Transforming Research Methods in the Social Sciences

Published by Wits University Press

Transforming Research Methods in the Social Sciences: Case Studies from South Africa.
Project MUSE. muse.jhu.edu/book/88982.

For additional information about this book
https://muse.jhu.edu/book/88982

For content related to this chapter
https://muse.jhu.edu/related_content?type=book&id=3356599
Introduction

South Africa, as with many other developing countries, is a consumer of psychological tests that have mainly been developed in the West. All too often the tests are yet to be standardised for the South African multi-ethnic populace. Good psychometric properties are not spatially and even temporally transferable. Using a psychological measure in a different context requires that a prior validation process be undertaken to deal with potential test bias and to circumvent undesirable social, economic and personal consequences with serious ethical undertones (Claassen, 1997). Regarding social and economic consequences, one need not recount the role of social scientists, including psychologists, in addressing the ‘racial’ question in South Africa, where psychological measures were used to bolster ideologies rather than objectively reflect the nature of the phenomenon under study (Louw & Danziger, 2000; Makhubela & Mashegoane, 2016).

It is not surprising that post-apartheid politicians in South Africa sought to curtail the hitherto unchecked use of tests and their subsequent abuse in decision-making. Section 8 of the Employment Equity Act (No. 55 of 1998) requires, amongst other things, that a test possess demonstrable validity and lack of bias for it to be used to make important decisions. Thus, validation, at least in terms of the law, takes centre stage in test use. Validation generally connotes evaluation of the psychological measure for its context-specific psychometric properties and suitability, before it is relied on for making decisions. The process is much more complex than back-translating the instrument and finding equivalent terms and phrases for original and special terminology. Scale validation ought to include evaluating the measurement equivalence (these terms are defined later) of the measure between the source and recipient groups.

There is serious risk of maleficence or, at the least, measurement error in using an instrument created in another country with the South African population without revalidating it. Psychological concepts can be understood distinctively across cultures, and construct-irrelevant variance in test performance (arising from a source such as group membership) is also possible (Xu & Tracey, 2017). The validation process of psychological measures is performed within a framework of models concerning the underlying trait (occasionally compared
among groups), and associated psychometric procedures and interpretations (Dimitrov, 2010). This involves, among other processes, the structural aspects of the construct, and justifiable relations between the construct and associated external criteria.

An instrument can be validated by means of different methods, including factor analytic procedures. This chapter focuses on the use of factor analysis (i.e. exploratory factor analysis [EFA], confirmatory factor analysis [CFA] and multigroup CFA) in the validation of a psychological measure, procedures to investigate internal relations among observed measures used to operationalise a theoretical construct domain (i.e. structural validity) and whether the factor loadings, intercepts and error variances of such a latent construct are equal across groups (i.e. measurement invariance, or MI). The concept of factor analysis is introduced, and then an example of the analysis is provided using data from South Africa. The Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965), a popular measure of global self-esteem, is validated with a South African sample to illustrate these methods. The reader is referred to a number of decision-making procedures that accompany the use of factor analysis in the test validation process.

Factor analysis as a validation method

Validity theory has substantially evolved over the last century in response to the increased use of assessments in society. This evolution has seen a shift to contemporary validity theorising and practice which, while still considering it necessary to demonstrate all types of validity, nevertheless regards predictive, concurrent and content validities to be fundamentally provisional (Dimitrov, 2010; Messick, 1995) and construct validity as the ‘whole of validity from a scientific point of view’ (Loevinger, 1957, p. 636). Practitioners wishing to investigate the construct validity (i.e. demonstration that a test actually measures the concept it purports to measure) of an instrument have the following statistical procedures at their disposal: interscale correlations, factor analysis and item response theory (Rasch model). Factor analysis, the focus of this chapter, has two main classes, namely EFA and CFA.

While there seems to be a general grasp of convergent and discriminant validity methods, confusion abounds among social science researchers as to which factor analytic method to use to answer which question and under what conditions. Many psychology practitioners in South Africa have little, if any, formal EFA and/or CFA instruction, leading to inadequate and curious use of the methods (e.g. Mthembu, 2015). This chapter elucidates some of the presumptions and uses of the techniques. It also pays special attention to multigroup confirmatory factor analysis (MGCF), a special method of executing CFA, because this critically important aspect of construct validity has not received widespread use in South Africa. MGCF is now considered the yardstick for determining the extent to which measures are equivalent for different groups (Chen, 2008). This is done
using a study on the RSES. This study examined the dimensionality and cross-cultural utility of the RSES in a relatively diverse sample of South African students. Notwithstanding the extensive application of the RSES across diverse cultural groups in southern Africa (Westaway, Jordaan & Tsai, 2015), studies are yet to settle the factorial validity and MI of the scale across groups of respondents.

We tested for the factorial validity and MI of the RSES across black \((n = 579)\) and white \((n = 291)\) university students. The analysis was conducted using both the Statistical Package for the Social Sciences (SPSS) version 23.0 and Analysis of Moment Structures (AMOS) version 23.0 (Arbuckle, 2014) software. EFA was examined using the principal axis factoring (PAF) method of estimation, while CFA employed the maximum likelihood (ML) estimation method to test the factor structure.

**Exploratory factor analysis**

The first step in the RSES study involved conducting the EFA. This technique, also referred to as common/principal factor analysis, has traditionally been used to explore the latent structure of a set of observed variables or of an area of functioning as examined by a particular measure, without hypothesising a predetermined structure (Schmitt, 2011). The approach also assists researchers to develop a structural theory: to determine the number of constructs underlying a set of items or observed variables, choose good measures of a construct/factor and define the content or meaning of latent constructs/factors. A related but distinct data extraction/variable reduction method is principal component analysis (PCA). Both PCA and EFA belong to the same family of analytical procedures, namely factor analysis (cf. Tabachnick & Fidell, 2007), leading to a misunderstanding. While PCA and EFA are at times mistakenly thought to be similar statistical methods, the former approach focuses on the analysis of all the variance of the observed variables/items, as opposed to the latter’s focus on only the common/shared variance among items (Schmitt, 2011). It has generally been accepted that under certain circumstances the differences between the two approaches are inconsequential and, as such, practitioners can choose either of the two. For example, in instances where the communalities of PCA and EFA are all close to one, analysis tends to yield similar structural outcomes. Nonetheless, each retains its specific role in test validation, as noted.

**Determining sample size**

There has generally been a difficulty in determining empirically derived sample-size requirements for factor analysis. It is common knowledge that factor solutions derived with relatively small sample sizes tend to be unstable and difficult to replicate (MacCallum & Tucker, 1991). The following rule-of-thumb sample-size guidelines have been proposed: participant-to-parameter ratio; subject-to-variable ratio \((10:1)\); Nunnally, 1978); absolute minimum sample required (a sample of not less than 100 participants and preferably more than 200;
Establishing factorial validity of the Rosenberg Self-Esteem Scale

Izquierdo, Olea & Abad, 2014); and variables-to-expected-factors ratio (Cattell, 1978). These heuristic sample-size estimation methods should be used with circumspection.

While it is generally accepted that sample-size requirements depend on the number of parameters in the model, it should also be noted that aside from the sample size, the degree of communalities and the extent of factor overdetermination are critical predictors of the precision and power of factor solutions (MacCallum, Widaman, Zhang & Hong, 1999). The total sample size for this study was 870 university students. This sample size was determined using various sample-size estimation criteria (Gaskin & Happell, 2014; MacCallum et al., 1999). Inaccurate sample-size estimations for factor analysis often lead to the over- or underselection of respondents. This not only has an impact on the dependability of the factor solutions produced but could also result in scientific and ethical consequences (i.e. inconclusive results, inaccurate interpretations of findings or statistically significant results, albeit with negligible clinical significance) (Gaskin & Happell, 2014).

Choosing the number of factors to extract/retain

Amongst conventional factor extraction methods, the Kaiser-Guttman criterion (eigenvalue > 1.00; Kaiser, 1960) and Cattell’s (1966) scree plot test are the most regularly used methods for retaining the number of factors in factor analysis and are default settings in many statistical analysis programs. The former tends to miscalculate the number of dimensions to retain (O’Connor, 2000), while the latter is subjective (e.g. eye-ballung eigenvalues plot for the elbow). Instead, the literature (O’Connor, 2000) favours the use of both parallel analysis (PA) and the minimum average partial (MAP) (Horn, 1965; Velicer, 1976) methods due to their precision, psychometric integrity and negligible inconsistency of the results (Zwick & Velicer, 1986). PA (Horn, 1965) focuses on the number of factors that explain more variance than on the factors resulting from random data.

Velicer’s (1976) MAP test attends to the relative number of residual systematic and unsystematic variance in a correlation matrix after extractions of increasing numbers of factors. Procedures and scripts for running these methods are available for SPSS, the Statistical Analysis System (SAS) and R (O’Connor, 2000). Researchers also have new methods such as the Hull method (Lorenzo-Seva, Timmerman & Kiers, 2011) and Ruscio and Roche’s (2012) comparison data (CD) at their disposal (although more empirical evidence is still needed on their utility). It is recommended that a combination of the traditional and more advanced procedures works better and that the factors produced must be interpretable and make theoretical sense (also see Zygmont & Smith, 2014). In our study, PA and MAP suggested that one component should be retained. For instance, Velicer’s MAP test revealed that a one-factor solution resulted in the lowest average squared correlation of $r^2 = 0.19$.

Choosing a data extraction method

Various data extraction procedures have been proposed to aid in the estimation of common factor analysis models. These methods include unweighted least
squares method, PAF, the ML method, alpha factor analysis, image factor analysis and generalised least squares method. Researchers need to select extraction procedures judiciously and in keeping with their objectives, because these methods produce different solutions (Gaskin & Happell, 2014). Extraction procedures vary in the assumptions that undergird them: aims, characterisation of uniqueness and mechanisms for calculating commonality values and factor scores (MacCallum, Browne & Cai, 2007). In this study the PAF method of estimation was chosen to examine the factor solution of the common/shared variance only, rather than the whole correlation matrix.

Selecting a factor rotation method

After extraction, the retained factors are normally rotated to simplify the factor solution and render it more interpretable. Unrotated factors tend to be ambiguous and indiscernible. Rotation seeks to attain optimal simple or parsimonious structures that are easier to interpret. A simple structure relates to a small number of factors that explain most of the observed variance in a larger set of items. Factor rotation can either be oblique, allowing solutions with correlated factors, or orthogonal, keeping factors uncorrelated. Within EFA, oblique rotation uses methods such as direct oblimin, quartimin and promax procedures, while orthogonal uses procedures such as varimax, quartimax and equamax (Thompson, 2004). The following rotation methods should be considered when aiming to allow for cross-loadings in the analysis: 1) larger cross-loadings: CF-equamax or CF-facparsim; and 2) fewer cross-loadings: geomin or CF-quartimax (Schmitt, 2011). Both rotations are default options in most statistical analysis programs. Safe to say the choice of rotational method depends largely on what rotation method is established within a particular field of study and the objectives of the analysis. For our study on the RSES, rotation was not necessary since only one factor was extracted.

Evaluating factor loadings

Factor loadings within EFA are considered to be meaningful and practically significant when > 0.30, although loadings > 0.40 are more preferable (Kline, 1994; Tabachnick & Fidell, 2007). Another important consideration here is that factor solutions should explain a substantial percentage of the total variance of the measured variables (see Streiner, 1994). A minimum item loading of about 0.30 amounts to about 10% overlapping variance with other items with the same factor. Parsimonious factor structures are solutions with consistently high communalities, item loadings > 0.30, and without free-standing, low-loading and cross-loading items and factors with less than three items (Costello & Osborne, 2005).

While item communalities of ≥ 0.80 are deemed to be high (Velicer & Fava, 1998), communalities of around 0.40 to 0.70 (low to moderate) are typically accepted in social science research. Item communalities below 0.40 would suggest that more factors should be investigated or that the said item is not associated to other items. Items are said to cross-load if they load at ≥ 0.30 on more than one factor. Item cross-loading may be suggestive of problems related to
item construction or the hypothesised factor solution. Factors containing less than three items are mostly poor and unstable, whereas those with ≥ 5 significantly loading items are advisable (Costello & Osborne, 2005).

Consistent with the PA and MAP results, the EFA of the RSES in our study revealed variance discontinuities that suggested a single latent factor. Collectively, the results indicated that a one-factor solution was optimal. The one factor accounted for 33.59% of the variance, with an eigenvalue of 3.36. Table 4.1 presents the factor matrix with factor loadings.

<table>
<thead>
<tr>
<th>Principal factors</th>
<th>Self-esteem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance explained</td>
<td>33.59%</td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>3.36</td>
</tr>
</tbody>
</table>

**Table 4.1 Factor matrix of the RSES**

<table>
<thead>
<tr>
<th>Item descriptor</th>
<th>Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSES 1</td>
<td>0.63</td>
</tr>
<tr>
<td>RSES 2</td>
<td>0.63</td>
</tr>
<tr>
<td>RSES 3</td>
<td>0.60</td>
</tr>
<tr>
<td>RSES 4</td>
<td>0.60</td>
</tr>
<tr>
<td>RSES 5</td>
<td>0.58</td>
</tr>
<tr>
<td>RSES 6</td>
<td>0.54</td>
</tr>
<tr>
<td>RSES 7</td>
<td>0.48</td>
</tr>
<tr>
<td>RSES 8</td>
<td>0.45</td>
</tr>
<tr>
<td>RSES 9</td>
<td>0.30</td>
</tr>
<tr>
<td>RSES 10</td>
<td>0.30</td>
</tr>
</tbody>
</table>

Source: Authors

**Confirmatory factor analysis**

CFA manifestly hypothesises a latent structure (based on theory and/or empirical data) and tests its fit with the observed variance-covariance structure of the observed indicators. It also allows for the evaluation of relative fit of competing structural models (Schmitt, 2011). Construct validity is demonstrated if the dimensionality of the scale is consistent with the trait the instrument alleges to measure, is also consistent with theory, and the variable loadings are large (Dimitrov, 2010). In our study, the validity of the RSES factor structure that was derived empirically with the EFA was further tested with CFA. The factor solution is schematically portrayed in Figure 4.1.

**Data quality and normality**

Data derived from interval or quasi-interval scales (e.g. Likert-type scales) are usually necessary for the successful use of factor analytic methods. Tetrachoric
Section One: Quantitative methods

(for dichotomous items) and polychoric (for polytomous items) correlations can also be used as input when working with Likert-type data instead of correlation matrices (Pearson correlation coefficients) (Gorsuch, 1974). Similarly, multivariate normality (reported, for example, using Mardia’s multivariate skewness and kurtosis tests/estimates) is a mandatory requirement when using factor analytic methods, especially when using parameter estimation methods like ML. Robust alternatives, using the ML parameter estimates with standard errors and a mean-adjusted Chi-square test (MLM) and robust maximum likelihood estimator (MLR) (in programs such as LISREL, Mplus, EQS and lavaan), are available for use and provide the option of relaxing multivariate normality (Byrne, 2012; Muthén & Muthén, 2012).

Alternative estimation methods, like the robust weighted least squares (WLS) (WLSMV or WLSM estimator in Mplus and lavaan), are available for use with Likert-type data, while current versions of AMOS offer the Bayesian approach for such a task (Lee, 2007). Monte Carlo studies show that satisfactory solutions are difficult to achieve when violations of assumptions, such as a normal distribution, are not accounted for. In the instance that this occurs, most statistical

---

**Figure 4.1 Hypothesised one-factor model**

Source: Authors
packages, like AMOS and EQS, have the option of bootstrapping or robust/corrected statistics (like the ones mentioned above) as a solution to this problem (Efron & Tibshirani, 1993). Bootstrapping is employed to obtain unbiased standard errors when multivariate normality is violated. A related issue is that of missing data. Because of the impact that missing data may have on the analysis, a method such as full-information ML should be used to correct for missing data (Jöreskog & Sörbom, 1996). Missing data could represent either sampling or measurement error and this might affect factor analysis results adversely, but this is entirely reliant on how these types of data are treated. Bootstrapping was used to obtain parameter estimates in the CFA of the RSES due to the multivariate non-normality of our data (i.e. Mardia’s multivariate kurtosis = 695.81; c.r. = 680.78).

Sample adequacy
While five to ten participants per variable is a commonly utilised guideline in CFA, Jöreskog and Sörbom (1989) recommend ten participants per parameter estimated. The number of parameters estimated in CFA increases as more variables are added, thus making the model more intricate. As such, Jöreskog and Sörbom’s (1989) suggestion is aimed at keeping models simpler when conducting CFA. Our research sample ($N = 870$) meets the recommended model-based sample-size estimation standards for CFA (Gagné & Hancock, 2006).

Examining model fit
ML estimation is a common estimation method to assess how well the specified model adequately represents the data. Practitioners have the following fit indices to evaluate for model adequacy: 1) absolute fit indices: based on how adequately the a priori/hypothesised model fits the data (Jöreskog & Sörbom, 1993); 2) comparative/incremental/relative fit indices: obtained by comparing the independence model to the hypothesised and respecified models; and 3) parsimonious fit indices: based on adjustments of absolute and comparative fit indices. Given the absence of consensus in the literature on preferred fit indices (e.g. Bentler, 1990; Hu & Bentler, 1995, 1999; Kline, 2005), it is prudent that several fit indices (absolute, comparative/incremental and parsimonious), modification indices, related expected parameter changes and residual error terms (Arbuckle & Wothke, 1999) be used to assess model fit.

Fit indices typically used to assess the goodness-of-fit of models include the Chi-square statistic to $df$ ratio ($\chi^2/df$; absolute fit), the comparative fit index (CFI; relative fit), the standardised root mean-square residual (SRMR; absolute fit), the normed fit index (NFI; relative fit), the Tucker-Lewis index (TLI; relative fit), Akaike’s information criterion (AIC; relative fit), expected cross-validation index (ECVI; relative fit) and the root mean-square error of approximation (RMSEA; absolute fit) along with its related 90% confidence interval. Models are accepted as providing good fit if $\chi^2/df < 1.5$, TLI and CFI $\geq 0.95$, RMSEA $< 0.06$ and NFI $\geq 0.90$ (see Bentler, 1990; Bentler & Bonett, 1980; Hu & Bentler, 1995; Kline, 2005). The AIC and the ECVI are evaluated to compare the relative fit of competing models; the lower the values on both indices, the superior the fit of the model to the data.
A word of caution regarding fit indices is that they may suggest that a model fits well, when in fact not all aspects of fit are good. Marsh, Hau and Wen (2004) warn against a strict adherence to cut-off criteria such as those of Hu and Bentler (1995), leading to incorrectly rejecting models that would otherwise be acceptable (type I errors). Researchers are expected to also take supporting theory into account when using fit indices. Modification indices (MI) or Lagrange Multiplier (LM) tests are also normally examined to identify the most significant and meaningful model modifications, and to improve the fit of the models that have poor fit (Hu & Bentler, 1995).

Goodness-of-fit statistics related to the test of our RSES structural model are shown in Table 4.2. Results revealed a well-fitting model to the data and all parameters were statistically significant and had the expected signs (Table 4.3).

### Table 4.2 One-factor model of the RSES structure: Goodness-of-fit statistics

<table>
<thead>
<tr>
<th>Model</th>
<th>One factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>57.20</td>
</tr>
<tr>
<td>df</td>
<td>23</td>
</tr>
<tr>
<td>TLI</td>
<td>0.96</td>
</tr>
<tr>
<td>CFI</td>
<td>0.97</td>
</tr>
<tr>
<td>RMSEA</td>
<td>0.04</td>
</tr>
<tr>
<td>90% RMSEA CI</td>
<td>0.02, 0.05</td>
</tr>
<tr>
<td>SRMR</td>
<td>0.03</td>
</tr>
</tbody>
</table>

*Source: Authors*

Notes: $\chi^2$ = Chi-square; df = degrees of freedom; CI = confidence interval

### Table 4.3 Structural path coefficients for the data

<table>
<thead>
<tr>
<th>Items</th>
<th>Coefficient (SE)</th>
<th>C.R.</th>
<th>Standardised coefficient</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSE 1</td>
<td>0.35 (0.03)</td>
<td>13.53</td>
<td>0.50</td>
<td>0.25</td>
</tr>
<tr>
<td>RSE 2</td>
<td>0.49 (0.03)</td>
<td>15.26</td>
<td>0.58</td>
<td>0.34</td>
</tr>
<tr>
<td>RSE 3</td>
<td>0.37 (0.05)</td>
<td>6.86</td>
<td>0.25</td>
<td>0.06</td>
</tr>
<tr>
<td>RSE 4</td>
<td>1.00</td>
<td></td>
<td>1.37</td>
<td>−0.79</td>
</tr>
<tr>
<td>RSE 5</td>
<td>0.52 (0.03)</td>
<td>17.86</td>
<td>0.61</td>
<td>0.38</td>
</tr>
<tr>
<td>RSE 6</td>
<td>0.57 (0.03)</td>
<td>18.66</td>
<td>0.66</td>
<td>0.43</td>
</tr>
<tr>
<td>RSE 7</td>
<td>0.29 (0.03)</td>
<td>10.62</td>
<td>0.39</td>
<td>0.15</td>
</tr>
<tr>
<td>RSE 8</td>
<td>0.31 (0.04)</td>
<td>7.34</td>
<td>0.29</td>
<td>0.08</td>
</tr>
<tr>
<td>RSE 9</td>
<td>0.45 (0.02)</td>
<td>18.22</td>
<td>0.63</td>
<td>0.39</td>
</tr>
<tr>
<td>RSE 10</td>
<td>0.44 (0.03)</td>
<td>15.74</td>
<td>0.58</td>
<td>0.33</td>
</tr>
</tbody>
</table>

*Source: Authors*

Notes: All path coefficients are significant at 5% level of significance; $N = 870$. 
Establishing factorial validity of the Rosenberg Self-Esteem Scale

However, the $R^2$ associated to eight of the ten observed indicators show that the factor accounts for a considerable part of the variance (between 15.5% and 79.2%). Only items 3 and 8 had negligible $R^2$ (0.06 and 0.08), suggesting the likelihood that these items do not measure the same latent trait as the other eight (at least in this student sample). From these results, we concluded that the hypothesised model of the RSES structure adequately represented data for non-clinical South African university students.

Multigroup confirmatory factor analysis

Evidence of MI (the extent to which an instrument’s items or subtests have equal meaning across groups of testees) is a prerequisite for the fair and ethical use of psychological tests and their test scores across groups. MI presumes that comparisons between groups are done on the basis of equivalence of the structure of the measure(s) being used. Relatedly, respondents from different groups should understand and respond to the items of the measure(s) in the same way, and there should be no systematic and artefactual way that their test scores differ on any items (i.e. construct-irrelevant variance). It is only when the measurement parameters (factor structure, factor loadings, indicator intercepts and residual [error] variances) are invariant across groups that the differences between them can be interpreted with some validity (Byrne, 2006). In essence, in this chapter we examine for the degree to which properties and interpretations of test scores of a particular trait generalise across groups of people (see Messick, 1995).

Although there are other approaches for examining multigroup invariance (e.g. EFA methods; Zumbo, Sireci & Hambleton, 2003), we prefer the robust MGCFA framework (Byrne, 2004; Cheung & Rensvold, 2002; Meredith, 1993), which addresses configural, metric and scalar invariance, as opposed to just the similarity of factor patterns across groups. Contrary to other approaches for invariance limited only to the analysis of covariance structures (COVS), the CFA procedure presented herein is based on the comparisons of mean and covariance structures (MACS) and this permits the examination of scalar equivalence through the comparisons of mean levels of factors (Chen, Sousa & West, 2005). Most Structural Equation Modeling (SEM) programs (e.g. EQS, AMOS, LISREL and Mplus) provide platforms to run the analysis, with the researcher having a choice of running the analysis manually or in an automated way.

The most preferred approach of MGCFA involves initially running a measurement model, in which baseline CFA models of the hypothesised factor model, established in the previous analysis stages, are first conducted. If this baseline model proves to be consistent with the data, analysis then proceeds to test the invariance of this model across subgroups, using a series of ordered steps (Byrne, 2006; Chen et al., 2005; Wu, Li & Zumbo, 2007). The initial model specified is for configural invariance: that is, the same factor structure is concurrently estimated for both groups without any equality constraints being set on the parameter estimates. If this model is established to be consistent with the data, the analysis proceeds to impose a series of stricter between-group constraints to examine for
factorial invariance. The second model (metric or weak factorial invariance) is then estimated in which the factor loadings are constrained to be equal between groups. This model allows differences in factor variances and error variances but forces equal loadings between groups. Results of this specified model remaining consistent with the data will, in addition to MI, also imply invariant between-group variance in the latent variables or attribute examined by the indicators.

The third model (scalar or strong invariance) further constrains the item intercepts to be equal between groups, consequently forcing equality of the variances/covariance matrices across the groups. The fourth model is residual variance invariance (strict factorial invariance), wherein the residual variances (uniqueness or measurement error), in addition to factor loadings and intercepts of latent variables, are constrained to be equal between groups. However, there are disagreements in the literature as to the necessity of this aspect of MI (Chen et al., 2005; Widaman & Reise, 1997). Table 4.4 illustrates a series of multigroup models, each constituting a gradually more constrained parameterisation than its precursor. These models are considered to be hierarchically nested.

Table 4.4 Steps for multigroup analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>Constraint parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Configural invariance</td>
<td>Similar factor structure across groups (no equality constraints imposed)</td>
</tr>
<tr>
<td>2. Metric invariance</td>
<td>Same factor loadings constrained to be equal across groups</td>
</tr>
<tr>
<td>3. Scalar invariance</td>
<td>Factor loadings and item intercepts constrained to be equal between groups</td>
</tr>
<tr>
<td>4. Strict (residual) invariance</td>
<td>Factor loadings, item intercepts and uniqueness or measurement error constrained to be equal across groups</td>
</tr>
</tbody>
</table>

Source: Authors

MGCFA uses similar fit indices as the normal CFA. However, nested models are compared in sets by calculating the differences in their overall CFI and RMSEA estimates to examine for MI (i.e. ∆; difference value). While the $\chi^2$ difference value ($\Delta\chi^2$) is also calculated, the literature (Cheung & Rensvold, 2002; Little, 1997; Marsh, Hey & Roche, 1997) suggests that the $\Delta\chi^2$ value is similarly sensitive to non-normality and sample size as the $\chi^2$ statistic, thus questioning its reliability when it comes to offering evidence for invariance. Evidence of invariance is based on the following criteria: the multigroup model exhibits an adequate fit to the data, and the CFI and RMSEA values between models are negligible ($\Delta\text{CFI} \leq 0.01$ and when $\Delta\text{RMSEA} \leq 0.015$) (Chen, 2007; Cheung & Rensvold, 2002). In addition to the aforestated process pertaining to full MI, we also have partial MI – a less strict procedure for making group comparisons. Within the partial MI framework, only a limited portion of parameters in the model are constrained to equal across the groups, while others are left to vary (see
Establishing factorial validity of the Rosenberg Self-Esteem Scale

Vandenberg & Lance, 2000). The precondition for evaluating partial MI requires that configural invariance and partial metric invariance (whereby only minority parameters are allowed to differ between groups) be met, before proceeding with other levels of MI (Milfont & Fischer, 2010).

The MI results for our RSES study are as follows:

**Baseline models:** Tests of the hypothesised RSES structure (Figure 4.1) revealed a good fit to the data for both black ($\chi^2 = 25.51; \text{CFI} = 0.99; \text{SRMR} = 0.02; \text{RMSEA} = 0.02$, with 90% CI = 0.00 to 0.04) and white ($\chi^2 = 59.13; \text{CFI} = 0.96; \text{SRMR} = 0.02; \text{RMSEA} = 0.07$, with 90% CI = 0.05 to 0.10) participants. All parameter estimates were viable and statistically significant.

**Tests for factorial invariance:** Hierarchically nested multigroup models were tested across black and white university students, each comprising more restricted parameterisation than its precursor (Vandenberg & Lance, 2000). Results from the related tests for invariance are summarised in Table 4.5.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>127.71</td>
<td>193.58</td>
<td>915.92</td>
</tr>
<tr>
<td>df</td>
<td>44</td>
<td>53</td>
<td>77</td>
</tr>
<tr>
<td>CFI</td>
<td>0.95</td>
<td>0.93</td>
<td>0.57</td>
</tr>
<tr>
<td>SRMR</td>
<td>0.03</td>
<td>0.03</td>
<td>0.22</td>
</tr>
<tr>
<td>RMSEA</td>
<td>0.04</td>
<td>0.06</td>
<td>0.11</td>
</tr>
<tr>
<td>90% RMSEA CI</td>
<td>0.03, 0.05</td>
<td>0.05, 0.06</td>
<td>0.10, 0.12</td>
</tr>
<tr>
<td>Model comparison</td>
<td>2 vs 1</td>
<td>3 vs 1</td>
<td></td>
</tr>
<tr>
<td>$\Delta^*\text{CFI}$</td>
<td>0.02</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>$\Delta^*\text{RMSEA}$</td>
<td>0.02</td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>

*Source: Authors*

**Notes:** $p < .001; \chi^2 = \text{Chi-square test}; df = \text{degrees of freedom}; \Delta^*\text{CFI} = \text{comparative fit index difference value}; \Delta^*\text{RMSEA} = \text{root mean-square error of approximation difference value}.

In Table 4.5, tests of the hypothesised RSES structure, the configural model, revealed a good fit to the data for both black and white ($\chi^2_{[44]} = 127.72; \text{CFI} = 0.95; \text{SRMR} = 0.03; \text{RMSEA} = 0.04$, with 90% CI = 0.03 to 0.05) participants. All factor loadings were viable and statistically significant. This model serves as the baseline against which all remaining models are compared in the process of determining evidence of invariance. Furthermore, all MGCFA results for MI across race show that all the first two nested models represented a good fit to the data (CFIs = 0.93–0.95, SRMRs = 0.03, RMSEAs = 0.04–0.06). Model 2 (metric invariance), in which all factor loadings (i.e., measurement weights) were equally constrained, also represented a good fit to the data, but the resulting $\Delta^*\text{CFI} = 0.02,$
while the ∆RMSEA value = 0.02. Moreover, the fit of the models deteriorated when we assumed equal variance in the factor loadings and item intercepts. The lack of satisfaction of metric invariance implies that items of RSES do not have equal salience across black and white participants. These results suggest that the RSES provides an assessment of self-esteem that is not equivalent across race groups in South African university students. The absence of MI indicates that practitioners should use the RSES with some caution given the lack of generalisability of the instrument’s properties across race in university students.

Conclusion and implications

This chapter discussed procedures commonly used for studying the construct validity of assessment instruments in psychology. Its applications were illustrated by using self-esteem data from South African university students. EFA was done using PAF. MAP and PA methods suggested that only one factor could be extracted. Subsequently, a single-structure model was tested with the CFA, and it fitted the data well. This is the model that was used to illustrate the use of MGCFA to establish MI between whites and blacks in South Africa. Ensuring that the data are suitable for analysis, including the test of their normality, dealing with missing values and making sure that data are large enough, will make it possible that MGCFA will be conducted and its results can be relied upon to decide whether there is MI or not.

Methods like MGCFA offer potential solutions to the measurement bias problems of most of the psychological tests used in South Africa. Moreover, contemporary MGCFA approaches include procedures such as MACS that are on a par with other renowned approaches to examining for MI, like item response theory (IRT).

Continued use of tests that are clearly biased and lacking in fairness is a major problem in a multicultural context such as South Africa, and knowingly doing so constitutes an ethical dilemma in itself. The Employment Equity Act places the burden of proving test validity on test administrators. Consequently, South African practitioners have been amply warned, and methods of validating and improving the psychometric properties of measures circulating in the market have been suggested (Van de Vijver & Rothmann, 2004; Van de Vijver & Tanzer, 2004). Neglecting to implement on a wide scale the recommendations and suggestions of commentators is in itself incomprehensible.

The methods of validation described in this chapter assist test users to avoid the usual pitfalls of transporting tests and their accompanying constructs to a recipient context. Construct, method (instrument) and item bias are minimised, with reasonable prospects of eliminating them. Although some of the problems associated with cross- and multicultural use of tests may not disappear easily, at least they are dealt with sufficiently. As an example, scalar equivalence ensures the comparability of the measure in spite of the persistence of mean differences across groups.
Establishing factorial validity of the Rosenberg Self-Esteem Scale

References
Arbuckle, J. L. (2014). *Amos 23.0 user’s guide*. Chicago, IL: IBM SPSS.


