Negative Item Response Bias in Education-Based Surveys - a Factor Modelling Approach

ALEXANDER O'RIORDAN

Stellenbosch Economic Working Papers: WP04/2021

www.ekon.sun.ac.za/wpapers/2021/wp042021

March 2021

KEYWORDS: Latent Construct Estimation, Negatively Item Response, Confirmatory Factor Analysis, Hierarchical Cluster Analysis
JEL: A21, C81, C83, I21, O12

ReSEP (Research on Socio-Economic Policy)
https://resep.sun.ac.za

DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH

SOUTH AFRICA

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE
BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

www.ekon.sun.ac.za/wpapers

Negative Item Response Bias in Education-Based Surveys - a Factor Modelling Approach

Alexander C. O'Riordan[a]

[a] *Department of Economics, Stellenbosch University*

Abstract

In applied survey-based research, it is common to encounter responses based on both positively and negatively worded questions. In practice, responses are typically recoded to ensure that the numerical values attached to the responses of positively and negatively worded questions are aligned. This is done under the assumption that the responses to negatively worded questions are perfectly reversed reflections of responses to identical or similar positively worded questions - that the variation is inversed. This paper tests this assumption within a framework of factor modelling using South African Grade 4 TIMSS data. It finds significant differences in the degree to which the different question orientations capture information about single latent constructs that a specific group of questions is designed to capture. And thus, a failure of the assumption.

Keywords: Latent Construct Estimation, Negatively Item Response, Confirmatory Factor Analysis, Hierarchical Cluster Analysis

JEL classification A21, C81, C83, I21, O12

1. Introduction

In survey-based research, data derives directly from individual responses to the items included in a specifically and intentionally designed questionnaire. While it is generally well understood that the sample of respondents needs to adhere to several requirements, such as random selection and broad representation of the population, the same is not true to the same extent for the specified format and orientation of questionnaire response items. This paper seeks to highlight the importance of the design and orientation (specifically, positive or negative wording) of survey items, and how these questionnaire item characteristics can affect the ultimate outcomes of empirical research. Whether used alone as an explanatory variable in regression, or as an input into a factor modelling procedure, it is vital that the information contained within each derived feature is reliable and accurately reflects the intended component of the specific data generating process as per its design.

It is common for questionnaire construction specialists to include negatively worded items in surveys. Moreover, it is also common to encounter several items, possibly of different design and orientation, that are intended to capture the same underlying process - many questions broadly for one piece

Email address: alexoriordan@sun.ac.za (Alexander C. O'Riordan)

of information. This is done, in part, to disrupt response sets and thus maintain active response engagement by the respondent (Marsh, 1986). The premise is that including items that are slightly dissimilar to those preceding it will induce the respondent to pause and think, and thus provide more accurate and holistic information. However, this is done under the assumption that responses to items of either orientation are perfectly aligned with one another - the observable variation equally well reflects underlying processes.[1] The broad purpose of this paper is to test this assumption and uncover a potential source of measurement error - differing question orientations.

The factor modelling approach is ideal for this purpose as its implicit intention is to model unobservable phenomena that determine observable outcomes - the unobservable processes that underlie observable responses. The basic premise of factor estimation is to use a group of informationally homogenous observable features to estimate a specified number of factors, typically a single factor, as is the case in this paper. Importantly, these procedures provide information on how well each individual feature, within the specified group of informationally homogenous features, fits the single specified factor. Therefore, we are able to quantify the relative strength with which each feature reveals information about the single specified factor. Moreover, this approach differs from contemporaneously common tests for sampling adequacy in that it emphasises group structure in relation to an underlying factor, rather than only the interrelationships within and among the group's individual features.[2]

The analysis employs Confirmatory Factor Analysis (CFA) to test the factor structure of two example latent constructs - self-efficacy and the enjoyment of mathematics. Two samples of response item groups are drawn from the South African 2016 Trends in International Mathematics and Science Study (TIMSS) grade 8 dataset, with each corresponding to one of the two studied latent constructs. CFA is a useful method in that it allows the researcher to impose a factor structure upon an identified latent construct and then test the adequacy of the hypothesised structure (Brown, 2015). That is, identification of the latent construct is discretionary. Methods such as Explanatory Factor Analysis and Principal Component Analysis, while very similar, are used primarily to uncover factor structure rather than test the adequacy of imposed structure. They are therefore less useful in this context. In addition to CFA, a thorough hierarchical clustering exercise is applied to the group of features associated with each latent construct. This is valuable in addition to CFA in that cluster analysis allows the data to speak for itself without the imposition of hypothesised structure or *a priori* design or consideration for underlying latent constructs. It is merely a statistical grouping exercise, one that can demonstrate which features share similar characteristics and which do not. Moreover, traditional measures of sampling adequacy such as Cronbach's Alpha are used. The idea supporting this addition is to investigate whether or not tests that are typically applied in practice will also reveal possible

[1] given that the responses to negatively worded items have appropriately been adjusted and recoded to ensure that the scale ordering is comparable.

[2] Sampling adequacy testing is the broad term given to procedures used to test the degree of homogeneity among a group of features (what we are essentially doing in this paper). It is a common step in procedures of factor estimation and modelling (Cerney & Kaiser, 1977). In this paper, we compare the results of the Confirmatory Factor Analysis to more typical methods of sampling adequacy in order to uncover any possible differences in the results. As traditional tests of sampling adequacy are generally relied upon in empirical work, any differences in the results would be a concerning finding.

informational differences based on question orientation.

The paper is structured as follows. First, the theoretical and practical effects of the inclusion of negatively worded items are discussed. Second, the data used is described. Third, the employed methodology is explained. Fourth, the empirical results are provided and discussed. Finally, conclusions are provided.

2. Theoretical Framework: The use of Negatively Worded Response Items

For the use of features based on negatively worded items (referred to as negative features hereafter) to be justified, their inclusion must first, introduce minimal additional noise in the form of measurement error into the data, and second, ideally provide beneficial effects in the form of response-set disruption and more accurate variation. That is, they must at a minimum have no negative effect and ideally have some positive effect. Using a factor estimation approach, it is possible to empirically investigate whether or not an individual feature is accurately measuring the underlying construct that it is designed to measure. Moreover, we can also use this approach to compare the extent to which individuals in a group of features reflect a single underlying construct. There are two main phenomena by which negative features can have harmful noise inducing effects, specifically when using responses provided by young adolescent students or those with underdeveloped literacy skills (Chiavaroli, 2017).[3] Each is discussed in turn.[4]

First, for positively and negatively worded items to be measuring the same underlying opinions and processes as positively worded items, respondents need to be capable of employing double negative logic. The mental processing required to understand and apply the logic of the English double negative is complex (Hunt, 1978). Such logic requires a relatively high level of verbal reasoning, which may not yet have developed for young students. For example, if a given student does enjoy mathematics, the question "I do *not* enjoy mathematics" requires a response of "I do *not* agree" while the question "I enjoy mathematics" requires a response of "I agree". If the respondent is incapable of applying the appropriate logic, their given responses may capture this inability in the form of measurement error that is essentially impossible to identify by analysing the response in isolation. Therefore, the responses to negatively worded items may be noisy reversed scores of the responses given to similar positively worded items. This noise results from the confusion and uncertainty experienced when answering a negatively worded item and the inappropriate or erroneous application of double negative logic.

The potential for noise, and its distribution within the sample, created by this limited verbal reasoning is exacerbated in the South African context, one in which the distribution of reading ability is extremely unequal (Spaull, 2013). Students in poorer schools generally receive a low-quality education and

[3] In the South African context, in many cases there exists a disconnect between student age and literacy skills. Many are several years behind the level of development typically associated with their age (Spaull & Hoadley, 2017)

[4] See Wong & Rindfleisch (2003); Weijters & Baumgartner (2013); and Hartly (2014).

will have more difficulty applying English double negative logic to appropriately answer a negatively worded item than would their wealthier peers. While not all students in the sample are tested in English, it is assumed that the double negative logic required in other testing languages is similarly distinct from that used when interpreting positively worded text. Therefore, if it is true that negatively worded items do induce noisy responses, their inclusion in questionnaires could induce a systematic source of bias into the data in the South African case.

The second phenomenon is relatively simple, It might be the case that respondents are simply ignorant to the nuance of negatively worded items and not notice their distinction (MacDonald, 2013). In this case, response scores will be perfectly reversed versions (once recoded) of the response scores for positively worded items. If this is true, the proposed positive effect of negatively worded items will be eliminated. However, and more importantly, if these responses are *exactly* opposite to those intended by the orientation of the item, as seems plausible, the recoding (reversing of the scores) of the relevant features should completely remove this inaccurate measurement. Conversely, if the aforementioned first phenomenon does occur, it is not necessarily the case that simply recoding the data will solve the problem. Moreover, the second phenomenon will not cause systematic bias as will the first. Therefore, of these two phenomena, the first is more concerning and is the central point of interest of this paper.

3. Data

3.1. Description of Data

The South African 2016 Trends in International Mathematics and Science Study (TIMSS) grade 8 dataset is used for the empirical analysis (N = 10370). Specific features are selected to estimate two factors - one, a measure of the enjoyment of mathematics, the other, a measure of positive self-efficacy toward mathematics. TIMSS is preferred for this paper due to its extensive and available student questionnaire information. Moreover, the grade 8 dataset is preferred to the grade 4 dataset as the slightly older students should be better suited for the purpose of this study. The two specific studied factors are chosen because each has nine features that can be used in their identification. Among the nine features used to estimate the enjoyment factor, two are negative. Of the nine features used to estimate the efficacy factor, five are negative.

The definition of a negatively worded item used here is relatively simple. These are items that, for a given response to have the same intended meaning, must be answered with an opposite scoring than would be used when answering a similar positively worded item. For example, if a student thoroughly enjoys mathematics, she would answer the following two example questions - 1) I do enjoy mathematics, 2) I do *not* enjoy mathematics - as follows. Question 1 would be answered with a response of "I agree", while question 2 would be answered with a response of "I do *not* agree". Importantly, on the Likert type scale used in the TIMSS questionnaire, these responses would be on opposite ends of the response scale. Question 2 is the negatively worded item and question 1 is the positively worded item.

Here it is again valuable to note an important distinction. The main premise of this paper is not to

investigate whether the inclusion of negative features will negatively affect the estimation of factors. It obviously will, the estimation procedures used are not able to distinguish between the responses to items of different orientation and design. They see only the responses to items in numerical terms and are based broadly on the correlation structure that exists among a specified group of features. Therefore, negative features must be recoded to reverse the numerical order of the values attached to the qualitative responses. Importantly, this paper does not investigate whether ignorance of this necessary recoding of features will affect factor estimation. Rather, it investigates the effect of appropriately recoded negative features.

3.1.1. Latent Constructs - Enjoyment of Mathematics and Self-Efficacy

Each of the two factors is estimated using nine features, each of which is based on a response item that pertains to the self-reported degree of either the enjoyment of mathematics (first factor), or a measure of mathematical self-efficacy (second factor). The items are answered along a Likert-type scale that ranges from 1 to 4 as follows; 1 - Agree a lot, 2 - Agree a little, 3 - Disagree a little, 4 - Disagree a lot. Table 3.1 lists the nine items from which the features used to estimate the enjoyment factor are derived. It also provides the code of the feature used in the analysis that follows. A code with a "p" indicates a positive feature while an "n" indicates a negative feature. Table 3.2 lists the nine items from which the features used to estimate the efficacy factor are derived.

Code	Item
p1	I enjoy learning mathematics
n1	I wish I did not have to study mathematics
n2	Mathematics is boring
p2	I learn many interesting things in mathematics
p3	I like mathematics
p4	I like any schoolwork that involves numbers
p5	I like to solve mathematics problems
p6	I look forward to mathematics class
p7	Mathematics is one of my favorite subjects

Table 3.1: Enjoyment Factor - Features and Codes

Code	Item
p1	I usually do well in mathematics
n1	Mathematics is more difficult for me than for many of my classmates
n2	Mathematics is not one of my strengths
p2	I learn things quickly in mathematics
n3	Mathematics makes me nervous
p3	I am good at working out difficult mathematics problems
p4	My teacher tells me I am good at mathematics
n4	Mathematics is harder for me than any other subject
n5	Mathematics makes me confused

Table 3.2: Efficacy Factor - Features and Codes

3.2. Correlation Analysis

Figure 3.1 displays correlation heatmaps for the nine features included in the two factor estimation procedures. The top two plots are for the efficacy factor, the lower two, the enjoyment factor. The figures to the left show the heatmaps for features that have not been recoded while the figures to the right include negative features that have been recoded. Positive features are noted with a "p", negative features are noted with an "n". A darker blue shade indicates a stronger positive correlation while a darker red shade indicates a stronger negative correlation.

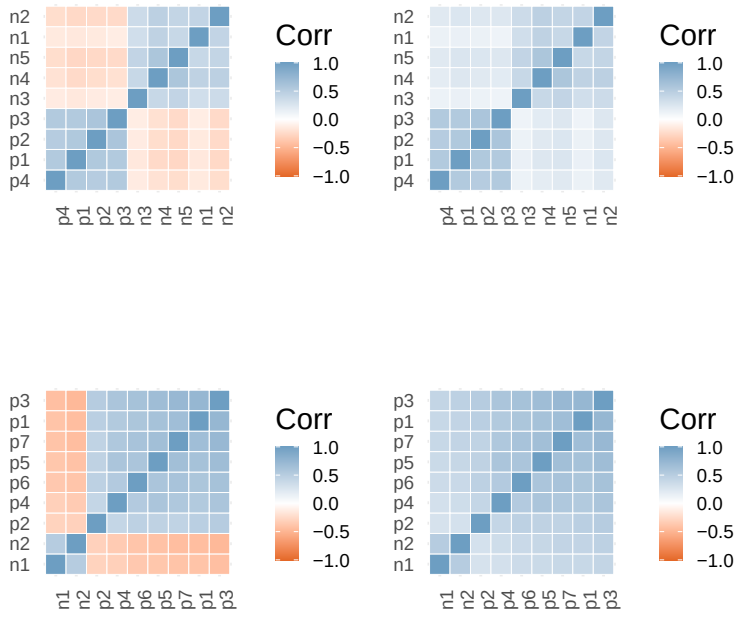

Figure 3.1: Correlation Heatmap - Efficacy Construct (Top) and Enjoyment Construct (Bottom)

From the top two plots of figure 3.1, it is evident that the negative features are negatively correlated with the positive features before being recoded. This is a simple representation of the aforementioned assertion that responses to similar items of differing orientation will be on opposite ends of the response scale. Viewed in isolation, this does reveal that most students in this sample are aware, to a certain extent, of the negative orientation of these items, thus indicating that the second phenomenon of ignorance to negatively worded items is not pervasive. A more interesting insight is revealed by the figure to the top right, which plots the correlations including the recoded negative features. It is evident that, even once recoded, the negative features remain clustered together and only weakly correlated with the positive features while being relatively strongly correlated with only one another.[5]

[5]The ordering of the features on the heatmaps in figures 3.1 is determined by a clustering algorithm. Therefore, features

Therefore, both the magnitude of correlation and the groupings made by the clustering algorithm indicate that these features are characteristically different to the others. That is, responses to items of differing orientation appear to be imperfectly aligned.

From the bottom two plots of figure 3.1, it is evident that the negative features are clustered together and negatively correlated with the positive features. Again, the more interesting finding is that once recoded, the negative features remain clustered together. Moreover, they are correlated strongly only with one another. If all nine features, associated with either factor, are truly capturing the same latent construct, there should be no noticeable difference in the degree of correlation or the grouping by the cluster analysis.

This initial look at the data reveals that the recoding of negative features may not be sufficient. The features with different orientations do appear to be consistently different from one another. In the following section, a more extensive clustering exercise is applied to the data to better understand the group structures that exist. Importantly, clustering can provide information about group structure and which features are similar to one another along certain dimensions. It does not at all provide information about factor structure and the relative ability of a group of features to reflect a specific latent construct or estimate a specific factor.

3.3. Hierarchical Cluster Analysis

Clustering is a statistical partitioning technique that groups a set of features based on their characteristics (Maimon & Rokach, 2010). In this case, correlation clustering is used, the correlation coefficient is the relevant measure of similarity between variables. Broadly, hierarchical clustering differs from other methods such as k-means clustering in that the number of final clusters is not specified *a priori*. Hierarchical clustering can be subdivided into two broad categories, agglomerative nesting (AGNES) and divisive analysis (DIANA). AGNES is a bottom-up approach in which each feature initially represents a distinct cluster. These individual clusters are iteratively merged into larger clusters until only a single large cluster exists and the full hierarchical structure is obtained. DIANA is a top-down approach that works in the exact opposite manner. The procedure starts with a single cluster which encompasses all of the features. This single cluster is iteratively sub-divided until each feature is its own distinct cluster (Maimon & Rokach, 2010).

The technique used here is a complete linkage agglomerative nesting hierarchical cluster algorithm based on Euclidean distances within the correlation matrix. Complete linkage is based on maximum inter-cluster dissimilarity, that is, the similarity of two clusters is the similarity of their two most dissimilar members. In simpler terms, in an agglomerative approach in which the algorithm starts with N clusters, the first two clusters to merge will be the two that are most similar, the two that are the closest by euclidean distance. There are now *N-1* clusters in total with one cluster containing two features. Let's assume that the next iteration results in the two features closest to the cluster

that demonstrate similar characteristics are positioned near to one another on the heatmap.

formed in the first iteration to merge into one. There are now two clusters that each contain two features and *N-4* clusters that contain only one feature. The concept of complete linkage is that the distance between these two clusters each with two features is a measure of the distance between the two features that are the furthest apart from one another within the two respective clusters. Therefore, the next iteration of cluster merging is determined by the relative closeness in Euclidean distance of the two most dissimilar features within each cluster. This iterative process will continue until one all-encompassing cluster exists.

3.3.1. Enjoyment Latent Construct

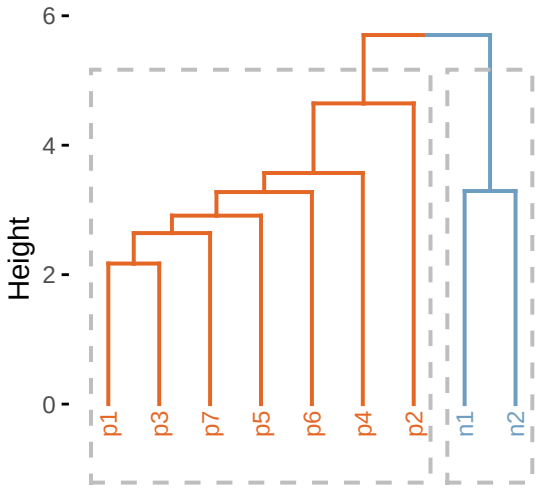

Figure 3.2: Enjoyment Latent Construct - Dendrogram

Figure 3.3 displays the dendrogram for the results of the cluster analysis of the nine features used to estimate the enjoyment factor. It is evident that the two negative features are grouped separately from the seven positive features. Boxes (partitions) are drawn around the dendrogram at a cutoff specified at two clusters. The cutoff need not be two, however, this serves to demonstrate the main partition in the data which separates the features associated with positively and negatively worded response items. This box partition reveals how the data would have been clustered if two final clusters were explicitly specified.

It is worth referring back to the procedure of agglomerative nesting, which is bottom-up. In creating partitions in the data, the clustering algorithm kept the features associated with positively and negatively worded items separate from one another, joining them only at the final iteration in which the final two clusters are forced to merge into one. Therefore, it is along the positive-negative orientation that the features are most dissimilar. Another meaningful insight can be gained by comparing the heights at which clusters are iteratively joined. The height measure indicates the relative distance

between the clusters on the two branches joined at that particular node. It is evident that the two negative features join one another only after p1, p3, p7, p5, and p6 have merged into one cluster. This indicates that the two negative features are more dissimilar to one another than are the first five most similar positive features. Therefore, not only are negative features dissimilar to the positive features, they are also relatively dissimilar to one another.

This significant separation created by the clustering algorithm is demonstrated more extremely by the cluster plot in figure 3.4. From Figure 3.4 it is evident first that there exist two main clusters, one comprised entirely of positive features, the other, entirely of negative features.[6] More importantly, these two clusters are far apart, and in clustering terms, highly dissimilar. While cluster 1 appears upon initial inspection to be large and spread out, with the exception of p2, its points are relatively close to one another along the x-axis, the dimension upon which the majority of the information exists. It is only feature p2 that shows significant dissimilarity to the other positive features. Referring back to Table 3.1, it is evident that this particular positive item is slightly dissimilar to the others. Both clusters are compact and relatively far from one another. Features derived from items with the same orientation are similar to one another while being dissimilar to features derived from items with the opposite orientation. Therefore, these features could be considered to have different characteristics.

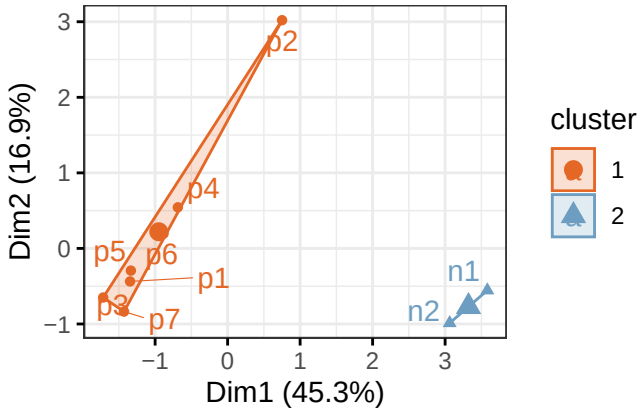

Figure 3.3: Enjoyment Latent Construct - Cluster Plot

3.3.2. Efficacy Latent Construct

The efficacy construct differs from the enjoyment construct in that it incorporates a larger number of negative features. Therefore, a cluster analysis applied to the features used for its identification can provide additional results that are not possible with features of the enjoyment construct. Figure 3.5 displays the results of the clustering algorithm in the form of four dendrograms. While the

[6] Note here that two clusters are explicitly specified. Therefore, it is not the number of clusters that is of interest, rather, it is the composition of each cluster that is important.

dendrograms themselves are identical, each of the four differs in the specified number of box partitions. There are specifications of 2, 3, 4, and 5 box partitions. Again, these box partitions reveal what the individual components of n clusters would be if n final clusters were to be specified.

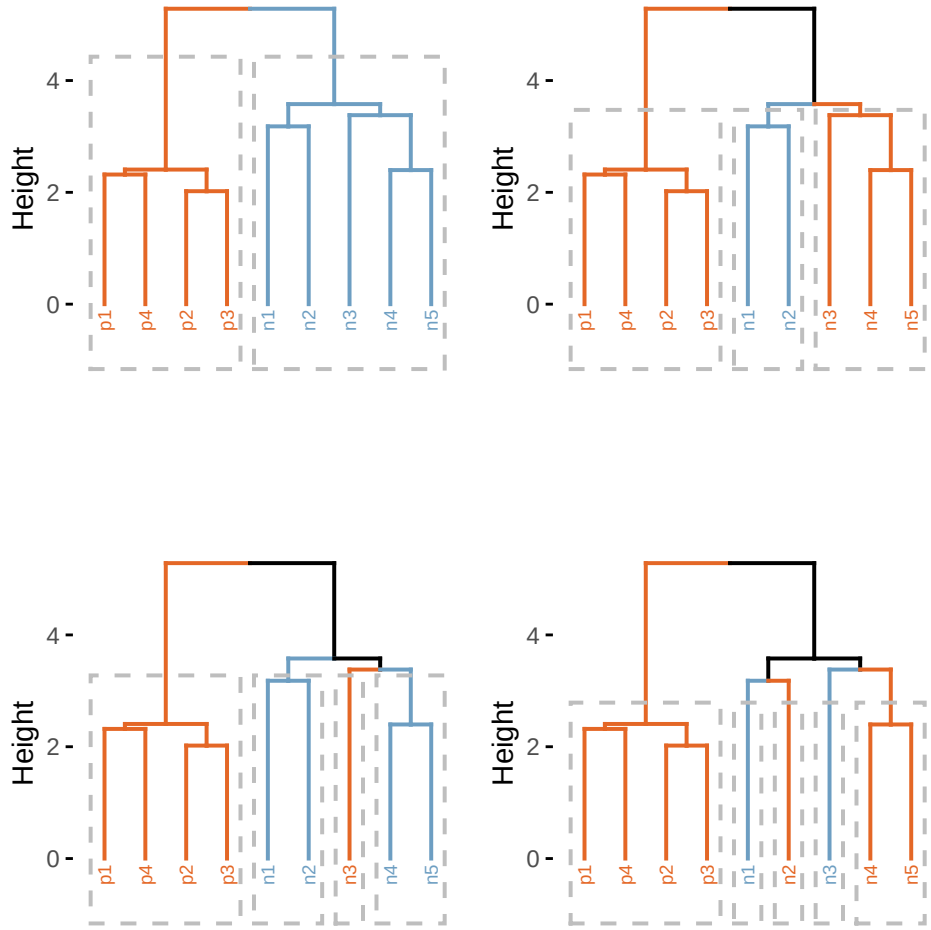

Figure 3.4: Efficacy Latent Construct - Dendrogram

From the dendrograms in figure 3.5, it is evident that the positive features remain in a single cluster, regardless of the box partition specification. This indicates that strong similarity exists between these features. Conversely, the negative features are split into separate clusters at each successively higher box partition specification. Therefore, within the initial single cluster that contains only negative features, the individual features are relatively dissimilar. This finding is corroborated by the height at which the four positive features merge into a single cluster. It is evident that the four positive features merge into a single cluster before even the first cluster containing two negative features is merged. What this indicates is that the two most dissimilar positive features are more similar to one another

than are the two most similar negative features. This finding is in itself interesting. It indicates that not only are negative features dissimilar to positive features, they are also relatively dissimilar to one another. Therefore, the information provided by the group of negative features appears to be internally inconsistent. The results shown in Figure 3.6, which have the same interpretation as those in Figure 3.4, further corroborate the interpretation that negative features are dissimilar to positive features, and are also relatively dissimilar to one another.

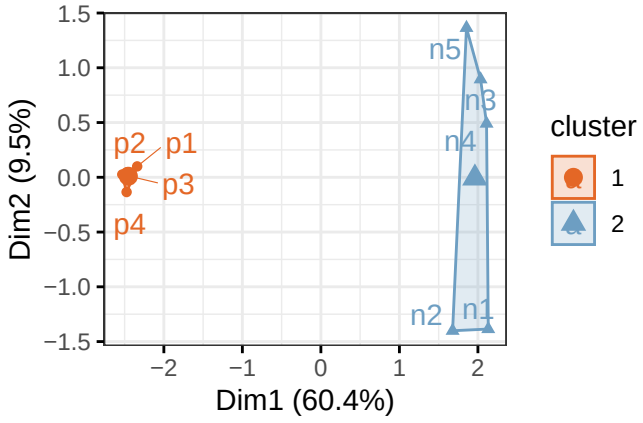

Figure 3.5: Efficacy Latent Construct - Cluster Plot

4. Methodology

4.1. Confirmatory Factor Analysis (CFA)

CFA is a statistical technique used to test and verify a proposed factor structure among a group of features (Suhr, 2006). It achieves this by estimating the common sources of covariance among N features (limited to $N = 1$ in this paper). The implicit assumption underlying the use of CFA is that the observable features are determined by a number of unobservable processes. Simply, observable features are outcomes of unobservable latent constructs. For example, a student will answer a question of mathematical enjoyment positively if she does enjoy mathematics. In this way, the positive response that we observe is an outcome of her innate unobservable enjoyment of mathematics.

In contrast to Principal Component Analysis, which explains maximum total variance, CFA explains maximum shared variance (covariance) among a set of features (Babyak & Green, 2010). Moreover, CFA makes a distinction between variance that is common to all N features (shared variance), and that that is unique to each feature (idiosyncratic variance). Therefore, CFA is a correlation-focused approach in which factors represent the common variance of N features, and the variance not explained is defined as feature-specific (idiosyncratic) variation.

CFA allows one to test for the existence of an *a priori* specified relationship among a set of features

and their proposed common underlying latent construct. It does this by assigning factor loadings to each employed observable feature. These loadings are measures of the degree to which individual features are determined by estimated factors - how much does each unobservable construct influence each observable feature. Therefore, if feature y loads highly onto factor x, we can infer that the observable response to the item from which feature y is derived is largely determined by the latent construct underlying, and represented by, the estimated factor x. Moreover, if features y, w and b all load highly onto factor x, we can infer that these three features share a common underlying latent construct - they are determined by the same underlying process or phenomenon. As indicated, in this paper, we investigate whether or not the observable features derived from positively and negatively worded items are equally determined by the same underlying unobservable latent construct.

The empirical procedure underlying CFA is based primarily on the Common Factor Model (Thurstone, 1947). The Common Factor Model is premised on the notion that there exist two types of latent constructs that influence observed item responses, and their derived features - shared and idiosyncratic (MacCallum, 2009). The CFA procedure models features as a linear function of shared and idiosyncratic influence (variance) by a specified number of factors. The following outlines the fundamental equation of CFA as originally proposed by Joreskog (1967).

$$\mathbf{y} = \mathbf{\Lambda x} + \mathbf{z} \tag{1}$$

The $\mathbf{y}$ vector contains the N observable features. $\mathbf{x}$ is a vector of m common factor scores, the single unobserved latent construct in this case ($m = 1$). $\mathbf{z}$ is a vector of N unique scores - idiosyncratic variance. $\mathbf{\Lambda}$ is an $N \times m$ matrix containing the factor loadings for each feature. From the above, it is evident that CFA decomposes each feature y into variance that is shared, and variance that is unique. Therefore, the results of CFA reveal the degree to which each of the N observed features is influenced by the unobserved common factor. The above fundamental equation depends on three critical assumptions.

$$E(x) = E(z) = 0 \tag{2}$$

$$E(xx') = \mathbf{\Phi} \tag{3}$$

$$E(zz') = \mathbf{\Psi} \tag{4}$$

The dispertion matrix of y is defined as

$$E(yy') = \mathbf{\Sigma\Lambda\Phi\Lambda'} + \mathbf{\Psi} \tag{5}$$

The equations listed above succinctly describe the fundamental premise of CFA. The dispersion matrix of y demonstrates the diagonalization procedure used to obtain the factor loadings, $\mathbf{\Lambda}$. The estimation of CFA is performed by maximum likelihood, with the maximization of the following likelihood function

$$F_{ml} = ln|\mathbf{\Sigma}| + tr(\mathbf{S\Sigma}^{-1}) - ln|\mathbf{S}| - m \tag{6}$$

Where matrix $\mathbf{S}$ contains estimates of variances and covariances of the components of y.

5. Empirical Results

In this paper, CFA is used to test hypotheses about factor structure using different model specifications based on feature orientation. Therefore, it is useful to think in terms of restricted and unrestricted models. In this case, an unrestricted model is one that contains all of the features associated with each factor, including the recoded features based on negatively worded items. The restricted models are those that contain a limited number of features of a single orientation. In total, five CFA models are estimated. Two models are based on the enjoyment factor and three models are based on the efficacy factor. The two models fit to the enjoyment factor are: one unrestricted model that uses all nine features, and one restricted model that uses only the seven positive features. Of the three models fit to the efficacy factor, one is unrestricted and two are restricted. Of the two restricted models, one uses only positive features while the other uses only negative features. Again, the unrestricted model uses all of the features. The performance of the individual features is investigated using measures of fit and the factor loadings. Appendix B contains information about the employed measures of fit.[7]

5.1. Estimated Factor - Enjoyment of Mathematics

Table 5.1 displays the measures of fit for the one restricted and one unrestricted CFA model fit to the features of the enjoyment factor. It is evident that the restricted model outperforms the unrestricted model across every measure of fit. This finding indicates that the model which includes only positive features outperforms the model that includes features of both orientations, with regard to estimating the factor of mathematical enjoyment. In line with the findings of earlier sections of this paper, this result indicates that the inclusion of negative features can be detrimental to factor estimation and that their inclusion in construct identification should be done with caution. That is, the two distinct feature orientations do not equally well reflect the underlying construct of mathematical enjoyment.

Fit_Measure	Unrestricted	Resctricted
Comparative Fit Index	0.962	0.988
Tucker-Lewis Index	0.949	0.983
Loglikelihood user model	-106906.5	-78618.96
Akaike Information Criterion	213849.01	157265.92
Bayesian Information Criterion	213922.25	157322.89
Root-Mean square error	0.080	0.057
Standardized Root Mean Square Residual	0.040	0.017
model chi-square	1838.670	489.535

Table 5.1: Enjoyment CFA - Measures of Fit

[7] In addition to the CFA results, Appendix A includes the results for several traditional measures of sampling adequacy - Chronbach's Alpha (both raw and standardized), Guttman's Lambda, and the Kaiser-Meyer-Olkin measure (Guttman, 1945; Cronbach, 1951; Kaiser & Rice, 1974). From the results of these tests, presented in tables 8.1 and 8.2, it is evident that they reveal no distinction between positive and negative features - that they are informationally identical. This is true for both constructs. Moreover, the overall measures, those that are used to test the sampling adequacy of a group of features as a whole, indicate that each test measure is well within the acceptable range. This is highly concerning as, in practice, a researcher would generally apply these tests and base their subsequent actions on their results alone. However, these results do ensure that the findings of this paper take on increased significance.

Table 5.2 displays the factor loadings of the features onto the enjoyment factor. Results for both the unrestricted and restricted models are shown. As is common, the loading of the first feature is scaled to 1. All loading scores are statistically significant though this information is not shown. From Table 5.2 it is evident that the two negative features have a similar and relatively weak loading onto the enjoyment factor. This indicates that these two features are determined to a lesser degree by the underlying construct of mathematical enjoyment than are the positive features. Interestingly, the feature coded as p2, which is based on the "I learn many interesting things in mathematics" item has the lowest loading. Recall this particular feature was also the most dissimilar positive feature according to the cluster analysis. This is likely due to the slightly different nature and design of the item which does not relate as directly as do the others to the actual enjoyment of mathematics. The findings shown in Table 5.2 indicate that the negative features do not load onto the enjoyment factor as well as the other features do, with the exception of feature p2. Therefore, it appears that there exists a weaker relationship between the negative features and the underlying latent construct of mathematical enjoyment when compared to the positive features. Again, the two distinct feature orientations do not equally well reflect the underlying construct of mathematical enjoyment.

Factor_Loadings	Unrestricted	Resctricted
p1	1.000	1.000
n1	0.839	-
n2	0.830	-
p2	0.743	0.742
p3	1.218	1.219
p4	1.004	1.009
p5	1.150	1.153
p6	1.084	1.083
p7	1.335	1.335

Table 5.2: Enjoyment CFA - Factor Loadings

5.2. Estimated Factor - Self-Efficacy

Table 5.3 shows the measures of fit for the one unrestricted and two restricted CFA models fit to the efficacy factor. The interpretation of the results is the same for those shown in table 5.1. The two restricted models are, in this case, one that contains only positive features and one that contains only negative features. In this case, the unrestricted model performs significantly worse than the two restricted models. An interesting characteristic of this model to consider is the relatively equal share of positive and negative features. The poor performance of the unrestricted model seems to corroborate the story that positive and negative features do not equally reflect the same underlying latent factor, noting again that the model is specified with a single factor. Therefore, the poor performance is possibly the result of forcing the model to estimate a single factor when the features used in the model are determined by two distinct underlying constructs captured by items of different orientation. This narrative is supported by the superior measures of fit for the two restricted models.

Fit.Measure	Unrestricted	Resctricted_Pos	Resctricted_Neg
Comparative Fit Index	0.655	0.995	0.978
Tucker-Lewis Index	0.540	0.984	0.956
Loglikelihood user model	-125701.29	-49829.69	-71689.85
Akaike Information Criterion	251438.58	99675.39	143399.7
Bayesian Information Criterion	251511.81	99707.95	143440.39
Root-Mean square error	0.191	0.062	0.076
Standardized Root Mean Square Residual	0.140	0.013	0.025
model chi-square	10220.532	80.542	301.142

Table 5.3: Efficacy CFA - Measures of Fit

Comparing only the two restricted models allows for further interpretations. While both restricted models outperform the unrestricted model on every measure, the restricted model with only positive features strongly outperforms the restricted model with only negative features. Here it is important to remember that we cannot explicitly specify the factor that is being estimated, we are only able to use theory to identify the latent construct and specify the number of factors. The CFA then estimated the single most important factor. However, if we assume that both sets of differently orientated items are designed to measure self-efficacy, which does seem very plausible, it appears that negative features perform significantly worse than do positive features with regard to reflecting the underlying construct of self-efficacy. This finding again corroborates previously posed interpretations that recoded negative features are a noisy reflection of positive features, and their grouping provides internally inconsistent information. This is possible evidence of confusion in response that is caused by negatively worded items.

Table 5.4 shows the factor loadings for one unrestricted and two restricted CFA models fit to the efficacy factor data. The interpretation of the results is the same for those shown in table 5.2. As before, the orientation of the features is revealed by the codes containing either “p” or “n”. From the unrestricted model, it is evident that all five negative features load onto the efficacy factor less strongly than do the four positive factors. In some cases, such as n1 and n3, this loading is significantly weaker. The restricted models, which each include only factors of a single orientation, have much higher factor loadings on average.

This finding indicates that, when features are separated into their distinct orientations, the CFA models perform better. When viewed in combination with earlier findings in this paper, the results in Table 5.4 appear to indicate that the combination of positive and negative features is harmful to the estimation of a single factor. That is, positive and negative features do not reflect the same underlying latent construct equally well. A stronger interpretation is that items of different orientations are not capturing the same underlying latent construct, the one may be capturing a noisy version of the other. It is a superior strategy to use features based on only one orientation of response items. A bolder interpretation would be that only positive features should be used. Importantly, One cannot generalise these findings indiscriminately. However, these findings do demonstrate that researchers should approach empirical work and factor modelling with an *a priori* understanding that similar issues may exist in the data that they plan to use.

Factor_Loadings	Unrestricted	Resctricted_Pos	Resctricted_Neg
p1	1.000	1.000	-
n1	0.588	-	1.000
n2	0.857	-	1.144
p2	1.123	1.159	-
n3	0.594	-	0.994
p3	1.157	1.211	-
p4	1.157	1.192	-
n4	0.871	-	1.416
n5	0.899	-	1.320

Table 5.4: Efficacy CFA - Factor Loadings

6. Conclusion

The results contained in this paper indicate that homogenous positive and negative features differ with regards to their ability to capture information of the same underlying latent constructs. That is, the two distinct orientations of questionnaire response items do not illicit identically aligned responses. Rather, the responses to negatively worded items appear to be noisy reversed reflections of those to positively worded items. This finding is of particular significance in the field of latent construct estimation, where the accurate estimation of factors and capturing of underlying latent constructs is essential for appropriate inference. Moreover, this paper finds that traditional measures of sampling adequacy fail to reveal this informational distinction based on response item orientation.

These findings reveal that the common practice of including negatively worded response items in survey questionnaires should be done with caution. Moreover, the findings reveal that the equally common practice of recoding features based on negatively worded items and continuing with subsequent estimation procedures may result in biased outcomes. This bias will result from the measurement error contained in features that, due to the nature and orientation of the item upon which they are based, do not adequately capture the information that they are intended to capture as per their design. Therefore, researchers should carefully consider the effects that the use of such features could have before proceeding with estimation. These results appear robust, at least in the South African context. The central findings are consistent across basic correlation analysis, hierarchical cluster analysis, and confirmatory factor analysis.

These results are particularly pertinent in the Developing Economy context in which large portions of the student population do not possess age-appropriate reading skills. If a large proportion of the analysed sample is incapable of adequately interpreting and answering survey response items, research conducted using this data could derive misleading results. However, one shortcoming of the research of this paper is the limited focus on only the mathematical performance of South African grade nine learners, and the use of only two example latent constructs.

There exists vast scope for future research that can leverage the methods used, and the results found, in this paper. For example, further study can include a wider array of example latent constructs, or the complexity of response item wording can be adjusted. Another productive area to research would

be cross-country comparisons. South Africa has a very unique education system, therefore, one cannot easily generalise results found in the South African context. Furthermore, research can better identify the source of the dissimilarity between responses to items of different orientations. There is much more that can be done with this research direction. Importantly, the findings from such research will have practical implications that can easily be incorporated into future work.

7. References

Brown, T.A. 2015. Confirmatory factor analysis for applied research. Guilford publications.

Cerny, C.A., & Kaiser, H.F. 1977. A study of a measure of sampling adequacy for factor-analytic correlation matrices. Multivariate Behavioral Research, 12(1): 43-47.

Chiavaroli, N. 2017. Negatively-worded multiple choice questions: An avoidable threat to validity. Practical Assessment, Research, and Evaluation, 22(1): 3 - 18.

Cronbach, L.J. 1951. Coefficient alpha and the internal strucuture of tests. Psychometrika, 16(3): 297-334.

Guttman, L. 1945. A basis for analyzing test-retest reliability. Psychometrika, 10 (4): 255-282.

Hartley, J. 2014. Some thoughts on Likert-type scales. International journal of clinical and health psychology, 14(1): 83-86.

Hunt, E. 1978. Mechanics of verbal ability. Psychological Review, 85(2): 109 – 130.

Jöreskog, K.G. 1967. Some contributions to maximum likelihood factor analysis. Psychometrika, 32(4): 443-482.

Kaiser, H. 1974. An index of factor simplicity. Psychometrika, 39: 31–36.

Kaiser, H.F. and Rice, J. 1974. Little jiffy, mark IV. Educational and psychological measurement, 34(1): 111-117.

MacCallum. 2009. Factor Analysis, in Millsap & Maydeu-Oliveras. Quantitative Methods in Psychology. 123 - 147.

Marsh, H.W. 1986. Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. Developmental Psychology, 22(1): 37 - 49.

McDonald, M.E. 2013. The Nurse Educator's Guide to Assessing Learning Outcomes. (3rd ed.). Burlington, MA: Jones and Bartlett.

Spaull, N. 2013. Poverty & privilege: Primary school inequality in South Africa. International Journal of Educational Development, 33(5): 436-447.

Spaull, N & Hoadley, U. (2017) Getting Reading Right In: Jamieson L, Berry L & Lake L (eds) South African Child Gauge 2017. Cape Town, Children's Institute, University of Cape Town.

Suhr, D.D. 2006. Exploratory or confirmatory factor analysis?.

Thurstone, L.L. 1947. Multiple factor analysis, Chicago, IL: University of Chicago Press.

Weijters, B., Baumgartner, H. and Schillewaert, N. 2013. Reversed item bias: An integrative model. Psychological methods, 18(3): 320 – 334.

Wong, N., Rindfleisch, A. and Burroughs, J.E. 2003. Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. Journal of consumer research, 30(1): 72-91.

8. Appendix A: Measures of Sampling Adequacy

Code	Raw_Alpha	Std_Alpha	Lambda	KMO
p1	0.88	0.88	0.88	0.93
n1	0.90	0.90	0.89	0.89
n2	0.89	0.90	0.89	0.90
p2	0.89	0.89	0.89	0.96
p3	0.87	0.87	0.87	0.91
p4	0.88	0.89	0.89	0.95
p5	0.88	0.88	0.88	0.94
p6	0.88	0.88	0.88	0.95
p7	0.87	0.88	0.88	0.93
Overall	0.89	0.9	0.9	0.93

Table 8.1: Enjoyment Construct - Measures of Sampling Adequacy

Code	Raw_Alpha	Std_Alpha	Lambda	KMO
p1	0.78	0.78	0.80	0.86
n1	0.79	0.80	0.81	0.85
n2	0.78	0.78	0.80	0.88
p2	0.78	0.78	0.80	0.84
n3	0.80	0.80	0.82	0.87
p3	0.79	0.78	0.80	0.83
p4	0.79	0.79	0.80	0.85
n4	0.78	0.78	0.80	0.82
n5	0.77	0.78	0.80	0.83
Overall	0.8	0.81	0.83	0.85

Table 8.2: Efficacy Construct - Measures of Sampling Adequacy

The value in analysing these traditional and more commonly used tests of sampling adequacy lies in their comparison with the novel methods used in this paper. In practice, researchers will generally perform one or more of the above tests to investigate whether or not each feature within a group of features derived from similar response items contains homogenous information (Cerney & Kaiser, 1977). It is a common step in factor modelling. A practical example is asset index creation. It is common for researchers to input several features based on response items regarding asset ownership into a PCA, and use its first component as a feature in subsequent analysis. Before running this PCA, it is common to perform a KMO test for sampling adequacy to investigate whether or not the proposed grouping of features is appropriately homogenous for use in PCA (Kaiser, 1974; Cerney & Kaiser, 1977). If the KMO statistic is within an acceptable range, the researcher can be assured that their proposed grouping of features can be used in PCA, and that the use of the first component in the subsequent analysis is justified and appropriate.[8] The test results in Tables 8.1 and 8.2, and results throughout the paper as a whole, indicate that this approach may be flawed.

[8]Note here that the definition of this acceptable range is inconsistent. Some sources indicate that a value greater than 0.5 is sufficient (Cerney & Kaiser, 1977). However, it is more commonly accepted that values between 0.75 - 0.9 are adequate.

From the results in Tables 8.1 and 8.2, it is evident that all nine features used to identify either of the two latent constructs are sufficiently homogenous according to all included tests for sampling adequacy. The results in these tables indicate that using these features in procedures of factor estimation, or interchangeably in regression, is appropriate and statistically justified. Simply, the traditional tests of sampling adequacy find no internal inconsistencies among either group of nine features. Importantly, inference regarding the use of either group of features as a whole is done using the overall measure. Each overall measure, for each group and test statistic, is above 0.8. Interestingly, the overall values for the group of features associated with the efficacy factor are lower than those of the enjoyment factor, noting that this grouping contains more negative features. However, viewed in isolation, the overall test statistics from the efficacy feature grouping do not raise concern. Therefore, these overall test statistic values are misleading when we consider the findings in the other sections of this paper.

The misleading nature of these overall test statistic values forms part of the rationale for the work done in this paper. Relying on these measures of sampling adequacy alone may be inappropriate in certain circumstances. The typical process of estimating a selected test of sampling adequacy and then continuing with regression or factor estimation is not sufficient, a more thorough examination of data should be performed. Furthermore, the test statistic values for the individual features reveal no distinction between the measures of sampling adequacy for features derived from items of different orientation. From these tests alone, there is no notable evidence that features may be informationally different from one another based on their orientation.

9. Appendix B: Measures of Fit

9.1. Model Test Statistic

This is the most basic test statistic of the CFA model, several other fit measures are based on some function of this measure. It is a test statistic derived as a function of the sample size and the fit function (F_{ml}). Therefore, the test captures the dispersion matrix, the variance-covariance matrix, and a trade-off with the sample size. Therefore, it is a measure of the overall fit and the discrepancy between the sample and fitted covariance matrices. It is a test for perfect fit, with a smaller value indicating a better fit. A value of zero would indicate a perfect fit. A perfect fit would indicate that the variance of all included features is perfectly explained by a single factor of underlying covariance. The model test statistics is calculated as

$$T = (n-1)F_{ml} \tag{7}$$

In asymptotically large samples, and given a sufficiently large m, T follows a χ^2 distribution with degrees of freedom equal to the number of unique variance and covariance in the variance-covariance matrix of the observed variables. The degrees of freedom are calculated as follows, where q is the number of freely estimated parameters and m is the number of features.

$$df = \frac{m(m+1)}{2} - q \tag{8}$$

9.2. Comparative Fit Index

The comparative fit index, as the name suggests, can only be used to compare the fit of two competing models, it is not an absolute measure. The Comparative Fit Index compares the fit of the estimated model with that of the null model. The null model, in this case, is the model with the worst possible fit. In the null model, features have zero covariance. The null model would be the model with the maximum possible Model Test Statistic described above. Therefore, it is expected that the estimated model will have a better fit than the null model, the Comparative Fit Index is a measure of how much better the estimated model is than the null model. A higher value is prefered. The Comparative Fit Index is calculated as

$$CFI = \frac{(T_{null} - df_{null}) - (T_{estimated} - df_{estimated})}{T_{null} - df_{null}} \tag{9}$$

9.3. Tucker-Lewis Index

The Tucker-Lewis Index is a comparative fit index that improves upon the omitted Bentlre-Bonett index in that it penalises additional parameters. Therefore, it is a fit index that favours a more parsimonious model. The Tucker-Lewis index, as well as the abovementioned Comparative Fit Index, provides information on how the estimated model fit improves upon the fit of the null model. A fit

measure of 0.95, for example, indicates that the estimated model improves upon the fit of the null model by 95%. The measure depends on the average correlations among the set of features. If the average correlation is low, the fit measure will be small. Therefore, a large value is prefered. The Tucker-Lewis Index is calculated as

$$TLI = \frac{(\frac{T}{df})_{null} - (\frac{T}{df})_{estimated}}{\frac{T}{df}_{null}} \tag{10}$$

9.4. Root-Mean Square Error

The Root-Mean Square Error is an absolute fit measure that assesses the lack of fit of a model. If two individual models are run, one a restricted version of the other, this measure can then be used to compare the relative fit of the two estimated models. As this measure indicates the lack of fit, a smaller value is prefered. A smaller value is prefered. The Root-Mean Square Error is calculated as follows, where N is the number of observations.

$$RMSE = \sqrt{\frac{(\frac{T_{estimated}}{N-1})}{df_{estimated}} - \frac{1}{N-1}} \tag{11}$$

9.5. Standardized Root Mean Square Residual

The Standardized Root Mean Square Residual is an absolute measure of fit, it is defined as the square-root of the difference between the residuals of the sample covariance matrix and the hypothesized model. It can be interpreted as the average standardized residual covariance (Maydeu-Olivares. 2017). Therefore, it is a measure of the covariance that remains after the influence of the single factor has been partialed out. A smaller value is preferred. The Standardized Root Mean Square Residual is calculated as

$$SRMR = \sqrt{\frac{S}{\frac{m(m+1)}{2+m}}} \tag{12}$$

9.6. Log-likelihood of Estimated Model

The maximum log-likelihood value is derived from the maximum likelihood estimation procedure of CFA. This measure is useful only when compared to that of competing models, it is a comparative and not an absolute measure. It is the value at which the numerical optimisation procedure applied to the log-likelihood function reaches its final iteration. This is the value at which the likelihood function is optimised. A blunt interpretation of this measure is that it reveals how likely it is that the observed features are produced by the fitted model. Therefore, a larger value is prefered.

9.6.1. Akaike Information Criterion

The Akaike Information Criterion is a measure of fit that is widely used and has application beyond the CFA literature. It is a purely comparative measure that has no interpretative value when viewed in isolation. It is a useful measure of fit in that it punishes over-parameterization, it is a measure that considers the trade-off between fit and parsimony. In this case, we would expect the measure to be biased toward the restricted models, as they will provide a relatively similar fit but with considerably more degrees of freedom. A smaller value is prefered. The Akaike Information Criterion is calculated as follows, where q is the number of freely estimated parameters and L is the log-likelihood value.

$$AIC = 2q + 2ln(L) \tag{13}$$

9.7. Bayesian Information Criterion

The Bayesian Information Criterion is theoretically very similar to the Akaike Information Criterion. However, the Bayesian Information Criterion places a stronger penalty of model complexity, it is more inclined to favour a parsimonious model than is the Akaike Information Criterion. A smaller value is prefered. The Bayesian Information Criterion is calculated as

$$BIC = ln(N)q - 2ln(L) \tag{14}$$

9.8. Measures of Fit Summary Table

	Fit_Measure	Preferred_Value
1	Comparative Fit Index	Larger
2	Tucker-Lewis Index	Larger
3	Loglikelihood user model	Larger
4	Akaike Information Criterion	Smaller
5	Bayesian Information Criterion	Smaller
6	Root-Mean square error	Smaller
7	Standardized Root Mean Square Residual	Smaller
8	model chi-square	Smaller

Table 9.1: Measures of Fit Summary