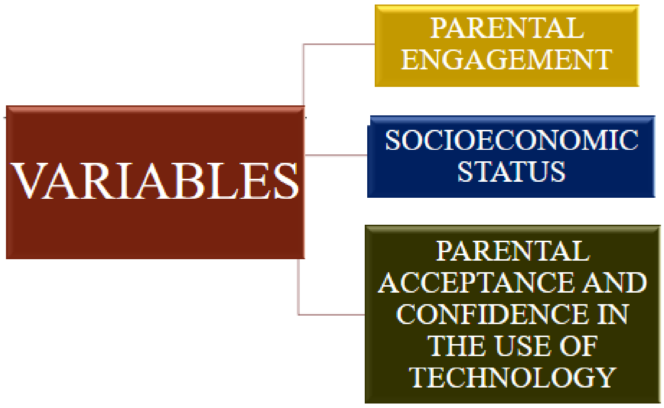
- Anonymous

The ubiquitous Covid-19 pandemic was a singular event in world history that was sure to severely impact every system of importance. Education, being the pillar of our society, did not escape this crisis unscathed. A few weeks saw the world switch from routine methods of instruction and schooling to either online platforms or not even that. The student of today is the citizen of tomorrow and it was indeed highly distressing to see the youth confined to closed doors and receive essential life lessons from a screen.

In such a situation, it is undoubtedly the responsibility of parents/guardians of a child to ensure that the gap in education resulting from this unforeseen event is not too large. Thus, we arrive at the topic of parental engagement and Covid-19's impact on it.

But at the very outset of our analysis, we are faced with an issue that demands our immediate attention. We realize that concepts such as parental engagement and impact of one event on a construct are hazy at best.We do not have the apparatus to evaluate them with the precision and reliability that should precede any inference. Our only hope is to qualitatively measure events directly affected by these constructs and this kind of data is precisely what we handle in this project. The focal point of our observation shall be the three variables which are mentioned in the next paragraph. We measure these variables on scales where the user reflects and evaluates their position on the scale. Then we utilise analytical procedures to get conclusions. We then interpret these conclusions.

To determine what measurable variables parental engagement could affect we must first have a clear idea of what it is. We consider Parental Engagement to be "the proactive engagement of parents in various activities and behaviours that aim to promote learning and development of their child."Several studies in the past have shown that parental engagement is crucial to the holistic development of the child and highly influences the scholastic approach, social behaviour and mental health of the child in question and shapes who he/she will become in the future. The vastness of this concept makes it tedious to comment upon every sphere in which it appears. We view parental engagement from the viewpoint of a global pandemic and the questions asked and the variables observed are framed keeping that in mind. In this project the variables that we examine are:



A brief description of the project is as follows:

The first part of the project is a descriptive analysis of all components of the survey. This includes the identification variables used to code the data, the three qualitative variables and the construction of scales.

Three scales were constructed and included in the data-set depending on the three variables and we describe these in the first section. These scales were created using Confirmatory Factor Analysis and Multi-Group Confirmatory Factor Analysis which are the primary topics that we explore in the second section. Additionally, we replicate the analyses in the data-paper using CFA and MG-CFA.

In the third section we present the data collected independently using the provided questionnaire, analyse this data and see if these inferences match with those provided in the data paper.

CHAPTER 2

DATA DESCRIPTION AND ACRONYMS

The detection of the first case of Sars-Cov-2 in the late of November 2019 in China and later in early March in several other countries had initiated urgent governance steps by the Ministries of Education to carry out various educational and learning activities remotely. Online learning platforms such zoom, google meet, moodle, Wikipedia, youtube played a vital role in effectuating this change.

But obviously this approach had several barriers like availability of electronic devices, lack of internet connectivity, and the umpteen distractions that the internet readily provides. The role of a parent in supporting their child's education which in normal circumstances may extend to emotional support and providing a healthy home environment, escalates, in such a scenario, to what we may call direct and indirect educational practices.

By direct educational practices we mean providing learning experiences to children such as: reading to children, using complex languages (for younger students), confirming that the work allotted by the school is completed in time, ensuring that the child engages in other activities like painting, singing, drawing, etc besides schoolwork. Indirect education practices relate to responsiveness and warmth in interactions and conversations, checking up on the mental health of the child and developing necessary IT skills.

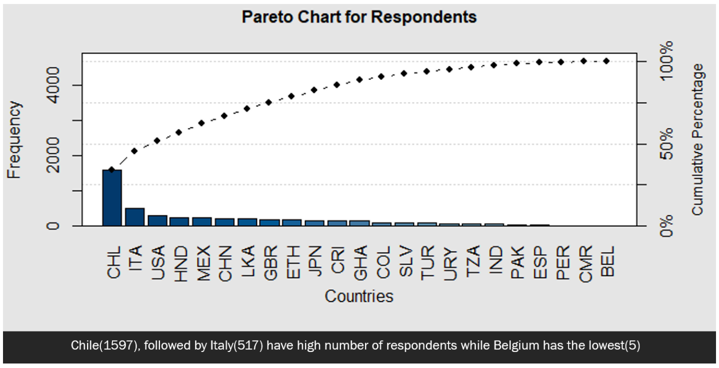
The data provided in this study allows researchers to embark on investigations to the above and other related areas and questions.

The ICIPES 2020 data files contain various identification variables that aid in the identification of the participant's salient characteristics. The codes and descriptions of the variables have been given in the table below:

NAME OF VARIABLE	DESCRIPTION
IDCNTRY	This variable indicates the country or participating education system. the data refers to an up to six-digit numeric code based on the ISO 3166 classification, with adaptations reflecting the participating education systems. This variable should always be used as the first linking variable whenever files are linked within and across countries.
CNT	This variable indicates the participant's three-letter alphanumeric code, based on the ISO 3166-1 coding, with adaptations reflecting the participating country
CNTPARID	This variable indicates the country's three numeric code, based on the ISO 3166–1 coding, plus a unique identifier for each respondent.
REGID	This variable identifies the specific region that each country belongs to. There are five geographical regions: Africa, East Asia, Europe, South Asia and America.
REG	This variable indicates the participant's three-letter alphanumeric code, based on the ISO 3166- 1 coding, with adaptations reflecting the participating geographical regions
URN	This variable identifies the specific questionnaire that was administered to each parent. This number was automatically provided by the Online Surveys tool.

The data provided in the data-paper was collected by means of an online survey with questions within a predetermined thematic framework. It is a collaborative effort of more than 20 institutions to investigate the ways in which parents and caregivers built capacity to engage with children's learning during the period of social distancing arising from the global COVID-19 pandemic. The survey was conducted across 23 countries. The questions were not set in order or in phrasing. The questions are semi-structured and are qualitative in nature. Online survey was a cost-efficient method of data collection. Obviously, we are faced with a selection bias since the respondents of the questionnaire will be individuals who have ready internet access—(electronic devices). Nevertheless, as the data paper claims, this was an effective way to get real data from the online population. A total of 4658 respondents (parents) answered questionnaires from the participating countries: Cameroon, Ethiopia, Ghana, Tanzania, China (i.e., Mainland, Hong Kong, and Macao), Japan, Belgium, Italy, Spain, Turkey, United Kingdom, India, Pakistan, Sri Lanka, Chile, Colombia, Costa Rica, El Salvador, Honduras, Mexico, Peru, Uruguay, the United States. Later, the 23 countries were split into five regions: Africa, East Asia, Europe, South Asia, America.

A pareto chart of the number of respondents by country has been given below.



CHAPTER 3

VARIABLES IN OUR DATA

3.1 Parental Engagement (PE/ENG)

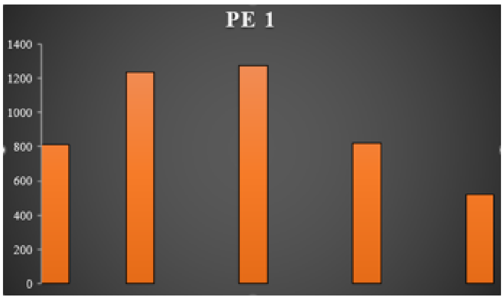
This variable aims to evaluate the direct engagement of the parent with the child's education. It ascertains whether the parent independently forms ideas about what the children need to learn or hold enough confidence in the educational institution that the child is enrolled in to do the same. The pandemic made home-schooling necessary and as we all know it is difficult to have as structured and organized learning environment at home as one might have in school. As a result, it is essential that parents set a fixed timetable to ensure productivity. It is also beneficial to the child if the parent and children actively participate in non-academic activities such as cooking, woodwork, online games, sports, etc. This would make the child view studies with an importance but at the same time realize that there are other things equally important.

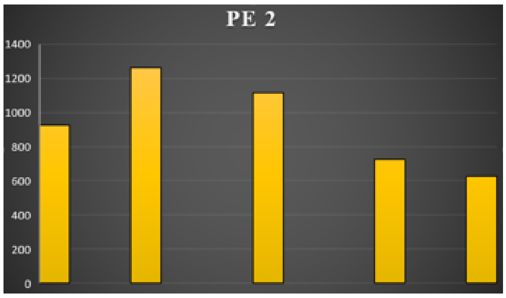
The parental engagement scale was constructed using the following questions: Q21_2, Q21_3, Q22_2, Q22_3, and Q22_6 from the data set. Always, Often, Occasionally, Rarely, Never (from 0 to 4)

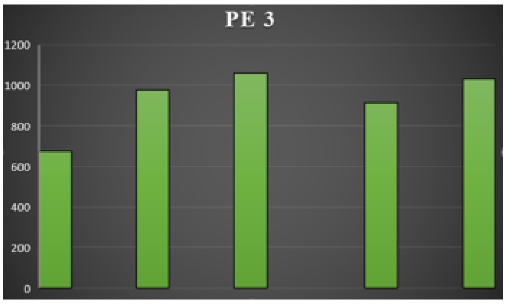
- Q21_2 Follow my ideas about what my children need to learn
- Q21_3 Mix my own ideas with the school's plan on what my children need to learn
- Q22_2 I list and prepare the activities myself before developing them with my child(ren)
- Q22_3 My children and I have a set home-schooling timetable.
- Q22_6 I develop with my children spontaneous learning activities not necessarily schoolrelated such as cooking, woodwork, online games, physical activities, etc.

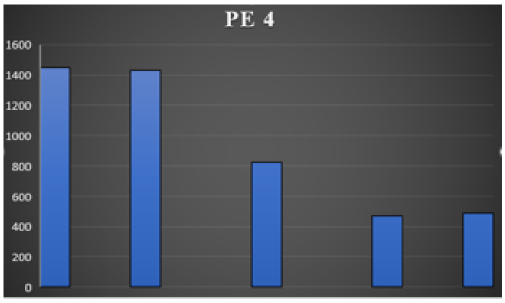
As can be observed these variables cannot be measured directly. A scale was constructed with options as Always (0), Often (1), Occasionally (2), Rarely (3), Never (4).
The values that were used in documentation and computing are given in parentheses.
Below are the histograms of the responses of parents for the five questions from 0 to 4. Here the

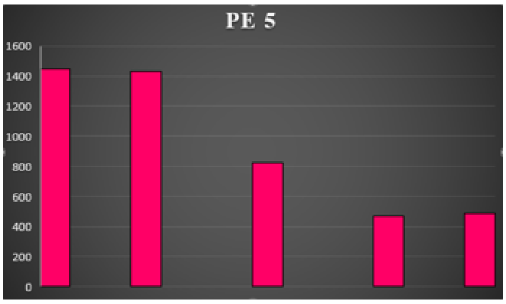
questions have been labelled as Q21_2 as **PE 1**, Q21_3 as **PE 2**, Q22_2 as **PE 3**, Q22_3 as **PE 4** and Q22_6 as **PE 5**.











As we can see from the above histograms on the five questions, most parents responded *often* or *occasionally* for the first three questions and most parents responded *always* or *often* for the last two quesions.

3.2 Socioeconomic status (SES)

It has been found in several studies that SES influences parental involvement in terms of a child's education. In the article Family Socioeconomic Status, Parent Expectation, and a Child's Achievement, the relationship between a family's SES and the educational expectations parents had for their children was researched. This study found that having a higher SES positively impacted parental involvement and therefore increased the expectations set by parents. The higher the parental income, the greater the expectancy that their child would attend and finish college (Stull, 2013). Since the expectancy is higher, parents from high-SES families also tend to invest a greater proportion of their time in the child's education to ensure the said expectation. On the other hand, studies have found that parents from a low-SES are completely dependent towards institutions such as public schools to provide education.

This certainly does not propel the growth of countries like India in the long run where majority of the population is from low-SES and is especially ruinous in the context of a global pandemic. To determine the SES of the parents, researchers have first asked them their primary source of income and then a range/ estimate for the monthly household income. Since our focus lies in education in the times of lockdown, internet connectivity plays a huge role in the quality of education received by the child. The next two questions ask for the number of usable electronic devices available in the household and the number of computers available per child.

Socioeconomic status has been constructed using the questions: Q5, Q7, Q13N and Q14.

- Q5 What do you do in your main job? (Eg. teach high school students, administrative jobs, manage a sales team). This was an open question that was recorded into an ordinal variable following the list of occupations described in the one-digit ISCO (International Standard Classification of Occupations).

  91 Elementary trades and related occupations
  92 Elementary administration and service occupations
  41 Administrative occupations
  42 Secretarial and related occupations
  61 Caring person

- Q7 In a normal month, what is your total household income? This variable was recorded by grouping the income level reported in deciles of income within each country.
- Q13N is composed of How many usable devices are there in the house? (Smartphones, tablets or iPads, laptops, desktops).
- Q14 How many computers per child have you got at home?

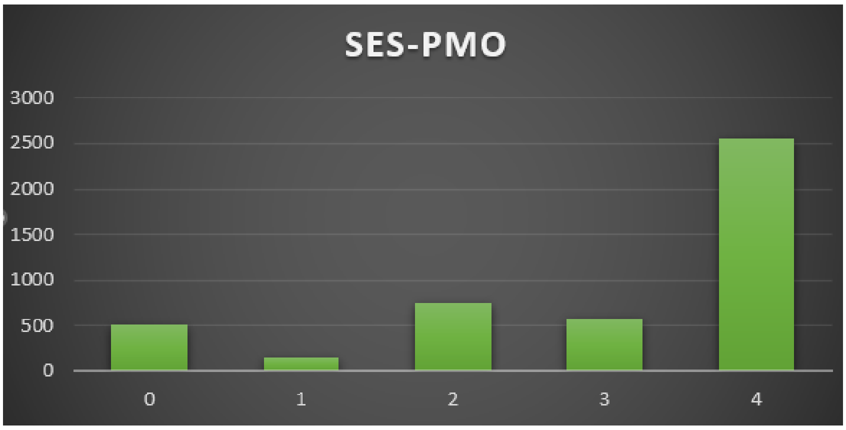
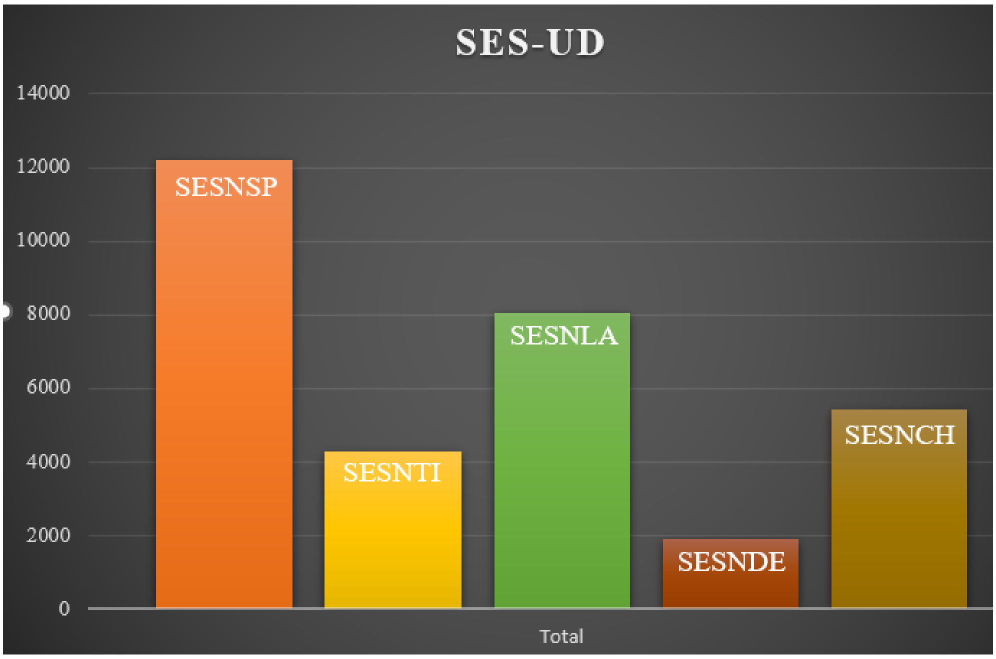


Figure 3.1: The above histogram is of the occupations of the parents who answered the survey. The occupations were classified into five groups as specified by the *International Standard Classification of Occupations*.



The above histogram specifies the number of electronic devices in a particular household with

1. **SESNSP** standing for number of usable smartphones
2. **SESNTI** standing for number of usable tablets or ipads
3. **SESNLA** standing for number of usable laptops
4. **SESNDE** standing for number of usable desktops
5. **SESNSP** standing for number of usable computers

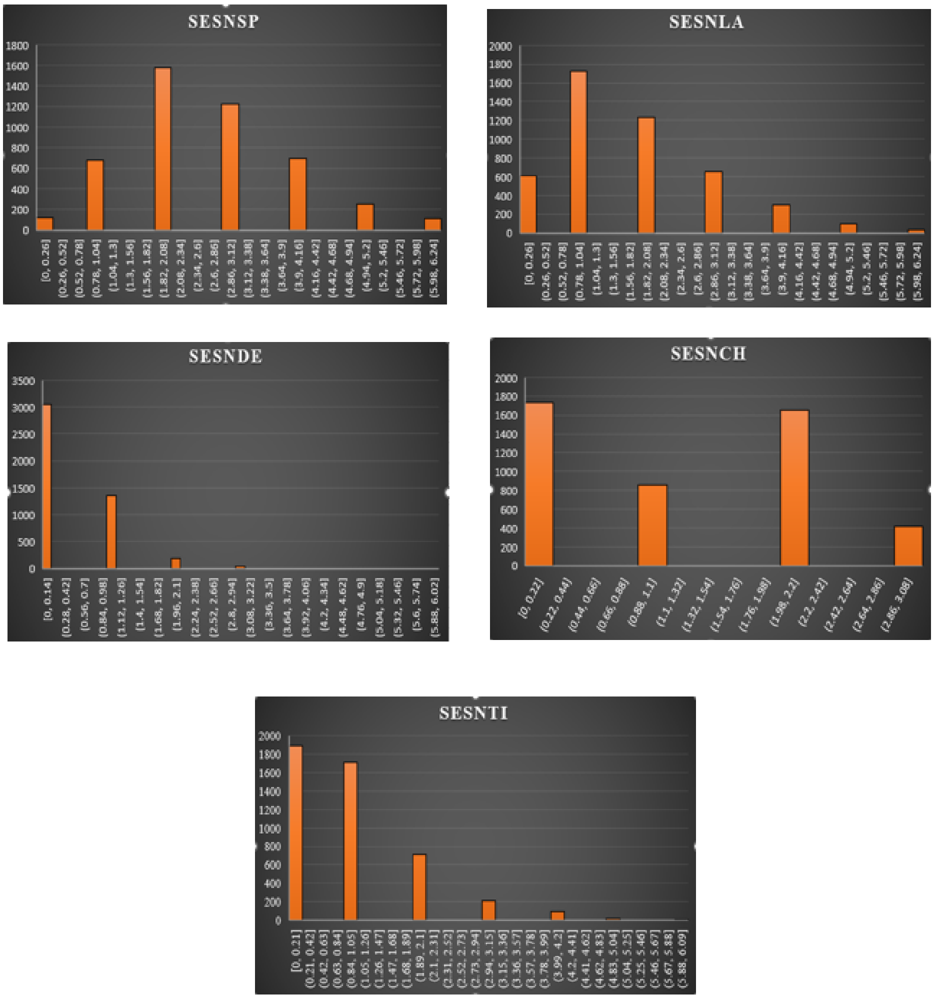


Figure 3.2: The above histograms reflect the actual responses of parents for the five variables specified on the previous page.

3.3 Parental acceptance and confidence in the use of technology (CON)

When a child attends school regularly, his/her mindset, behaviour, interpretation and decisions are affected by the viewpoint prevalent at his/her school as well as home. This also aids in forming a firm perspective since the individual in question is provided with options as to what to believe in. But the constancy of a home environment can be detrimental in the sense that the child develops a one-dimensional outlook towards the world and may not be ready to accept ideas which are not accepted by his/her parents. Parental acceptance and confidence in the use of technology is crucial because if the parents have the necessary IT skills to support an online education and the belief that such an education is useful then the child will also put faith in the schoolwork provided and give it the necessary importance. This variable was evaluated as a scale.

Parental engagement scale was constructed as a second-order construct, with constructs measuring the parents' level of parental acceptance and confidence in the use of technology as 'tools', 'for social purposes' and 'self- perceived capacity'. The items asked parents about the frequency with which they carry out different activities using technology (response options: Always, Often, Occasionally, Rarely Never), and how confident they felt carrying out these activities (response options: Not at all confident, Slightly confident, Moderately confident, Quite confident, Extremely confident).

Parental acceptance and confidence in the use of technology= tool + social + capacity.

- tool = Q22_1 + Q24_1 + Q24_5
- social=Q21_4 + Q21_5 + Q21_6 + Q24_12;
- capacity=Q24_2 + Q24_3 +Q24_4 + Q24_6 + Q24_7 + Q24_8 + Q24_9 + Q24_10+Q24_11+ Q21_7

CHAPTER 4

ANALYTICAL STRATEGY

4.1 Structural Equation Modelling

"Without data you are just another person with an opinion"

The world that we live in, on the surface seems intimidating due to the unvarying and firm structure that it possesses, but students of the fields of statistics, languages and mathematical philosophy will know that the very framework of every aspect of our lives is unsteady. Taking languages for example. The meaning of every word we know can be traced down to some intuitive understanding of a concept or idea and as the quote above aptly states, such convictions are redundant when drawing inferences reliably. Every assumption must be quantified and every hunch should be followed upon in an organized fashion. This is where structural equation modelling comes into play. In disciplines such as behavioural and social sciences and econometrics, where many of the fundamental ideas cannot be expressed in a precise manner, SEM is a vital tool to propel both experimental and observational research.

Formally, Structural Equation Modelling (SEM) is an umbrella term for a diverse set of multivariate techniques employed to confirm and evaluate certain pre-assumed causal relationships. It definition of SEM was articulated by the geneticist Sewall Wright, economist Trygve Haavelmo and cognitive scientist Herbert A.

Although Structural Equation Modelling has often been criticised for its tendency to put faith in models without establishing external validity, mathematical formulation issues and potential philosophical bias, its advantages lie in formalizing those concepts that we cannot easily characterize or document. This, along with the fact that it is user-friendly and accounts for errors, make SEM an essential component of the toolkit of statistical research.

4.2 Latent Variables

Latent variables (meaning 'lie hidden' in Latin) are variables whose presence is assured of by our understanding but cannot be measured due to practical hurdles. These are also called hidden or hypothetical variables because they correspond to abstract concepts like categories, behavioural and mental states or data structures. Latent variables are often inferred indirectly through a mathematical model from other observable variables which can be directly measured.

Latent variables serve to reduce the dimensionality of the data (which means we reduce the number of attributes of our collected data to see only those that will be useful in our study and forming our model). They are highly useful when several attributes in a data set can be linked to some concept of intuitive understanding. For example, when people eat at a restaurant, the happiness they link with that experience, is a variable, which in theory, can be measured using a scale but should not be for the simple reason that different people might evaluate the significance of each scale point differently. A more systematic, albeit tedious, way of evaluating this construct is to observe the variables it is affected by like satisfaction with the food, hospitality of the staff and ambience and document these on a scale. Then we can estimate the amount that each of these affect our variable under study and derive the likelihood. Similarly, variables such as happiness of a student at a particular college, intelligence, depression, etc are examples of latent variables.

Since latent variables are measured indirectly, we introduce an error term while computing it to account for the various kinds of errors that might occur during measurement.

While measuring a latent variable η we divide it into two parts

$$\eta = X + \varepsilon$$

X is the observed variable and ε is the error term.

Continuing our analogy of happiness and as a motivating example, we can think η to be happiness while X to be some value which we observe as an answer for the question designed for happiness (e.g. X can be answer for the question: On scale of $1 - 10$ rate your happiness while studying at ISI) and ε is the error term.

Errors are mainly be of two kinds: Systematic and random.

Systematic Errors

The errors whose source of origin is known and which can be rectified are known as systematic errors. They arise as a result of imperfection of the apparatus, for example if a machine has a recurring error at certain finite number of values, we can always rectify it at those values, or if a miscalibrated scale measures weights as higher than they are we can always scale our recorded values accordingly by observing the result obtained for an object whose weight is known prior. These errors can also arise from a faulty question structure for example, while measuring the importance of breakfast if one of the questions is “Do you eat breakfast everyday?”, some respondents may reply with a negative even if they have skipped breakfast on only one day in recent times. A better framed question in this scenario would be “How many days a week do you usually have break?” and then we provide necessary numerical options.

We can think of systematic errors as a consistent or proportional difference between the true and observed values. Systematic errors taken together have a non-zero mean.

Random Errors

As the name suggests, these errors have sources whose origins are either unknown or cannot be determined by the researcher. Extending one of our previous example, if an individual was told to take a survey on a restaurant’s services on a particularly bad day, they would be more likely to be severe and thus give the restaurant a lower score in individual components.

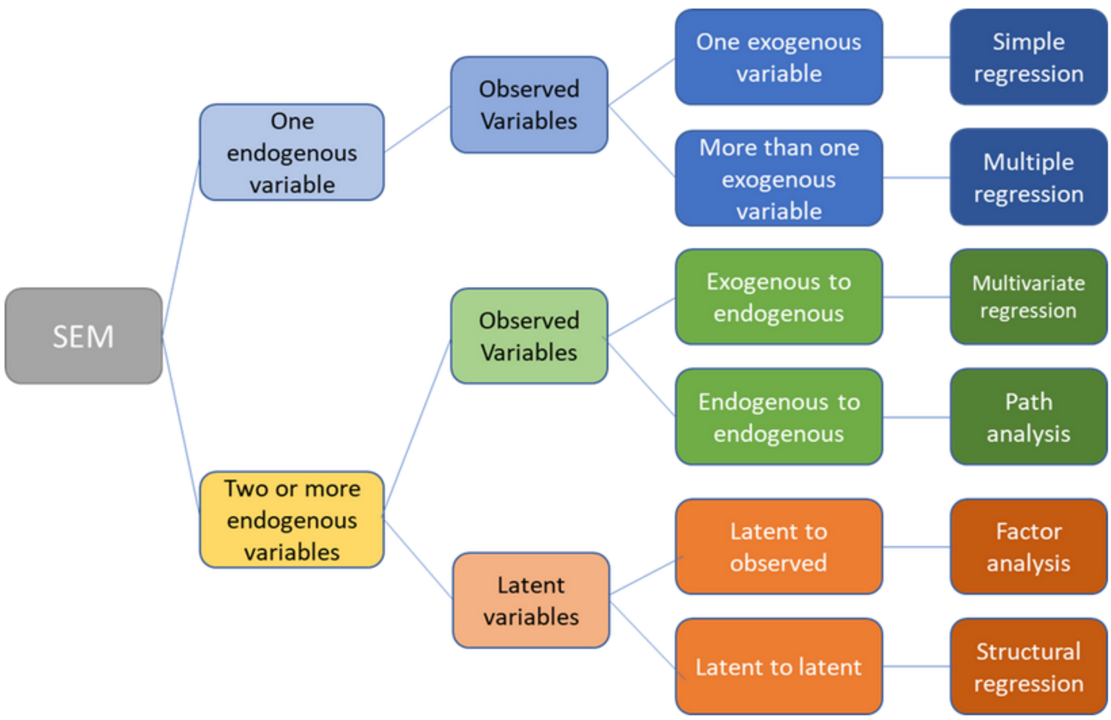
We assume the following about random error:

- $\mathbf{E}(\varepsilon) = 0$, in essence, the mean of the random errors is 0.
- ε_i for each observed variable X_i is independent.
- $\mathbf{Var}(\varepsilon) < \infty \; \forall i$, that is, the error has finite variance.

4.3 Path Diagrams

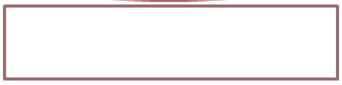
The structural aspect of our model implies associations between variables that represent the ideology under investigation. Variables are either exogenous implying that their variance is not dependent on any other variable in the model or endogenous, meaning their variance is affected by other variables in the model. Informally, exogenous variables can be viewed as those that have arrows pointing from them and endogenous variables are those that have arrows pointing towards them. Variables having arrows both pointing from them and towards them are known as mediating variables. Equations involving all of these form an integral part of SEM. They are a mathematical translation of what is observed and they are estimated with statistical algorithms based on matrix algebra and generalized linear or non-linear models. These equations arise from the experimental or observational data.

Now that we have known what latent, exogenous and endogenous variables are, we can summarize the types of structural equation models through the following flowchart:



The pattern of these mathematical associations among variables can be visualized by means of a path diagram. A path diagram can be best understood by its components:

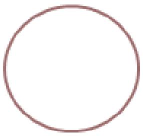
Ovals representing latent variables.



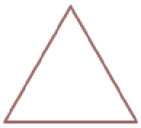
Rectangles indicating the observed variables.

Single headed arrow shows regression or directional path implying causality in the model. The head of the arrow shows the direction of the relation. Usually we have an arrow from the latent variable to the observable as we say, latent variable is an underlying variable influencing the variables we can observe.

Double headed arrow shows covariance or non-directional path.



Circles showing the error or variance terms.

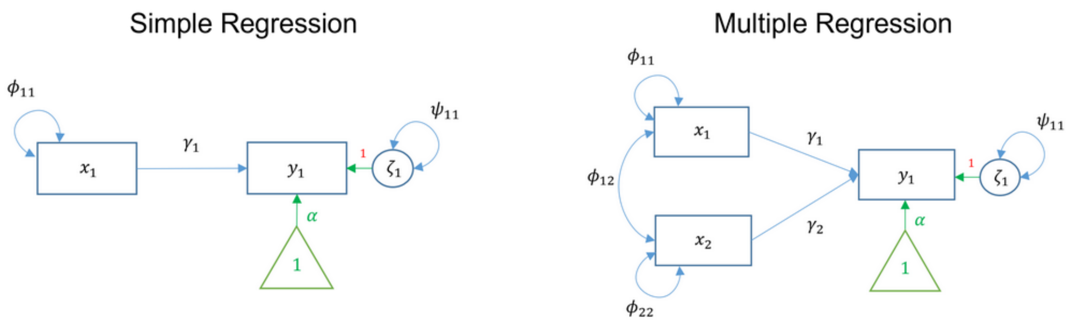


Triangles indicating intercept of a variable, value of which is found out by solving regression equations between dependent and independent variables.

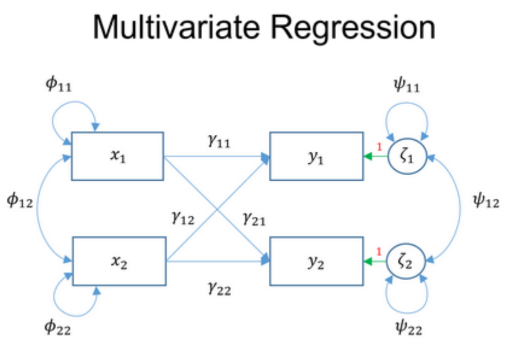
This is highly useful when mathematical rigor and equations would overpower the interlinks the researcher is attempting to evaluate. Path diagrams give one a bird's eye view as to how the various aspects of the data are tied together in the system.

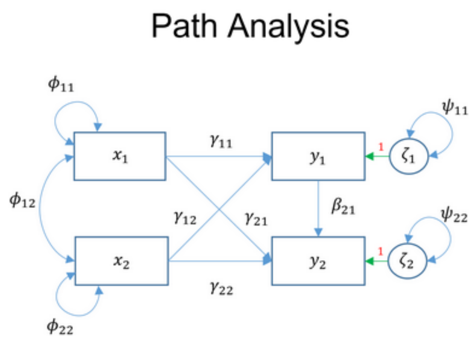
A few models that can be represented by path diagrams are as follows:

- Simple regression and multiple regression both involve one endogenous variable. In simple linear regression, there is one exogenous variable that predicts one endogenous variable. In multiple regressions, multiple exogenous variables (with covariance) predict a single endogenous variable. The exogenous variables each have variance. The endogenous variable has an intercept (the triangle) and residual variance.

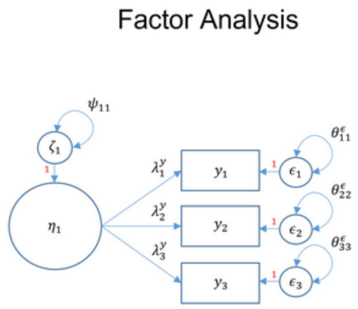


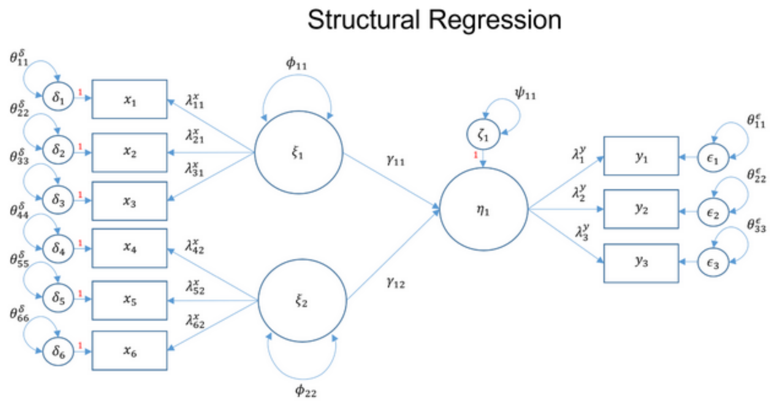
- Multivariate regression and path analysis both have two or more observed endogenous variables. The main difference is that multivariate regression has only exogenous variables predicting endogenous variables. However, in path analysis, endogenous variables can also predict other endogenous variables.





- Factor analysis and structural regression both have two or more endogenous variables. The main difference is that factor analysis looks at how a latent variable can predict observed variables. Structural regression can use latent variables to predict other latent variables.





In our project we lay more emphasis on Factor Analysis.

4.4 Factor Analysis (FA)

Factor Analysis is a statistical technique to describe variability, inter-dependencies in the observed variables in terms of unobserved/latent variables, in this context, known as factors. In a sense, we group the data based on what we find common between our variables. Factor analysis simplifies the data that we handle and thus enables us to draw inferences with greater rapidity. It is of great utility in psychology, biology, marketing, operations research, etc where several observed variables are thought to reflect a few manageable latent variables. Observed variables are modelled as linear combinations of potential factors and errors are taken into account. The coefficient of a potential factor in a linear combination is known as the factor loading and it quantifies the extent to which the variable is related to a given factor. We usually take a large number of observed variables as compared to the latent variables to make the model over-identifiable.

Key terms in FA:

1. **Observed Variable**: The observed variable is the factor that we use to measure a construct. This includes the data we record during research.

2. **Latent Variable**: As we have explained above, latent are variables that can only be inferred indirectly through a mathematical model from other observable variables that can be directly observed or measured.

3. **Factor Loading**: Factor loading is a number that measures the correspondence between the observed variables and latent variables. The values of factor loadings are usually between zero and one and higher values indicate stronger correlation between the relevant variables.

There are different methods we use in various stages of factor analysis from the data set.

1. *Principal Component Analysis* is a method for reducing the dimensionality (number of useful observable variables are extracted out of sample as a whole) of the data, while also increasing its interpretability and minimising information loss. It does so by creating new uncorrelated variables that successively maximize variance.

2. *Common Factor Analysis* uses covariance matrices to determine which variables have the highest amount of correlation and grouping these into a single factor.

3. *Image Factoring* is based on the correlation matrix again and uses predicted variables using ordinary least squares regression method.

4. *Maximum liklihood method* assumes that the data is normally distributed. A reliability coefficient is proposed to indicate quality of representation of interrelations among attributes, which tells us whether to reject or accept a factor solution in our model.

5. *Alfa factoring* and *weight square* are other methods.

Types of Factor Analysis

There are essentially two types of factor analysis which we will explore in detail in the forthcoming sections.

1. *Exploratory Factor Analysis:* Here the researcher does not make any assumptions about prior relationships between factors. Every variable is assumed to be related to every factor. This can help us in analysing the causality structure that exists in the model.

2. *Confirmatory Factor Analysis:* Here, the researcher assumes that variables are related to specific factors and uses pre-established theories to confirm their expectations of the model.

Assumptions of Factor Analysis

1. There are no outliers in the data.

2. The sample size is greater than the number of factors.

3. Factors and thus observed variables should be continuous.

4. There is the assumption of linearity, that is, correlations are linear and the observables are linear combinations of latent variables.

5. Reasonably high correlations between observed variables should be present.

4.5 Multiple Linear Regression

Multiple linear regression is used to estimate the relation between two or more explanatory variables and one dependent variable. It is highly useful to estimate the value of the dependent variable at a certain data point and also to find the strength of the association between the dependent and explanatory variables. In essence, multiple linear regression is an extension of simple linear regression because it involves more than one explanatory variable.

4.5.1 Formula and calculations of MLR

$$y_i = \beta_0 + \beta_1 x_{i1} \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where for $i = n$ observations
y_i = dependent variable
x_i = explanatory variables
β_0 = y-intercept
β_p = slope coefficients for each explanatory variables
ε_i = the model's error term

4.5.2 Assumptions of MLR

1. *Homogeneity of variance*: The size of the error that we might make in our predictions of the dependent variable does not change significantly across values of the explanatory variable, i.e., the variance of residuals is constant at very point in the model.
2. *Independence of observations*: The observations in the dataset are collected using statistically valid sampling methods and there is little or no correlation between the observed variables. If the correlation is too high between two particular variables, only one of them should be used in our analysis, or both of them are not used at all.
3. *Multivariate Normality*: In MLR, we assume that the residuals follow normal distribution
4. *Linearity*: The best-fitting line is a straight line/plane/hyperplane.

4.5.3 Limitations of MLR

Since linearity is assumed, MLR is not useful when the dependent and explanatory variables have non-linear relationships. Also when a number of the explanatory variables are correlated highly among themselves, we usually assume that an underlying factor is influencing them to have such behaviour, such as which cannot be measured directly, i.e., a latent variable.

4.5.4 Multiple Linear Regression vs. Factor Analysis

The primary point of disparity between multiple linear regression and factor analysis arises in the way we view the data. In multiple linear regression, we estimate certain relationships between a dependent variable and two or more explanatory variables. Both the dependent and explanatory variables are observables here. In factor analysis, the observable variable is assumed to be a linear combination of latent variables, whose influences on the observed variable, we need to find out.

4.6 Exploratory Factor Analysis

In statistics and data modelling, exploratory factor analysis is a statistical technique to uncover the underlying structures prevalent in a dataset when the researcher has no a priori hypothesis about factors or patterns of measured variables. It is commonly used when developing a scale and it serves to identify a set of latent constructs concealed behind the observable variables.

Exploratory data analysis seeks to reveal structure, or simple descriptions in data. We look at numbers or graphs and try to find patterns. We pursue leads suggested by background information, imagination, patterns perceived, and experience with other data analyses

-*Persi Diaconis*

The word exploratory is indication enough that EFA is essential to determine or explore the relations between the given data. Confirmatory Factor Analysis which is the technique that we utilise in this paper, allows the researcher to test the hypothesis that a relationship exists between the observed variables and their underlying latent factors.

EFA is based on the common factor model where manifest variables are expressed as a function of common factors, unique factors and errors of measurement. This is not something we will be needing in our project.

However, given just a set of observable variables with no idea how to come up with a model about how to fit the data, how do latent variables come to existence?

We first check whatever we know about our data: especially the correlation matrix of our sample. If there is low correlation between any two different observed variables, our data has no underlying factor that explains our data. However if there is a high enough correlation between two observed variables, we can assume the existence of an underlying factor that explains the variables. There are a number of tests to see whether our data is fit for carrying out EFA like Bartlett's sphericity test and the Kaiser-Meyer-Olkin(KMO) test, all based on observing the correlation matrix of the sample data.

Once we have checked that our data is fit for carrying out EFA, we try to extract the latent variables that can explain our data sufficiently well. This is done by some methods like Maximum Likelihood(ML) and Principal Factors(PF). We usually assume that the number of latent variables which explain our sample data should not exceed the number of observed variables. The procedure of EFA relies on the extraction of eigenvalues and eigenvectors as they summarise variance in the given correlation matrix. For practical purposes, it is useful to view eigenvalues as representing the variance in the observables explained by the successive latent variables/factors. To select the minimum number of latent variables required to explain a certain percentage (more than 75-80 % is considered sufficiently good) of total variance of our sample data, we use tests like scree plot test, parallel test, etc. which again performs tests on the eigenvalues of the correlation matrix.

We then try to calculate the factor loadings, which give a measure of the correlation between an observed variable and a latent variable. This is done by some calculations done using eigenvectors and eigenvalues, or by base change (multiplying on the left of the correlation matrix by an invertible matrix, and multiplying its inverse on the right). Axis rotation is another last procedure carried out to obtain a further simpler solution (enhance the factor loadings) of our model. We shall again not give too much detail about.

Till now we do not know what kind of latent variables are these, i.e., what they should be named as like Happiness or Depression, that explains our observed variables. The naming of the latent variables is done by the statistician based on real-life scenario and logic, by noting their relationship with the nature of observed variables.

We present the following example to clarify what we have explained so far:

1. We have put forward a survey asking interested participants to rate six personality traits they see in themselves on a scale of 1 to 5. These traits being *outgoing(1), sociable(2), hard-working(3), dutiful(4), warm-hearted(5)* and *helpful(6).*

2. Suppose we obtain the following correlation matrix A where the entry $(A)_{ij}$ gives the correlation between observables i and j.

$$\begin{bmatrix} 1 & & & & & \\ 0.58 & 1 & & & & \\ -0.13 & 0.07 & 1 & & & \\ -0.02 & 0.33 & 0.34 & 1 & & \\ -0.04 & 0.27 & 0 & 0 & 1 & \\ -0.22 & -0.26 & 0.03 & -0.25 & 0.52 & 1 \end{bmatrix}$$

   Note that we have kept the matrix lower triangular so as not to repeat the values, or else the matrix should be symmetric. Every variable should be perfectly correlated with itself and so, 1's appear along the diagonal. This matrix will be giving us eigenvalues and eigenvalues required for our calculation. However it is useful to note that we have a notable correlation between observables *outgoing(1)* and *sociable(2)* and between *warm-hearted(5)* and *helpful(6).*

3. Next we see the table which gives the eigenvalues of the correlation matrix obtained corresponding to the latent variables (here a maximum of 6 as it was a 6×6 matrix) and the percentage of variance explained by each latent variable:

Latent variable	Eigenvalue	Percentage of variance	Cumulative percentage
1	1.87	31.21	31.21
2	1.48	24.71	55.93
3	1.36	22.7	78.62
4	0.67	11.23	89.86
5	0.4	6.7	96.56
6	0.21	3.44	100

   Here the eigenvalues have been arranged in descending order. The percentage of variance is given by the ratio of the eigenvalue and the sum of all eigenvalues, multiplied by 100, that is, the percentage the eigenvalue contributes to the total sum of eigenvalues. By various tests, we select 3 latent variables out of the six, after observing they explain 78.62 % of the total variance of our sample data.

4. Next by certain operations based on these eigenvalues and eigenvectors, we get the percentage of variance in observed variable explained by the three selected latent variables. This gives us the factor loadings between a latent variable/factor and an observable.

	Component 1	Component 2	Component 3
outgoing	0.67	0.19	-0.54
sociable	0.8	0.47	-0.12
hard-working	0.16	0.12	0.79
dutiful	0.54	0.07	0.66
Warm-hearted	-0.17	0.91	0
helpful	-0.66	0.61	0.02

Note that, taking the square of the values in the columns and adding these values up gives the amount of variance of all observable variables explained by that component/latent variable. For example, for component 1,

$$0.67^2 + 0.8^2 + 0.16^2 + 0.54^2 + (-0.17)^2 + (-0.66)^2 = 1.8706 \sim 1.87$$

And if we observe the table above, this is the latent variable 1 which had the greatest eigenvalue 1.87. Similar calculations work for components/latent variables 2 and 3.

5. Once the factors/ latent variables are extracted, we rotate the factors to foster their interpretability. This is because, given any multiple factor model such as this, there exists an infinite number of equally good-fitting solutions (each represented by a different factor loading matrix). We specifically aim for a solution where each factor loads highly (high or not: determined hypothetically and not by any strict convention) on some particular observables and negligibly on the rest of them. We achieve this by factor rotation, which is, just a mathematical transformation (more precisely we take the 6×6 matrix of factor loadings here and obtain another factor loading matrix similar to the one we have already obtained by change of basis operation). There are two types of rotation:

   - orthogonal: factors are constrained to be uncorrelated (columns of base-change matrix are orthogonal)
   - oblique: factors are allowed to be correlated (columns of base-change matrix are not orthogonal)

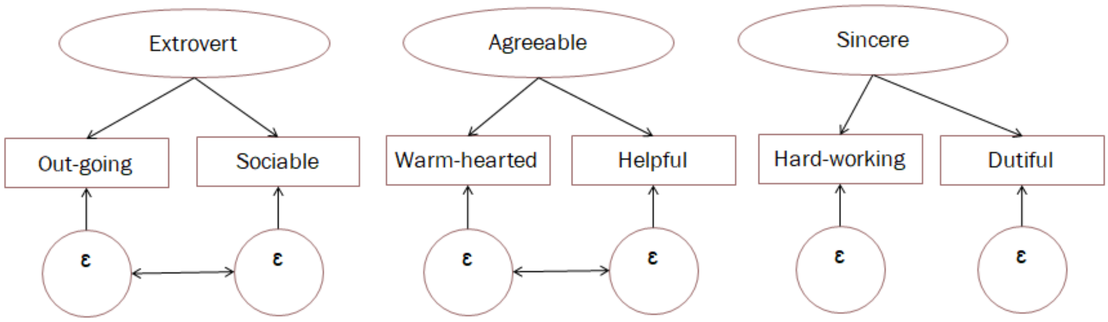
   We have carried out orthogonal rotation to obtain:

	Component 1	Component 2	Component 3
outgoing	0.84	-0.14	0.23
sociable	0.9	0.07	0.25
hard-working	-0.13	0.05	0.8
dutiful	0.22	-0.16	0.81
Warm-hearted	0.23	0.89	0.09
helpful	-0.3	0.84	-0.11

   Factor rotation does not change fit of the solution model to the given sample data.

6. We started out with just our correlation matrix and observed, every latent variable explains some part of the observables (all eigenvalues are non-zero). But three factors are sufficient and now, factor 1 explains variance in *outgoing* and *sociable* more than others (higher factor

loadings), factor 2 loads higher on *warm-hearted* and *helpful* than others while factor 3 loads higher on *hard-working* and *dutiful* than others. Our EFA model, found out from sample model is:



4.7 Confirmatory Factor Analysis

In statistics, confirmatory factor analysis is a form of factor analysis which is commonly used in social science research. It is used to test whether the measures of a latent variable are consistent with the researcher's understanding of the latent variable. The foremost objective of CFA is to ascertain whether the data is accordant with a hypothesized measurement model. CFA is different from EFA in the sense that the researcher constraints certain relationships in the dataset based on a priori hypotheses. This forces the model to be consistent with the developed theory. Also EFA operations are all carried out using the correlation matrix of sample data, whereas here in CFA, we look into the variance-covariance matrix of sample as well as our hypothesised model. Model fit measures are then used to assess how well the proposed model captures the required covariances.

Steps in Confirmatory Factor Analysis

1. **Specify the latent variable:** We begin confirmatory factor analysis by specifying and defining the latent variable that we want to analyse and comment upon. Establishing a baseline for the latent variable makes it possible for us to evaluate the accuracy of the observed variables. Definitions and explanations of the relevant variables in this project have been provided towards the beginning.

2. **Collect the data:** We gather the information we want to use in our confirmatory factor analysis. Large sample size ensures accurate analysis.

3. **Establish consistent parameters:** Using our statistical modelling software, we then establish standardized parameters to evaluate the latent and observed variables. We decide what measurement system we want to use as the standard and allow the software to convert all other values to that measurement.

4. **Compute the data:** We use *AMOS* and *R* to compute factor loadings for our data.

5. **Interpretation:**We review the factor loadings to determine how well each observed variable relates to the latent variable. We also compare the model implied variance-covariance matrix with the sample variance-covariance matrix and see how close they are.

4.7.1 Model identification in CFA

We have our hypothesized model, some unknown parameters we need to estimate, and our sample data. What should we do now?

We should find out whether our data provides us with enough information so as to estimate the parameters in our hypothesised model. We need to see if our model is *under-identified, just-identified* or *over-identified*.

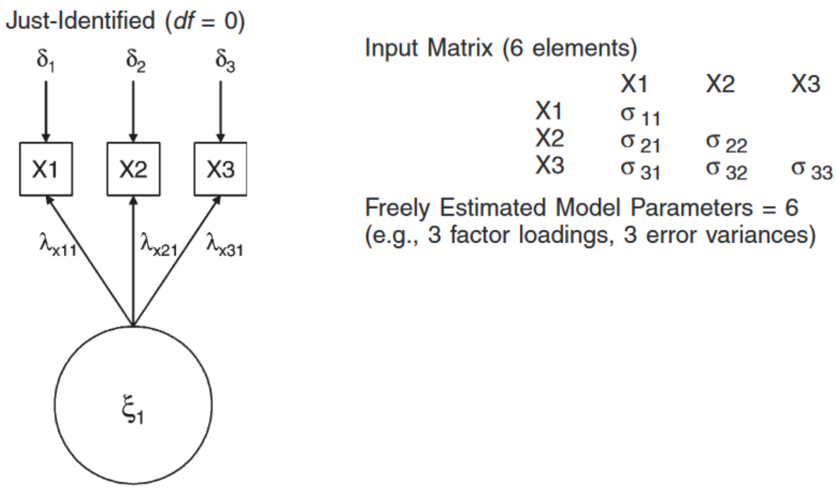
The number of known parameters are the known values in the variance-covariance matrix of our sample data. If we have p observable variables in our dataset, then the number of known parameters are the variances of the variables and the pairwise covariances, which are $\frac{p(p+1)}{2}$ in number (i.e. the upper or lower triangular matrix entries of the variance-covariance matrix). Now what we do not know (unknown parameters) are:

- variances and covariances of the factors in our model
- factor loadings which the factors load on the observable variables
- random errors associated with each observed variable

The degree of freedom of our model is given by:
df = number of known parameters - number of unknown parameters
This leads to the following three cases:

1. When the number of known parameters is equal to the number of unknown parameters, i.e. the degree of freedom of our model is 0, we call it a just identified model.

   For example consider the following model, where the variance of the latent variable (denoted by ξ_1) is fixed to 1. We then have three factor loadings (denoted by the λ's) and the three error terms (denoted by the δ's) as our unknown parameters:



   The factor loadings in this case can easily be calculated by considering the following three equations involving three unknowns:

$$\lambda_{x11} \times \lambda_{x21} = \sigma_{21}$$

$$\lambda_{x11} \times \lambda_{x31} = \sigma_{31}$$

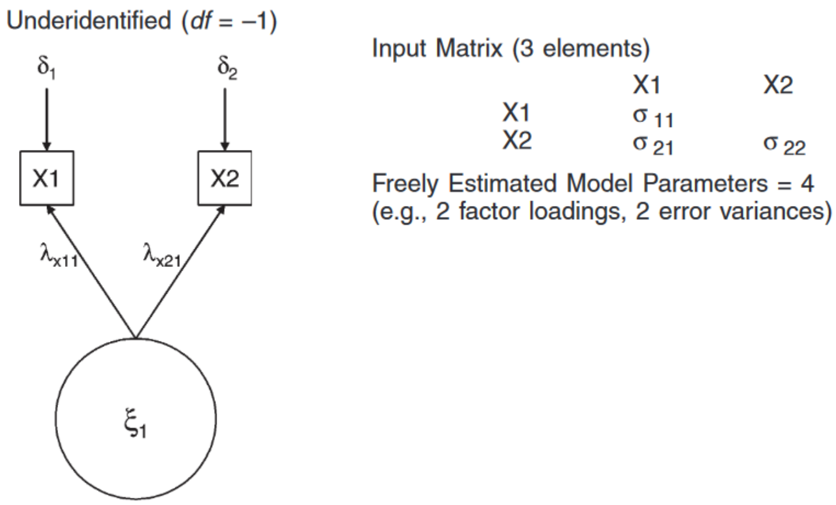
$$\lambda_{x31} \times \lambda_{x21} = \sigma_{32}$$

Now we can calculate the error variance by subtracting the square of the factor loading from the variance of the latent variable, here 1:

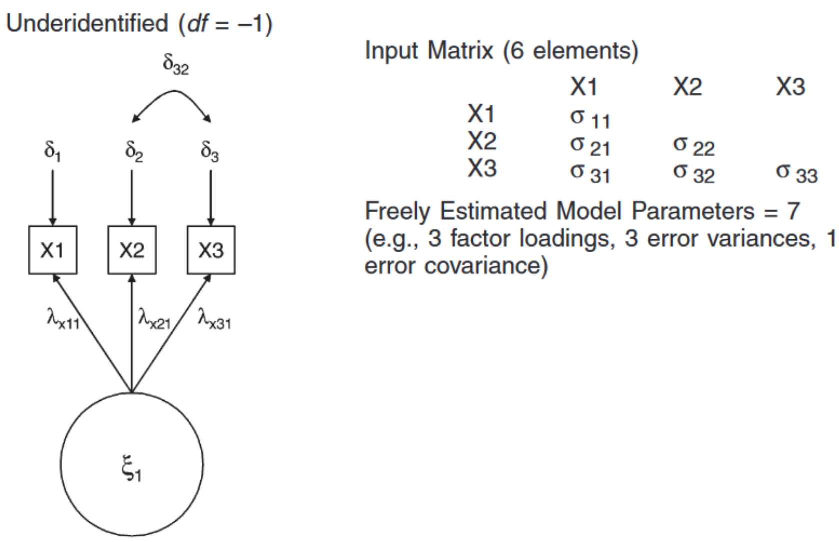
$$\delta_1 = 1 - \lambda_{x11}^2$$
$$\delta_2 = 1 - \lambda_{x21}^2$$
$$\delta_3 = 1 - \lambda_{x31}^2$$

2. When the number of known parameters is less than the number of unknown parameters we need to estimate, we call it a just identified model. If it makes sense, the degree of freedom of our model is then negative.

   For example, consider the following model, even after we fix the variance of the latent variable to 1, we have 3 known parameters and 4 unknown parameters:

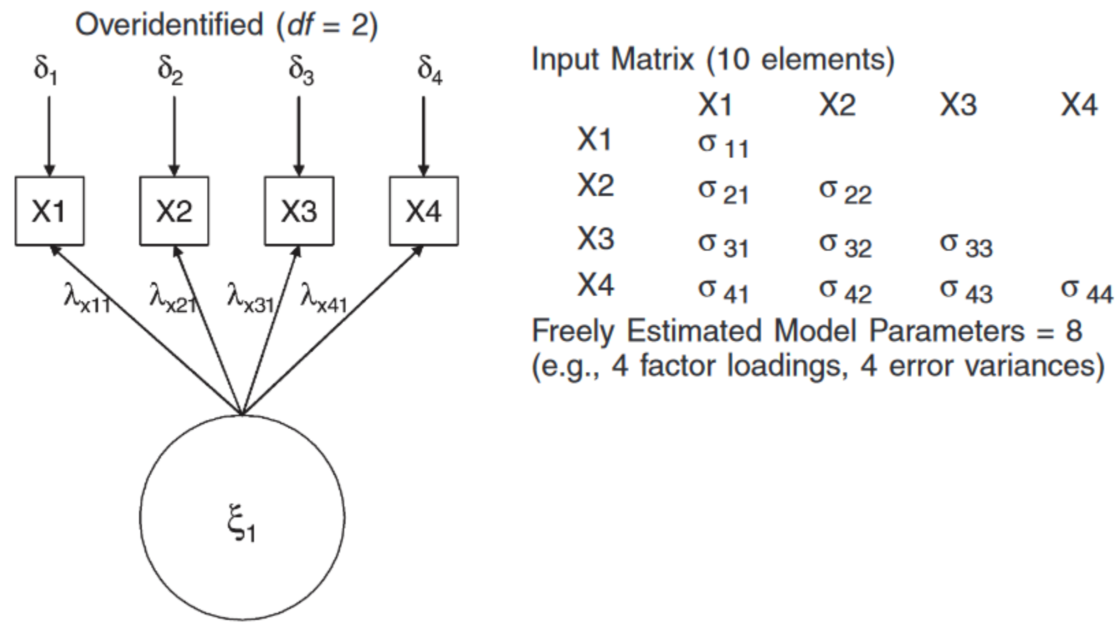


   Another example is this model, where we also have an error covariance (which actually represents the covariance between our observables):

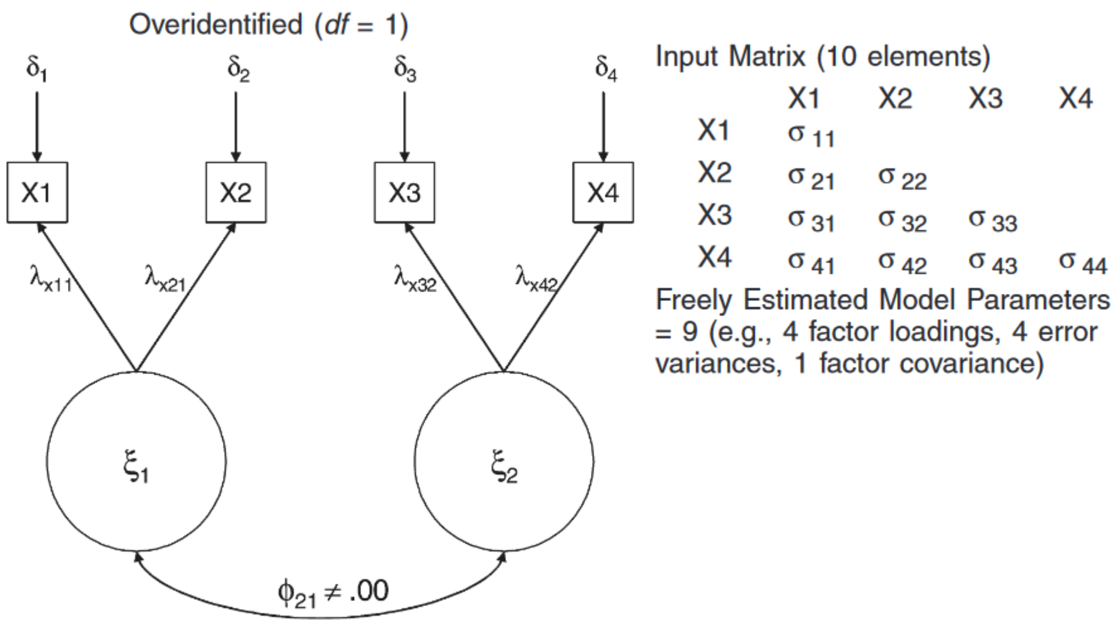


3. When the number of known parameters is greater than the number of unknown parameters, i.e. the degree of freedom of our model is positive, we call it a over-identified model.

For example, consider the following model, where again, we have fixed the variance of the latent variable to 1. We now have an over-identified model as the number of known parameters is 10 and that of unknown ones is 8:



Another model here would serve as an example; note that we have two latent variables in our model with a non-zero covariance between them:



To estimate our parameters, we can make an under-identified model just-identified or even over-identified. This is usually done by the following methods:

- *Variance standardization method* - Constrain variance of latent variable to 1 : This yields a standardized solution.

- *Marker method* - Constrain the factor loading of one item to 1 : we have the freedom of choosing which factor loading to set to one and calculate other factor loadings in terms of that. However, the variable which is hypothesised to have greater correlation with the latent variable is usually set to 1 by convention, which is because of the simple reason that setting the factor loading of the least correlated variable to 1 would cause the other factor loadings, calculated on the basis of the fixed one, to blow up.

In all our previous examples as well as in the models we have come up with, we have used the first method.

Now that we our model and we have our estimated parameters, we find our model-implied variance-coavraince matrix:

$$\Sigma(\theta) = \Lambda\Psi\Lambda' + \Theta_\delta$$

where Σ is the model implied variance covariance matrix
θ is composed of the parameters:
Λ - The factor loadings
Ψ - The variance covariance matrix of the latent factors (i.e. the variance of η if we have just one latent variable which is a scalar, and usually this scalar again is fixed to 1)
Θ_ε - The variance covariance matrix of the residuals.

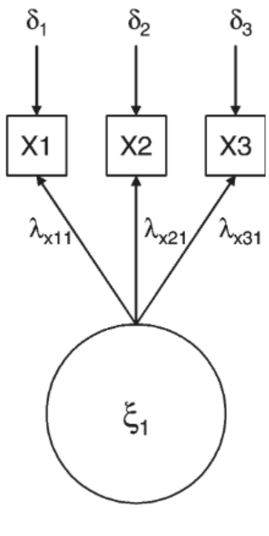
Now as statisticians, we actually deal with estimates of our data, so the model-implied variance-covariance matrix with real data is usually depicted by:

$$\Sigma(\hat{\theta}) = \hat{\Lambda}\hat{\Psi}\hat{\Lambda}' + \hat{\Theta}_\delta$$

For example, in a three-item one-factor model, we have the model implied variance-covariance matrix as:

$$\Sigma(\theta) = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} \begin{pmatrix} \psi_{11} \end{pmatrix} \begin{pmatrix} \lambda_1 & \lambda_2 & \lambda_3 \end{pmatrix} + \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{pmatrix}$$

In the model we have dealt with in our previous examples, namely:



We get our matrix as:

$$\Sigma(\theta) = \begin{pmatrix} \lambda_{x11} \\ \lambda_{x21} \\ \lambda_{x31} \end{pmatrix} \begin{pmatrix} \xi_1 \end{pmatrix} \begin{pmatrix} \lambda_{x11} & \lambda_{x21} & \lambda_{x31} \end{pmatrix} + \begin{pmatrix} \delta_1 & 0 & 0 \\ 0 & \delta_2 & 0 \\ 0 & 0 & \delta_3 \end{pmatrix}$$

And now that we have our model-implied variance-covariance matrix, we have the obvious urge to compare it with the variance-covariance matrix of the sample data. This is exactly what we do in comparing our model and checking how close it is to our sample data. We shall delve into more about how we define this measure of closeness and other such measures in our next section *Fit Indices.*

In our project we conduct Confirmatory Factor Analysis using the statistical software *AMOS(Analysis of Moment Structures)*, apart from our usual *R programming language.*

4.8 Fit Indices

Subjects such as social sciences, economics, healthcare, etc are of paramount importance but the variables analysed in these subjects are often latent, i.e. representing some construct of which we have a firm intuitive understanding but little to no ability of measurement. These include measurement of happiness of the population of a country, effect of placebo, and in our present case, parental engagement.Structural Equation Modelling has drastically altered the way we observe and work with latent variables, providing researchers with methods to analyze data in ways that are impossible under general linear models. Many modern scales and measures utilise structural equation modelling to align measures with underlying latent constructs. The general procedure for such analysis is as follows: Researchers create a model, collect the data, and then test whether the model fits the data. There are several ways to evaluate a model's fit to the collected data, but the prevalent one is the use of fit indices. The fit indices are typically normalized and all the values are usually between 0 and 1.
We usually compare our model with **Null Model.** Null model basically assumes that every factor is uncorrelated to every other factor.

4.8.1 Absolute Fit Indices

Absolute fit indices determine how well does **a priori model** fits the sample data and demonstrates which of the proposed models has the most superior fit. These measures provide the most fundamental indication of how well the proposed theory fits the data.

Chi-Square Test

Chi-Square is a standard probability distribution which is used to indicate the closeness of a model with the original data. The test statistic can be computed using the Maximum Likelihood Estimation where we estimate the value of statistic for given data with the highest probability. A smaller chi square implies that model is more plausible, that is, fits the data well. The chi-squared test indicates the difference between observed and expected covariance matrices. Values closer to zero indicate a better fit; smaller difference between expected and observed covariance matrices.

The assumptions for this test is

- Large sample size
- Multivariate normality i.e the joint sampling distribution of the factors should be multivariate normal

Since the sample size is large in our case study, we can make these assumptions.

One disadvantage of the chi-squared test of model fit is that researchers may fail to reject an inappropriate model in small sample sizes and reject an appropriate model in large sample sizes. Also, a sizeable chi-squared test with a corresponding small p-value indicates that the model does not fit the data. To overcome these problem, other measures of fit have been developed.

SRMR (Standardized Root Mean Residual)

It is a residual based fit index. To explain this fit index we recall linear regression. The *SSE* in linear regression is calculated by summing the squares of residual while in *SRMR*, as is suggested by the name, we compute the average and then take out the square root of the average. Standardization scales down the metric of different observed variables to get values of SRMR between 0 to 1. *Hu* and *Bentler* suggested in 1999 that SRMR value less than 0.08 indicates a good fit. The lower the value of *SRMR* the more plausible is the model. This absolute fit index can be indicated as follows:

$$SRMR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}(\frac{s_{ij}-\hat{\sigma}_{ij}}{s_{ii}s_{jj}})^2}{p(p+1)/2}}$$

where

- s_{ij} is the ij th component of the sample covariance matrix
- $\hat{\sigma}_{ij}$ is the ij th component of the covariance matrix of hypothesized model $\Sigma(\hat{\theta})$
- p is the number of observed variables

RMSEA (Root Mean Square Error Approximation)

This fit index is similar to the one above, but the only difference in here we calculate the root mean squares of errors. Root Mean Square Error of Approximation (RMSEA) is a measure that attempts to correct the tendency of chi-square statistics to reject models with large samples. It avoids issues of sample size by analyzing the discrepancy between the proposed model, with optimally chosen parameter estimates, and the population covariance matrix. RMSEA is considered very good if it is equal to or less than 0.05, good between 0.05 and 0.08, mediocre between 0.08 and 0.10 and unacceptable if it is higher than 0.10. It is estimated by the formula:

$$RMSEA \sim \sqrt{max(\frac{\chi^2_{proprosed\ model} - df_{proprosed\ model}}{df_{proposed\ model} \times (N-1)}, 0)}$$

where N is the sample size and *df* the degrees of freedom. Additionally, RMSEA provides a one-sided test with the following hypotheses:

- H_0: the $RMSEA$ equals 0.05 (a close-fitting model), and so if the p-value $\geq$ 0.05 (not statistically significant) the fit of the model is close.
- H_a: the $RMSEA$ is higher than 0.05 and so if the p-value $<$ 0.05 the fit of the model is bad.

Relative Fit Indices

Relative fit indices, also called *incremental fit* or *comparative indices*, includes a factor that represents **deviations from a null model**. It compares the chi-square for the proposed model to a null model. This null model almost always contains a model in which all of the variables are uncorrelated, and as a result, has a very large chi-square (indicating poor fit). It is considered very good if the nearest is 1 and bad if it is less than 0.9. Some relative fit indices are as follows:

Normed Fit Index (NFI)

It is an indepennce-model-based fit index.It compares the error between the covariance matrix of hypothesized model and the null model.

$$NFI \sim 1 - \frac{\chi^2_{proposed\ model}}{\chi^2_{null\ model}}$$

Tucker-Lewis Index (TLI)

It is also an independence-model-based fit index. The major issue with NFI is it sometimes underestimates the parameter i.e gives the value which is significantly lower than the actual parameter (termed as negative bias). To get around this Tucker and Lewis came up with an upgraded version namely Tucker-Lewis Index (TLI). TLI is basically the normed fit index with the problem of negative bias resolved. Tucker-Lewis index (TLI) is also known as a non-normed fit index (NNFI). It is a combination of a measure of parsimony with a comparative index between the proposed model and the null model. It is considered very good if it is equal to or greater than 0.95, good between 0.9 and 0.95, suffering between 0.8 and 0.9 and bad if it is less than 0.8.

$$TLI \sim \frac{\frac{\chi^2_{null\ model}}{df_{null\ model}} - \frac{\chi^2_{proprosed\ model}}{df_{proprosed\ model}}}{\frac{\chi^2_{null\ model}}{df_{null\ model}} - 1}$$

CFI (Comparative Fit Index)

It is also an independence-model-based fit index. Comparative fit index (CFI) analyzes the model fit by examining the discrepancy between the data and the proposed model while adjusting for the issues of sample size intrinsic in the chi-squared test, and the normed fit index. It is considered very good if it is equal to or greater than 0.95, good between 0.9 and 0.95, suffering between 0.8 and 0.9 and bad if it is less than 0.8.

$$CFI = 1 - \frac{\lambda_m}{\lambda_b} \sim 1 - \frac{max(\chi^2_{proprosed\ model} - df_{proprosed\ model}, 0)}{max(\chi^2_{null\ model} - df_{null\ model}, 0)}$$

The λ terms are known as non-centrality parameter. The non-centrality parameter is calculated as the normalized difference between the parameter given by the model of null hypothesis and model of alternative hypothesis. Here the λ_m is the NCP of the hypothesized model while λ_b is the NCP of the baseline model.

4.8.2 Cronbach's alpha

Cronbach's alpha is a measure of internal consistency, that is, how closely related a set of items are as a group. It is considered to be a measure of scale reliability. Cronbach's alpha can be written as a function of the number of test items and the average inter-correlation among the items. Below, for conceptual purposes, we show the formula for the Cronbach's alpha:

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$$

Here N is equal to the number of items (observed variables) , $\bar{c}$ is the average inter-item covariance among the items and $\bar{v}$ equals the average variance.

One can see from this formula that if one increases the number of items, there is an increase in Cronbach's alpha. Additionally, if the average inter-item correlation is low, alpha will be low. As the average inter-item correlation increases, Cronbach's alpha increases as well (holding the number of items constant).

Technically speaking, Cronbach's alpha is not a statistical test – it is a coefficient of reliability (or consistency). In general, a score of more than 0.7 is usually okay. However, some authors suggest higher values of 0.90 to 0.95.

CHAPTER 5

CONFIRMATORY FACTOR ANALYSIS USING R

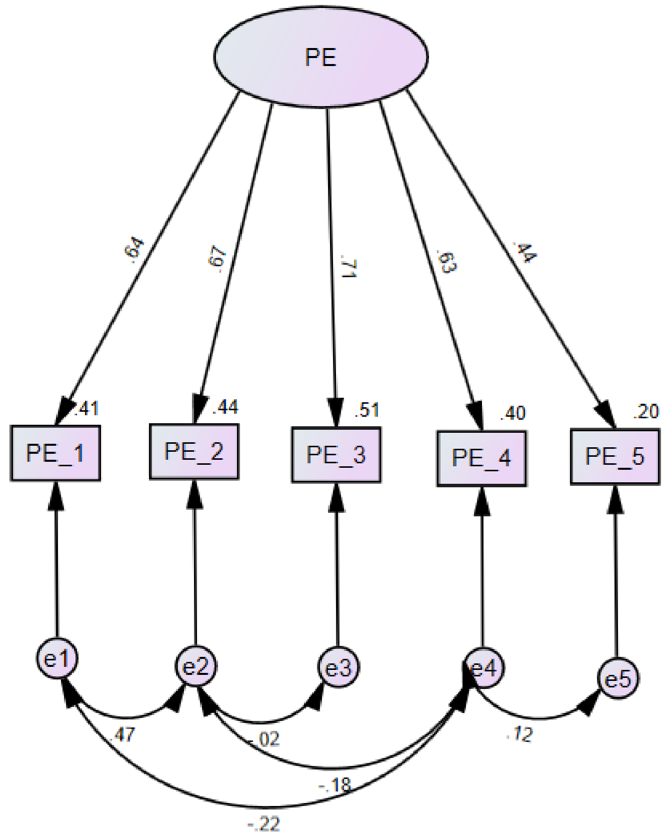
We redesigned the models for PE and SES scale and then compared the model we got by our hypothesis with the one given in the data paper. This was done since the diagrams were unconventional in the data paper as well as there was scarcity of information on first order CFA. Moreover the hypothesis which the authors of our paper used was also not properly defined.

5.1 Assumptions of Data Paper

- All the three scales are independent of each other i.e each of them would be having separate models.
- The variance of each of the scales is 1 and mean is 0.
  The above assumption which was originally given in data paper indirectly hints upon the constraints to be put in the model.
  We have kept degrees of freedom unchanged in every model and also respected the assumptions while redesigning the models.

Every section would be divided in few number of subsections which will clearly explain the plausibility of the hypothesis we framed, conventions as well as the statistics. (Code and data would be attached as hyperlinks towards the end of the paper)

5.2 Parental Engagement



All the values are the standardized estimates. The latent variable PE is seen to load the highest on PE_3 (factor loading 0.71 and variable variance 0.51), followed by PE_2, PE_1, PE_4 and lastly by PE_5.

5.2.1 Convention

The above diagram has been drawn with help of AMOS. The correlation of error in AMOS simply means that the observed variables corresponding to those of error terms are correlated.

5.2.2 The Observed Variables

- PE_1: Follow my own ideas about what my children need to learn

- PE_2: Mix my own ideas with the school's plan on what my children need to study.

- PE_3: I list and prepare the activities myself before developing them with my child(ren).

- PE_4: My children and I have set a homeschooling timetable.
- PE_5 I develop with my children spontaneous learning activities not necessarily school related such as cooking, woodwork, online games, etc.

5.2.3 Our Hypothesis

While designing CFA model it is a prime thing to take care of mathematical fit as well as practical reasoning of the model, as mentioned in the previous sections we impose certain scaling in CFA and while doing that we might end up getting a better model by correlating two terms which might not be correlated in real life and vice-versa.
A ~~ B means that A is assumed to be correlated with B.

- PE_1 ~~ PE_2 If a parent is following his/her ideas in homeschooling then it is fair to assume that he/she will mix his/her ideas in school's plan too.
- PE_1 ~~ PE_4 Similarly while following their own ideas parents need to create a homeschooling time-table hence this two would also be correlated.
- PE_2 ~~ PE_4 By above two arguments one would surely expect this correlation.
- PE_3 ~~ PE_4 Before creating a time-table parents would surely prepare a rough list of activities which would be performed in homeschooling.
- PE_1 ~~ PE_3 As the parents are trying to follow their own ideas it will surely effect the list of activities but as we can see that it is not highly correlated but the reason to keep this correlation was to maintain the degrees of freedom and it also won't affect the accuracy of model much as mathematically and practically this was the best constraint at this level.
- No correlation with PE_5 Notice that PE_5 is about activities other than schooling while others were about the schooling activities so it is expected that it won't be correlated with any of the other PE's.

5.2.4 Statistics

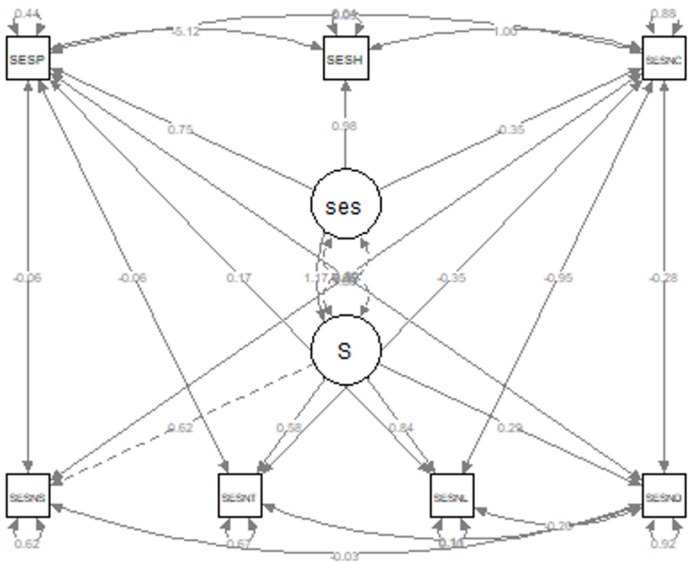
The hypothesis is now practically justified, now the next step is to check the mathematical fit of the model, following the previous sections we present the statistics i.e the fit indices (both absolute and relative) along with comparing them with the fit indices of the model of the data paper.

	Redesigned Model	Original Model
Chi-squared	54.32	508.122
SRMR	0.022	0.056
RMSEA	0.044	0.147
TLI	0.978	0.796
CFI	0.989	0.898

We can clearly infer that the model which we constructed fits the data better.

Conclusion: The model which we designed for the PE scale is mathematically as well as practically a better model.

5.3 Socioeconomic Status



5.3.1 Convention

The above diagram was produced in R using the "**lavaan**" and "**SEMTools**" packages. The convention is as follows:

- The dotted straight line indicates the fixed factor loading.
- The curved dotted line from the variable/factor to itself is the error corresponding to it.
- Other conventions are same as of path diagrams.

5.3.2 Observed Variables

- SESPMO: What do you do in your main job?
- SESHMI: In a normal month, what is your total household income?
- S: is composed of How many usable devices are there in the house? (Smartphones(SESNSP), tablets or iPads(SESNTI), laptops(SESNLA), desktops(SESNDE)).
- SESNCH: How many computers per child have you got at home?

5.3.3 Our Hypothesis

For estimating S we have fixed the factor loading of SESNSP as there is no standard convention of fixing it but it is preferred to fix the factor loading of variable which will affect the latent most. In this case mobile phone will practically affect S most since it is the cheapest as well as a compact device amongst all, which a family can afford even if they cannot afford the latter.

- SESPMO ~~ SESHMI
  The income has to share a lot of dependency on the profession.

- SESPMO ~~ SESNCH

- SESPMO ~~ SESNSP

- SESPMO ~~ SESNTI

- SESPMO ~~ SESNLA

- SESPMO ~~ SESNDE
  For the above five relations: occupation will directly impact amount of electronics in house as people with occupations like business, trade etc. generally needs more electronic items for maybe communicating purpose compared with posts like teacher, researcher, etc.

- SESHMI ~~ SESNCH
  The household income will directly affect the number of PCs in a house.

- SESNCH ~~ SESNDE

- SESNCH ~~ SESNLA
  Above two relation one will directly expect since the number of PCs will depend on number of desktops and laptops in house.

- SESNCH ~~ SESNSP

- SESNDE ~~ SESNSP

- SESNDE ~~ SESNTI
  The above three may have some dependencies but one of the minor reasons to include these is to maintain degree of freedom as that of the original paper.

- SESNDE ~~ SESNLA
  This relation one expects practically and sort of with negative correlation which the model confirms.

- SESNCH ~~ SESNTI
  This we can expect since the tablets especially can take place of computers in home.

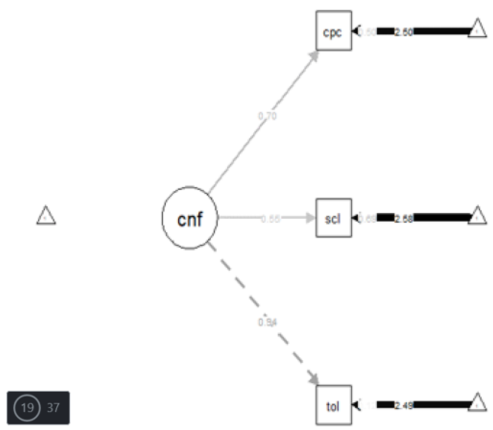
5.3.4 Statistics

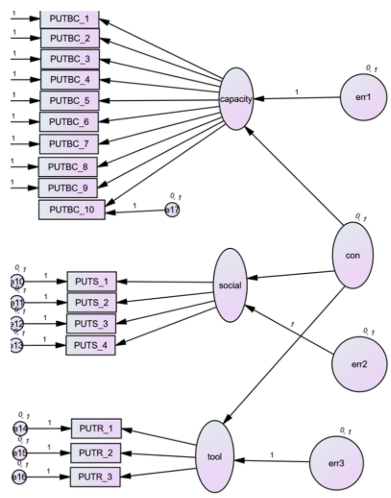
After the practical validity of the hypothesis, we present the fit indicating the mathematical fit.

	Redesigned Model	Original Model
Chi-squared	8.882	19.38
SRMR	0.006	0.015
RMSEA	0.026	0.046
TLI	0.988	0.977
CFI	0.999	0.992

5.4 Parental Acceptance and Confidence in the use of technology

The comparison below shows that the model doesn't give any hints about the first order CFA, moreover the degrees of freedom inside paper is also 0 hence it is not only tedious but also nearly impossible to check all the possibilities and we do not have any basis to compare the models since fit indices will take all the just identified values i.e (Chisq=0,SRMR=RMSEA=0,CFI=TLI=1).





CHAPTER 6

MULTI-GROUP CONFIRMATORY FACTOR ANALYSIS

After performing CFA we get a model which fits the whole data in this case across all countries as a one whole sample. But what about countries as individual samples or regions etc? MG-CFA helps us analyse this question. It shows us whether the sample amongst different groups also follow the same model or not. Elaborating this analogy for our paper we got an accurate model for all 23 countries as whole, now the goal is to check whether the model is also a good fit across the regions namely (Eastern-Asia (EAS), Southern-Asia (SAS), Africa (AFR), Europe (EUR), America (AMR)) and across all the countries individually.

6.1 Types Of Invariance

The term is self explanatory, there are different notions of invariance each imposing some kind of invariance along different groups we combinely look at all of them to make any inference. The four types of invariance which we will look are as follows:

- Configural Invariance (fit.configural)
  Here we fix the model and analyse it independently on every group.

- Metric Invariance (fit.loadings)
  Here we fix the factor loadings from the original model and compare the fit indices.

- Scalar Invariance (fit.intercepts)
  We fix the intercepts as well as the factor loadings and then compare the indices.

- Strict Invariance (fit.residuals)
  We fix the intercepts, factors as well as the variances of residuals across the group.

6.2 Results Of MG-CFA

Hereby we present the results of MG-CFA across the groups as mentioned above and also across the countries across the regions.

6.2.1 PE Scale

Across the 5 groups

```
Measurement invariance models:

Model 1 : fit.configural
Model 2 : fit.loadings
Model 3 : fit.intercepts
Model 4 : fit.residuals
Model 5 : fit.means


Chi-Squared Difference Test

               Df   AIC   BIC   Chisq Chisq diff    RMSEA Df diff Pr(>Chisq)
fit.configural 25 77759 78248  91.993
fit.loadings   41 77768 78153 133.243      41.25 0.039592      16  0.0005099 ***
fit.intercepts 57 77844 78125 241.151     107.91 0.075534      16  1.119e-15 ***
fit.residuals  57 77844 78125 241.151       0.00 0.000000       0
fit.means      61 77904 78158 309.001      67.85 0.125915       4  6.452e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Fit measures:

                 tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural 0.974 0.027 0.987        NA         NA        NA
fit.loadings   0.978 0.039 0.982     0.004      0.013     0.005
fit.intercepts 0.969 0.047 0.964     0.009      0.008     0.018
fit.residuals  0.969 0.047 0.964     0.000      0.000     0.000
fit.means      0.961 0.057 0.952     0.008      0.010     0.012
```

The above indices shows the fit of model across the 5 regions.

EAS

The below image shows the invariance of model across the countries of Eastern Asia (China, Japan).

```
Chi-Squared Difference Test

               Df    AIC    BIC  Chisq Chisq diff   RMSEA Df diff Pr(>Chisq)
fit.configural 10 5792.4 5910.3 30.407
fit.loadings   14 5787.0 5889.1 32.944      2.537 0.00000       4  0.6380403
fit.intercepts 18 5832.1 5918.6 86.092     53.148 0.25565       4  7.936e-11 ***
fit.residuals  18 5832.1 5918.6 86.092      0.000 0.00000       0
fit.means      19 5842.7 5925.2 98.687     12.595 0.24835       1  0.0003867 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Fit measures:

                 tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural 0.886 0.057 0.943        NA         NA        NA
fit.loadings   0.925 0.061 0.947     0.038      0.004     0.004
fit.intercepts 0.789 0.100 0.810     0.135      0.039     0.137
fit.residuals  0.789 0.100 0.810     0.000      0.000     0.000
fit.means      0.766 0.124 0.778     0.023      0.024     0.032
```

SAS

The below image shows the invariance of model across countries of Southern Asia (India, Pakistan, Sri-Lanka).

```
Chi-Squared Difference Test

                Df    AIC    BIC  Chisq Chisq diff   RMSEA Df diff Pr(>Chisq)
fit.configural 15 4592.3 4758.6 37.902
fit.loadings   23 4582.7 4719.5 44.333     6.4305 0.00000       8   0.599129
fit.intercepts 31 4591.1 4698.4 68.767    24.4341 0.14381       8   0.001937 **
fit.residuals  31 4591.1 4698.4 68.767     0.0000 0.00000       0
fit.means      33 4592.1 4691.9 73.686     4.9189 0.12121       2   0.085484 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Fit measures:

                  tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural 0.831 0.078 0.915        NA         NA        NA
fit.loadings   0.897 0.087 0.921     0.066      0.009     0.006
fit.intercepts 0.865 0.114 0.861     0.032      0.027     0.061
fit.residuals  0.865 0.114 0.861     0.000      0.000     0.000
fit.means      0.863 0.123 0.850     0.002      0.009     0.011
```

AFR

The below image shows the invariance of model across countries of Africa (Cameroon, Ethiopia, Ghana, Tanzania).

```
                Df    AIC    BIC   Chisq Chisq diff   RMSEA Df diff Pr(>Chisq)
fit.configural 20 5567.7 5804.3  50.065
fit.loadings   32 5577.8 5767.1  84.187     34.122 0.13912      12  0.0006452 ***
fit.intercepts 44 5601.5 5743.4 131.815     47.628 0.17655      12  3.627e-06 ***
fit.residuals  44 5601.5 5743.4 131.815      0.000 0.00000       0
fit.means      47 5617.1 5747.2 153.441     21.626 0.25531       3  7.804e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Fit measures:

                  tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural 0.846 0.111 0.923        NA         NA        NA
fit.loadings   0.833 0.121 0.866     0.013      0.010     0.057
fit.intercepts 0.795 0.142 0.775     0.037      0.021     0.091
fit.residuals  0.795 0.142 0.775     0.000      0.000     0.000
fit.means      0.768 0.178 0.727     0.028      0.035     0.048
```

AMR

The below image shows the invariance of model across countries of America (Chile, Colombia, Costa Rica, El Salvador, Handuras, Mexico, Peru, Uruguay, USA).

```
Chi-Squared Difference Test

                 Df   AIC   BIC  Chisq Chisq diff    RMSEA Df diff Pr(>Chisq)
fit.configural  45 42887 43688 118.26
fit.loadings    77 42876 43487 170.84     52.577 0.045594      32    0.01239 *
fit.intercepts 109 43056 43477 414.90    244.063 0.146367      32  < 2.2e-16 ***
fit.residuals  109 43056 43477 414.90      0.000 0.000000       0
fit.means      117 43078 43452 453.11     38.206 0.110481       8  6.897e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Fit measures:

                  tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural 0.952 0.039 0.976        NA         NA        NA
fit.loadings   0.964 0.053 0.969     0.012      0.015     0.007
fit.intercepts 0.916 0.075 0.899     0.047      0.022     0.070
fit.residuals  0.916 0.075 0.899     0.000      0.000     0.000
fit.means      0.915 0.085 0.889     0.002      0.009     0.010
```

EUR

Across Europe the estimated matrix is not positive semi-definite which can be handled but with some more computational efforts, but above indices suggests that the model is invariant across 18 countries hence it would be fair to assume the model is invariant across remaining 5 countries.

6.2.2 SES Scale

Similar results for SES Scale

Across the 5 groups

```
Chi-Squared Difference Test

                Df    AIC    BIC    Chisq Chisq diff   RMSEA Df diff Pr(>Chisq)
fit.configural   5 108854 109963   51.273
fit.loadings    29 109100 110052  345.680     294.41 0.10579      24  < 2.2e-16 ***
fit.intercepts  49 110340 111162 1626.169    1280.49 0.25020      20  < 2.2e-16 ***
fit.residuals   65 110546 111263 1863.592     237.42 0.11724      16  < 2.2e-16 ***
fit.means       73 111357 112022 2690.541     826.95 0.31887       8  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Fit measures:

                 tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural 0.833 0.015 0.992        NA         NA        NA
fit.loadings   0.803 0.047 0.946     0.030      0.032     0.046
fit.intercepts 0.420 0.102 0.729     0.383      0.055     0.216
fit.residuals  0.501 0.112 0.691     0.081      0.010     0.038
fit.means      0.353 0.191 0.551     0.148      0.079     0.141
```

EAS

```
Chi-Squared Difference Test

               Df    AIC    BIC   Chisq Chisq diff   RMSEA Df diff Pr(>Chisq)
fit.configural  2 7898.5 8165.8  14.213
fit.loadings    8 7894.6 8138.3  22.291      8.078 0.04292       6   0.232455
fit.intercepts 13 7911.1 8135.1  48.740     26.450 0.15106       5  7.299e-05 ***
fit.residuals  17 7919.9 8128.1  65.547     16.807 0.13050       4   0.002107 **
fit.means      19 8069.5 8270.0 219.216    153.669 0.63512       2  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Fit measures:

                  tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural  0.089 0.053 0.957        NA         NA        NA
fit.loadings    0.733 0.066 0.949     0.645      0.012     0.007
fit.intercepts  0.590 0.085 0.873     0.144      0.019     0.076
fit.residuals   0.574 0.096 0.827     0.016      0.012     0.046
fit.means      -0.573 0.207 0.289     1.147      0.111     0.539
```

SAS

```
Chi-Squared Difference Test

               Df    AIC    BIC    Chisq Chisq diff    RMSEA Df diff Pr(>Chisq)
fit.configural  3 5776.5 6153.6   4.4195
fit.loadings   15 5780.1 6112.8  31.9708     27.551 0.114221      12   0.006431 **
fit.intercepts 25 5822.5 6118.3  94.4273     62.457 0.229801      10  1.239e-09 ***
fit.residuals  33 5827.4 6093.5 115.2525     20.825 0.127040       8   0.007627 **
fit.means      37 5826.3 6077.7 122.1553      6.903 0.085473       4   0.141115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Fit measures:

                 tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural 0.921 0.020 0.996        NA         NA        NA
fit.loadings   0.812 0.081 0.955     0.109      0.061     0.041
fit.intercepts 0.538 0.121 0.817     0.274      0.040     0.138
fit.residuals  0.585 0.141 0.783     0.047      0.020     0.034
fit.means      0.617 0.137 0.775     0.032      0.004     0.008
```

AMR

```
Chi-Squared Difference Test

                Df   AIC   BIC    Chisq Chisq diff   RMSEA Df diff Pr(>Chisq)
fit.configural   9 58500 60315   82.887
fit.loadings    57 58598 60128  276.807     193.92 0.09913      48  < 2.2e-16 ***
fit.intercepts  97 58973 60266  732.233     455.43 0.18323      40  < 2.2e-16 ***
fit.residuals  129 59248 60351 1070.564     338.33 0.17592      32  < 2.2e-16 ***
fit.means      145 59992 61000 1846.780     776.22 0.39192      16  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Fit measures:

                 tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural 0.597 0.034 0.981        NA         NA        NA
fit.loadings   0.811 0.066 0.943     0.213      0.032     0.038
fit.intercepts 0.679 0.093 0.835     0.132      0.028     0.108
fit.residuals  0.642 0.117 0.756     0.037      0.024     0.079
fit.means      0.424 0.202 0.558     0.218      0.084     0.197
```

EUR and AFR

The same problem as well as the convergence issue of configured model and the small sample size across the some countries in Africa and Europe (which also serves as a minor reason for configured model to not converge).

CHAPTER 7

INDEPENDENTLY-COLLECTED DATA

The same questionnaire was recreated and we were able to collect data of around 55 independent samples but only constrained to India.

7.1 CFA on the data collected

Below are the statistics:

PE

Fits	Values
Chi-Squared	9.055
RMSEA	0.121
SRMR	0.076
TLI	0.943
CFI	0.972
Cronbach's Alpha	0.88

SES

Fits	Values
Chi-Squared	1.255
RMSEA	0.000
SRMR	0.026
TLI	1.197
CFI	1.000
Cronbach's Alpha	0.57

The fact that cronbach's alpha is lower for SES since the scale for income was not defined as well as the sample size is extremely small compared to the original data.

7.1.1 Combining The Data

We combined our data with the original data and did the CFA across regions since the countries won't get affected by the only data of India therefore there was no reason to conduct CFA across the countries.

PE

```
Measurement invariance models:

Model 1 : fit.configural
Model 2 : fit.loadings
Model 3 : fit.intercepts
Model 4 : fit.residuals
Model 5 : fit.means


Chi-Squared Difference Test

               Df   AIC   BIC   Chisq Chisq diff    RMSEA Df diff Pr(>Chisq)
fit.configural 25 78653 79143  93.855
fit.loadings   41 78655 79040 127.585     33.730 0.032996      16   0.005907 **
fit.intercepts 57 78700 78981 204.989     77.404 0.061406      16  4.869e-10 ***
fit.residuals  57 78700 78981 204.989      0.000 0.000000       0
fit.means      61 78756 79011 268.904     63.915 0.121313       4  4.355e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Fit measures:

                 tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural 0.974 0.026 0.987        NA         NA        NA
fit.loadings   0.980 0.037 0.984     0.006      0.011     0.003
fit.intercepts 0.975 0.044 0.972     0.005      0.006     0.012
fit.residuals  0.975 0.044 0.972     0.000      0.000     0.000
fit.means      0.968 0.054 0.961     0.008      0.010     0.011
```

SES

```
Measurement invariance models:

Model 1 : fit.configural
Model 2 : fit.loadings
Model 3 : fit.intercepts
Model 4 : fit.residuals
Model 5 : fit.means


Chi-Squared Difference Test

               Df    AIC    BIC    Chisq Chisq diff   RMSEA Df diff Pr(>Chisq)
fit.configural  5 110039 111150   51.524
fit.loadings   29 110293 111247  353.521     302.00 0.10668      24  < 2.2e-16 ***
fit.intercepts 49 111526 112350 1626.624    1273.10 0.24811      20  < 2.2e-16 ***
fit.residuals  65 111744 112463 1876.668     250.04 0.11988      16  < 2.2e-16 ***
fit.means      73 112549 113215 2696.852     820.18 0.31583       8  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Fit measures:

                 tli  srmr   cfi tli.delta srmr.delta cfi.delta
fit.configural 0.834 0.015 0.992        NA         NA        NA
fit.loadings   0.800 0.048 0.945     0.034      0.034     0.047
fit.intercepts 0.425 0.102 0.732     0.375      0.054     0.213
fit.residuals  0.502 0.114 0.692     0.077      0.011     0.040
fit.means      0.358 0.186 0.554     0.144      0.073     0.138
```

It is fascinating to observe that the model of data with higher internal consistency i.e higher cronbach's alpha is more consistent across the groups while one with lower internal consistency i.e lower cronbach's alpha is less consistent across the groups.

CHAPTER 8

CODE, DATA AND BIBLIOGRAPHY

8.1 Code and Data

- The original data with slight modification can be found **here (d.xlsx)**

- The independently collected data can be found **here (PES.xlsx)**

- The combined data can be found **here (e.xlsx)**

- The Region wise data (G1-G5) can be found **here**

- Our code for the models (R file) can be found **here**

8.2 Bibliography

We have taken help from few websites for this project. You can click on the following links to access the websites.

- **bookdown.org**
- **Research Gate**
- **Wikipedia (CFA)**
- **Wikipedia (SEM)**
- **Towards Data science**
- **https://stats.oarc.ucla.edu**
- **https://stats.oarc.ucla.edu/spss/**

- **Youtube (1)**
- **Youtube (2)**
- **Github**
- **Cronbach's alpha**

Books that have helped us understand concepts in thorough detail:

Confirmatory Factor Analysis for Applied Research - Second Edition - Timothy A. Brown