AN INVESTIGATION INTO THE MEASUREMENT INVARIANCE AND MEASUREMENT EQUIVALENCE OF THE SOUTH AFRICAN PERSONALITY INVENTORY ACROSS GENDER GROUPS IN SOUTH AFRICA

Sonja van der Bank

Thesis presented in partial fulfilment of the requirements for the degree of Master of Commerce in the Faculty of Economic and Management Sciences at Stellenbosch University

Supervisor: Prof C Theron

December 2019

DECLARATION

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signed: Sonja van der Bank

December 2019

ABSTRACT

Personality assessments are commonly used as predictor measures in employment selection due to substantial empirical evidence proving that personality constructs explain and predict employee performance and behaviour in organisational settings. Before conclusions can be made that inter-group differences in observed scores are caused by valid cross-group differences in the latent personality variables being assessed, the possibility of measurement bias being the cause must be nullified. Measurement bias refers to group-related error in the measurement of a specific construct carrying a specific constitutive definition. In this sense measurement bias refers to two hierarchically related questions, namely (a) whether the same construct, carrying a specific constitutive definition, is measured across groups, and if so (b) whether the same construct is measured in the same way across groups (i.e. whether a specific standing on the latent variable being assessed is associated with the same expected observed score or probability of achieving a specific observed score across groups).

Measurement bias comprises method bias, construct bias and item bias. The current study utilised a stringent definition of item bias that states that item bias occurs if the regression of observed item responses on the underlying latent dimension the item is designated to reflect, differs in terms of intercept (uniform bias), and/or slope (non-uniform bias) and/or error variance (error variance bias) across groups. When conceptualising measurement bias from the perspective of mean and covariance structure (MACS) analysis, the terms measurement invariance and measurement equivalence are typically used. Both measurement invariance and equivalence pertains to the question whether the slope, intercept or error variance of the regression of the item responses on the latent personality dimensions being measured differ across groups.

Dunbar et al. (2011) proposed a clear distinction between measurement invariance and measurement equivalence. Measurement invariance investigates whether a multigroup measurement model in which the factor structure (i.e. number of personality factors and the items' loading pattern on the factors) is constrained to be identical across multiple groups and in which (a) no parameters are constrained to be equal across the groups, (b) some parameters are constrained to be equal across the groups, fits the data obtained from two or more samples closely (Dunbar et al., 2011). The five hierarchical levels of measurement invariance include configural invariance, weak invariance, strong invariance, strict invariance and complete invariance (Dunbar et al., 2011). Measurement equivalence, investigates whether a multigroup measurement model in which the structure but no parameters is constrained to be equal across groups fits the data of multiple groups significantly better than a multigroup measurement model in which the structure and specific parameters are constrained to be equal across groups. Dunbar et al. (2011) also proposed four hierarchical levels of measurement equivalence, namely metric equivalence, scalar equivalence, conditional probability equivalence and full equivalence

The current study investigates the measurement invariance and measurement equivalence of the South African Personality Inventory (SAPI) across gender groups in South Africa. The SAPI demonstrated a lack of construct bias and a lack of non-uniform bias. The SAPI measured the same construct across the two samples groups, but the item content of some items were perceived and interpreted differently between the

two gender groups. Metric – partial scalar - partial conditional probability equivalence was demonstrated. Consequential implications and recommendations relating to the study findings for the test developers and human resource practitioners are discussed.

OPSOMMING

Persoonlikheidsassesserings word algemeen gebruik as voorspellers in seleksie van werknemers as gevolg van oortuigende empiriese bewyse wat daarop dui dat persoonlikheidskonstrukte werknemerprestasie en -gedrag verklaar en voorspel. Voordat daar egter gevolgtrekkings gemaak kan word dat intergroepverskille in waargenome tellings veroorsaak word deur geldige kruisgroepverskille in die latente persoonlikheidsveranderlikes wat geassesseer word, moet die moontlikheid van metingsydigheid uitgeskakel word. Metingsydigheid verwys na groepverwante foute in die meting van spesifieke konstrukte wat 'n spesifieke konstitutiewe definisie dra soos bepaal deur die toetsontwikkelaar. Metingsydigheid verwys in hierdie konteks na twee hiërargies verwante vrae, naamlik (a) of dieselfde konstruk, wat 'n spesifieke konstitutiewe definisie dra, oor groepe gemeet word, en indien wel, (b) of dieselfde konstruk op dieselfde wyse oor groepe gemeet word (d.w.s. of 'n spesifieke vlak op die geassesseerde latente veranderlike, oor groepe geassosieer word met dieselfde verwagte waargenome telling of waarskynlikheid om 'n spesifieke waargenome telling te behaal).

Metingsydigheid bestaan uit metodesydigheid, konstruksydigheid en itemsydigheid. Die huidige studie handhaaf 'n streng definisie van itemsydigheid wat daarop dui dat itemsydigheid plaasvind indien die regressie van waargenome itemresponse op die onderliggende latente dimensies wat die item aangewys is om te reflekteer, verskil in terme van afsnit (eenvormige sydigheid) en/of helling (nie-eenvormige sydigheid) en/of foutvariansie (foutvariansiesydigheid) oor groepe. Wanneer metingsydigheid vanuit die perspektief van gemiddelde en kovariansie-struktuur (MACS) analise gekonseptualiseer word, word die terme meting-invariansie en meting-ekwivalensie tipies gebruik. Beide meting-invariansie en -ekwivalensie hou verband met die vraag of die afsnit, helling en/of foutvariansie van die item -ntwoorde se regressie op die latente persoonlikheidsdimensies, verskil tussen groepe.

Dunbar et al. (2011) beklemtoon 'n duidelike onderskeid tussen meting-invariansie en meting-ekwivalensie. Meting-invariansie ondersoek of 'n multigroepmetingsmodel waarin die faktorstruktuur (d.w.s. die aantal persoonlikheidsfaktore en die items se ladingpatroon op die faktore) beperk word om identies te wees oor verskeie groepe en waarin (a) geen parameters beperk word om gelyk te wees oor die groepe, (b) sommige parameters beperk word om gelyk te wees oor die groepe, die data wat uit twee of meer steekproewe verkry word pas (Dunbar et al., 2011). Die vyf hiërargiese vlakke van meting-invariansie sluit in konfiguratiewe invariansie, swak-invariansie, sterk-invariansie, streng-invariansie en volledige invariansie (Dunbar et al., 2011). Meting-ekwivalensie ondersoek of 'n multigroepmetingsmodel waarin die struktuur maar geen parameters beperk word om gelyk te wees oor groepe, die data van veelvuldige groepe beduidend beter pas as 'n multigroepmetingsmodel waarin die struktuur en spesifieke parameters beperk word om gelyk te wees oor die groepe. Dunbar et al. (2011) het ook vier hiërargiese vlakke van meting-ekwivalensie voorgestel, naamlik metriese ekwivalensie, skalaar-ekwivalensie, voorwaardelike waarskynlikheid ekwivalensie en volle ekwivalensie.

Die huidige studie ondersoek die meting-invariansie en meting-ekwivalensie van die Suid-Afrikaanse Persoonlikheidsinventaris (SAPI) oor geslagsgroepe in Suid-Afrika. Die studie-resultate toon dat die SAPI 'n

gebrek aan konstruksydigheid en 'n gebrek aan nie-eenvormige sydigheid demonstreer. Die SAPI het dieselfde konstruk vir die twee groepe gemeet, maar die iteminhoud van die enkele items is verskillend waargeneem en geïnterpreteer tussen die twee geslagsgroepe. Metriese - gedeeltelike skalaar - gedeeltelike voorwaardelike waarskynlikheid ekwivalensie is gedemonstreer. Na aanleiding van die studie-resultate word implikasies en aanbevelings vir die toetsontwikkelaars en menslike hulpbronpraktisyns bespreek.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ACKNOWLEDGEMENTS

Prof Callie Theron, thank you for believing in each of your students, including me. Thank you for challenging me to apply my mind and changing the world, one assessment at a time. Your passion for the field of industrial psychology and the scientific foundations thereof is contagious. I salute you!

Prof Deon Meiring, thank you for entrusting me with the honour of evaluating the measurement invariance and measurement equivalence of the SAPI. Thank you for making the archival data available for me to conduct this research study.

Francois van der Bank my husband, best friend and chief cheerleader. You have encouraged me to push through, when my resilience was depleted. You reminded me of my purpose and true identity. I could not have done this without your love, support, patience and motivation. Thank you for all your sacrifices to allow me time to work on this research.

My children, Greeff and Karli van der Bank, thank you for understanding and allowing me to complete this research. May my example encourage you to strive to become the best you.

I want to thank my parents, Boet and Marthie Greeff, for teaching me the value of education and tertiary studies. Thank you for your practical and emotional support. Mom, I can still hear you shouting: “Go-go-go!!”

I want to thank my extended family for their endless encouragement and support during this research.

Then lastly, I want to honour the Lord Jesus Christ, for giving me a vision, hope and a future. Thank you for teaching me that my identity is in You. May this research glorify Your Name!

CHAPTER 1: INTRODUCTION

1.1. Introduction

In the global economy, organisations compete for market share to achieve sustainability in their three primary strategic objectives, namely environmentally friendly activities (planet), social responsibility regarding their employees and the broader community that the organisation serves and within which it functions, and lastly to maximise the return on investment (profit) (McWilliams, Parhankangas, Coupet, Welch, & Barnum, 2016; Palmer & Flanagan, 2016). To achieve and maintain their competitive advantage, organisations strive to function effectively and sustainably by allocating its limited resources optimally.

The importance of human capital in this striving lies in the fact that people manage and activate other production factors, thereby subsequently determining the utility effectiveness of all other resources (Marx, as cited in Moyo, 2009; Theron, 1999). The human resource (HR) function contributes to the achievement of the organisation's objectives through acquiring, allocating and managing the human resources to optimise the workforce's performance (Theron, 1999). The HR function applies various HR practices and interventions such as recruitment, selection, remuneration, training and development, and performance management to acquire and manage the workforce's performance (Noe, Hollenbeck, Gerhart, & Wright, 2010).

Selection plays a critical role in the value that the HR department adds to the organisation as it determines the flow of employees into and through the organisation, aiming to enhance the workforce's performance (Noe et al., 2010; Theron, 2007). Selection is constituted by decisions that are taken about the potential and current workforce. Wrong selection decisions yield high costs associated with recruitment and training (found to be up to $12,000 per employee in the American hospitality industry), loss in productivity (due to less employees and a steep learning curve when new employees are appointed), legislation, adverse impact and absenteeism to name but a few that could have been avoided with excellent selection decisions (Noe et al., 2010; Tracey & Hinkin, 2010). Optimal selection decisions maximise the utility payoff, which is the return on investment in the selection instrument (Theron, 2007). Selection decisions are based on the results of various selection methods. Howard and Thomas (2010) provided the following simple taxonomy for practitioners to distinguish between selection methods, although it might be deemed as an oversimplified classification: demonstrations of behaviour (e.g. administrative or interactive simulations), descriptions of behaviour (e.g. reference checks and career achievement records), or making inferences about behaviour[1] (e.g. personality tests).

The objective of selection decisions is to appoint the person who will eventually demonstrate an optimal level on the performance construct (η). The job performance level to be attained on-the-job serves as basis for the selection decision (Theron, 2007). Thus, the ideal would be to determine applicant suitability based on on-the-job demonstrations of performance (i.e. criterion). However, the level of the performance criterion is only attainable once the person is appointed (Theron, 2007). Yet, the likelihood of applicants performing with a certain level of proficiency can be inferred or predicted (either mechanically or clinically) from observed

[1] All assessment methods involve stimuli that elicit behaviour that reflects a person's standing on a latent variable. The behaviour that is elicited either directly reflects η or ξ, or indirectly through recall of behaviour in which η or ξ expressed itself. Inferences are therefor always made about η or ξ.

scores of off–the-job performance measure obtained in a content valid simulation of the job or in an alternative job highly similar to the target job (i.e. via a content orientated approach to selection) or from observed scores of person-cantered determinants of on-the-job performance (i.e. via a construct orientated approach to selection) (Binning & Barrett, 1989). The off–the-job performance measure or the measures of the person-cantered determinants of on-the-job job performance serve as predictor measures from which estimates of future on-the-job (or criterion) performance are derived clinically or mechanically (Theron, 2007). Such a prediction is only justified to the extent that (a) the predictor measures are shown to correlate with a performance measure, (b) the extent to which both the predictor construct(s) and performance construct(s) are reliably and construct validly measured by the respective measurement instruments, and lastly the manner in which the performance construct(s) and the latent construct(s) measured by the predictors are related in some specified manner is validly understood (Nunnally, as cited in Binning & Barrett, 1989). Personality assessments are commonly used as predictor measures in a construct-orientated approach to selection due to substantial empirical evidence proving that personality constructs explain and predict performance and behaviour in organisational settings (Fakir & Laher, 2015; Ones, Dilchert, Viswesvaran, & Judge, 2007).

1.2. Personality Assessment as Predictor During Employee Selection

South African industrial psychologists use personality assessments predominantly during selection to determine person-environment fit (Fakir & Laher, 2015). Despite harsh criticism against the use of personality assessments in selection half a century ago, the past two decades have seen a renaissance in statistical evaluations and meta-analyses in personality research in the work context (Barrick & Mount, 2005; Guion & Gottier, 1965). Research results demonstrate the relationship between personality and organisational outcomes, substantiating the case for using personality in employee selection to such an extent that Barrick and Mount (2005, p. 363) claim that “the statement that general mental ability predicts job performance better than personality is not entirely true”.

Research has identified personality as a successful predictor of work-related behaviours and organisational outcomes such as increased task performance, group success, organisational citizenship behaviour, job satisfaction and leadership effectiveness, as well decreased counterproductive behaviour, turnover, absenteeism, tardiness (Barrick & Mount, 2005; Hough, 2003). Big Five personality constructs successfully predicted both subjective and objective career success with (uncorrected) correlations ranging up to .49, with a joint multiple correlation of .60 when predicting occupational status and income (Judge, Higgins, Thoresen, & Barrick, 1999). Meta-analytic results have shown that the Big Five personality constructs together explain up to 25% of the variance in leadership, whilst demonstrating satisfactory correlations between leadership outcomes and individual personality constructs (Neuroticism ρ = -.24; Extraversion ρ = .31; Openness to Experience ρ = .24; Agreeableness ρ = .08; Conscientiousness ρ = .28) (Judge, Bono, Ilies, & Gerhardt, 2002). Another study found Conscientiousness, Emotional Stability and Openness to Experience to predict leader emergence (Conscientiousness ρ = .33; Emotional Stability ρ = .24, Openness to Experience ρ = .24) and leader effectiveness (Conscientiousness ρ = .16; Emotional Stability ρ = .22, Openness to Experience ρ = .24) (Colbert, Barrick, & Bradley, 2014). Colbert et al. (2014) found that a higher mean conscientiousness in a top management team resulted in a higher lagged financial performance across the entire organisation, as compared to top management teams with a lower mean Conscientiousness. In their meta-analysis

investigating the relationship between personality (in particular the Big Five constructs as a combined set) and different performance dimensions, Ones et al. (2007) reported multiple correlations between personality and individual overall performance (R = .27), counterproductive work behaviour (R = .45) and (team level performance with R = .60), to name only a few performance outcomes. In another recent meta-analysis, job level moderated the relationships of Emotional Stability and Ambition as predictors of overall adaptive performance with the two personality constructs, having a stronger influence on adaptive behaviour for managerial positions, than for employees (Huang, Ryan, Zabel, & Palmer, 2014). Huang et al. (2014) reported that ambition was the strongest predictor of proactive forms of adaptive performance, whilst Emotional Stability was the strongest predictor of reactive forms of adaptive performance. Meta-analytic results lead Barrick and Mount (2005) to conclude that of all the personality constructs, Conscientiousness and Emotional Stability remain important across different jobs since these constructs address emplyees' motivation (i.e. employees' "want to do a task") and competence (i.e. employees' "can do a task"). Furthermore, Emotional Stability has been shown to predict typical performance, whilst Opennes to Experience predicts maximal performance and Extraversion is shown to predict both typical and maximal performance (Barrick & Mount, 2005).

Not all authors share in this enthusiasm regarding the use of personality in personnel selection. Morgeson et al. (2007a, 2007b) argue that personality assessments predict overall job performance with very low validity (ranging from -.02 to .15), although they admit that slightly higher correlations are reported with contextual performance than with task performance. Notwithstanding this critique on personality testing in employee selection, several authors responded to refute Morgeson et al.'s (2007a, 2007b) statements. For instance Ones et al. (2007) report that the validity coefficients between personality and performance criteria range between .11 and .49, despite applying quite conservative corrections, while Tett and Christiansen (2007) insist that meta-analytic estimates for personality assessments are impressive and dependable. Adhering to Morgeson et al.'s concerns, it remains critical to ensure that personality assessments that are used in employee selection predict performance with satisfactory psychometric properties and be used in a responsible manner to prevent the potential negative impact that psychometrically questionable assessments might have in the workplace (Guion & Gottier, 1965; Moyo, 2009).

A case in point is the irresponsible and inappropriate manner in which psychometrically questionable psychological assessments were used in South African context in prior to the 1990's. Apartheid legislation supported this misuse of psychological assessments, leading to conclusions about intergroup differences without considering culture, socio-economic and other factors (Foxcroft & Roodt, 2010). However, since democracy in 1994, legislation such as the Employment Equity Act (EEA) (Republic of South Africa (RSA), 1998), supported by other authoritative guidelines such as the *International Guidelines for Test Use* (International Test Commission, 2001), have purposefully addressed such human rights violations by prohibiting the use of psychological assessments that are psychometrically questionable or biased against any subgroup. These regulatory changes have placed test developers under pressure to subject psychological assessments, and the manner in which they are used to inform decision-making, to sophisticated scientific analyses by assessing the psychometric appropriateness and relevance of assessments and the inferences derived from assessments to the South African context. For instance, test developers are required to empirically test the permissibility of their claim that observed scores from

personality assessment have the same meaning in terms of the underlying latent variables across different groups and that the instrument is not biased against any group, by investigating the assessment's measurement invariance and measurement equivalence. Some authors advocate that inferences from personality assessments may only be considered scientifically rooted when empirical evidence supports measurement invariance and measurement equivalence, and that without such evidence the scientific foundation for construct-referenced inferences across different groups will be considered severely lacking (Dunbar, Theron, & Spangenberg, 2011; Steenkamp & Baumgartner, 1998)[2].

1.3. Measurement bias, Measurement Invariance and Measurement Equivalence

It is acknowledged that inter-group differences in observed scores might be due to valid cross-group differences in the latent variable being assessed. However, before conclusions can be made about valid cross-group differences, the possibility of measurement bias and structural[3] bias must be ruled out. Measurement bias refers to group-related error in the measurement of a specific construct carrying a specific constitutive definition. In this sense measurement bias refers to two hierarchically related questions, namely (a) whether the same construct, carrying a specific constitutive definition, is measured across groups, and if so (b) whether the same construct is measured in the same way across groups (i.e. whether a specific standing on the latent variable being assessed is associated with the same expected observed score or probability of achieving a specific observed score across groups). Measurement bias in the latter sense refers to unwanted but systematic group-related sources that cause differences in observed scores that are not reflected in differences in the underlying latent construct measured (Meiring, Van de Vijver, Rothmann, & Barrick, 2005; Van de Vijver & Leung, 2001; Van de Vijver & Tanzer, 2004). Van de Vijver and Leung (2001) differentiate between three different types of measurement bias, namely: construct bias, method bias and item bias. Construct bias occurs when the construct that elicits the behavioural response to the items comprising the test differs (i.e. constructs are not identical) across groups, or when the behaviours that denote the construct of interest differ across groups (Van de Vijver & Leung, 2001). Construct bias therefore exists when the factor structure (or measurement model) implied by the constitutive definition of the construct being assessed and the design intention underlying the test is unable to satisfactorily account for the observed inter-item covariance matrix in all groups. Method bias in turn occurs when methodological strategies lead to a change in mean scores between groups, that are often erroneously interpreted as valid cross-cultural differences (Byrne & Watkins, 2003; Van de Vijver & Leung, 2001). Method bias is caused by sample incomparability, differential response to the instrument format (e.g. stimulus familiarity or response style) and lastly, administration bias which relates to discrepancies in the manner which the assessment is administered to the respondents (Byrne & Watkins, 2003; Van de Vijver & Leung, 2001). Finally, item bias, also known as differential item functioning, occurs when the meaning attached to item content is differentially interpreted across different groups (Byrne & Watkins, 2003; Van de Vijver & Leung, 2001). Test items serve as stimuli that elicit behavioural responses that denote a specific latent dimension of a construct. An item is a valid indicator of a specific latent dimension when the item responses correlate with the standing on the latent dimension (i.e. the regression of the item on the latent dimension has a statistically significant (positive or negative) slope). Item bias occurs if the regression of observed item responses on the underlying latent

[2] It is acknowledged though that the absence of construct and item bias in predictor measures is not a sufficient condition to ensure the absence of predictive bias in the criterion inferences that are clinically or mechanically derived from these predictor measures

[3] It is acknowledged that in order to unequivocally claim that an assessment is not biased structural bias will also need to be investigated. However, in the interest of parsimony the scope for the current study is limited only to measurement bias.

dimension the item is designated to reflect, differs in terms of intercept, slope and/or error (or residual) variance across groups. Three forms of item bias are distinguished, namely non-uniform, uniform and error variance bias. Item bias can be defined more leniently as tests where the expected item score, given a specific standing on the latent variable being measured ξ_c $E[X|\xi_c]$, differs across gender groups [i.e. $E[X|\xi_c; Group_{female}] \neq E[X|\xi_c; Group_{male}]$. Item bias can also be defined more stringently as tests where the probability of achieving a specific critical item score x_P or higher, given a specific standing on the latent variable being measured ξ_c $P[X \geq x_c|\xi_c]$, differs across gender groups [i.e. $E[X \geq x_c |\xi_c; Group_{female}] \neq E[X \geq x_c |\xi_c; Group_{male}]$. In terms of the stringent definition item bias will occur if the regression of the item response on the latent dimension being measured differs in terms of slope, and/or intercept and/or error variance. In terms of the more lenient definition item bias will occur if the regression of the item response on the latent dimension being measured differs only in terms of slope and/or intercept. The current study utilised the more stringent definition of item bias.

Measurement bias can be conceptualised and investigated from two perspectives. The discussion thus far approached the conceptualisation of measurement bias from the perspective of (classical and item response) measurement theory. The terms construct bias and uniform, non-uniform and error variance bias are typically used when approaching bias from a measurement theory perspective. It is however also possible to conceptualise and investigate measurement bias from the perspective of mean and covariance structure (MACS) analysis. The terms measurement invariance and measurement equivalence are typically used when approaching measurement bias from a MACS perspective. The terms measurement invariance and measurement equivalence are typically used interchangeably in literature. Theron (2016) concurs with Dunbar et al. (2011), in advocating a clear distinction between measurement invariance and measurement equivalence. These authors argue that two sets of questions emerge when differentiating between measurement invariance and measurement equivalence[4]. Both measurement invariance and equivalence pertains to the question whether the slope and/or intercept and/or error variance of the regression of the item responses on the latent personality dimensions being measured differ across groups. Measurement invariance in addition also pertains to the question whether a multigroup measurement model's factor structure (i.e. number of personality factors and the items' loading pattern on the factors) is identical across multiple groups. The criterion in terms of which the answers given in response to the questions posed in terms of measurement invariance are evaluated presents a more lenient evaluation of differences in the slope and/or intercept and/or error variance of the regression of the item responses on the latent personality dimensions being measured. The criterion in terms of which the answers given in response to the questions posed in terms of measurement equivalence are evaluated presents a more stringent evaluation of differences in the slope and/or intercept and/or error variance of the regression of the item responses on the latent personality dimensions being measured. Measurement invariance investigates whether a multigroup measurement model in which the factor structure (i.e. number of personality factors and the items' loading pattern on the factors) is constrained to be identical across multiple groups and in which (a) no parameters are constrained to be equal across the groups, (b) some parameters are constrained to be equal across the groups, fits the data obtained from two or more samples closely (Dunbar et al., 2011; Theron, 2016). Dunbar et al. (2011) proposed five hierarchical levels of measurement invariance which includes configural

[4] Chapter 3 will elaborate on the concept of measurement invariance and measurement equivalence.

invariance (measurement model structure is equal across groups), weak invariance (measurement model structure and slopes of the regression of items responses on the latent dimension being measured are equal across groups), strong invariance (measurement model structure, slopes and intercepts of the regression of items responses on the latent dimension being measured are equal across groups), strict invariance (measurement model structure, slopes, intercepts and error variances of the regression of items responses on the latent dimension being measured are equal across groups) and complete invariance[5]. Measurement equivalence investigates whether a multigroup measurement model in which the structure but no parameters is constrained to be equal across groups fits the data of multiple groups significantly better than a multigroup measurement model in which the structure and specific parameters are constrained to be equal across groups. Dunbar et al. (2011) also proposed four hierarchical levels of measurement equivalence, namely metric equivalence (the difference in fit between the configural invariance and weak invariance multigroup measurement models is not statistically or practically significant), scalar equivalence (the difference in fit between the configural invariance and strong invariance multigroup measurement models is not statistically or practically significant), conditional probability equivalence (the difference in fit between the configural invariance and strict invariance multigroup measurement models is not statistically or practically significant) and full equivalence[6]. Under the strict interpretation of measurement bias a finding of a lack of bias in the SAPI would be obtained when strict measurement invariance and conditional probability measurement equivalence are demonstrated (Meredith, 1993; Steenkamp & Baumgartner, 1998; Theron, 2016; Van de Vijver & Leung, 2001).

1.4. Gender Differences in Personality

Although often used interchangeably as synonyms in everyday language, the terms *sex* and *gender* represent different concepts in the social sciences. Personality differences between males and females have been widely investigated. Whereas a person's sex is a biological term that refers to the anatomy of the reproductive system, gender refers to the different societal roles with which the individual identifies (Wikipedia, 2018). It is acknowledged that people might identify with different societal roles (i.e. genders as in the latter definition from Wikipedia). However, the current study will refer to the sex differences (i.e. biological meaning of male and female) when referring the male, female and/or gender.

Two recent studies reported contradicting results regarding gender differences in personality. Whereas Samuel, South and Griffin (2015) found significant differences between males and females on the Big Five personality constructs, Zell, Krizan and Teeter (2015, p. 10) in turn reported "compelling support for the gender similarities hypothesis". Although there are contradicting research results, the majority of studies indicate gender differences in personality, albeit with small to very small effect sizes (Costa, Terracciano, & McCrae, 2001; Eagly, Johannesen-Schmidt, & Van Engen, 2003; Samuel et al., 2015; Zell et al., 2015). In addition to the effect sizes, the magnitude of personality differences between gender groups also differs

[5] Complete invariance is not really of interest to a measurement bias study since the differences in the covariances between the latent dimensions of the construct being measured does not impact either the lenient or stringent definition of (item) bias. It could possibly be argued that differences in the covariances between the latent dimensions of the construct being measured holds implications for construct bias given that the connotative meaning of the construct lies in the internal structure of the construct. The internal structure attributed to a construct is not fully explicated by simply specifying the number and identity of the latent dimensions. The nature of the correlational (and structural) relations that are thought to exist between the latent dimensions given the conceptualisation should also be specified. This line of reasoning moreover points to the need to structural invariance and equivalence analyses as part of the evaluation of construct bias.

[6] Full equivalence is not really of interest to a measurement bias study for the same reasons argued under footnote 5.

across culture groups and across personality domains (Costa et al., 2001; Zell et al., 2015). For instance, Zell et al. (2015) reported the counter intuitive finding that gender differences in personality are most pronounced in cultures where the traditional gender roles are minimised.

In addition to Samuel et al.'s (2015) report, other studies have also reported significant differences between genders on the Big Five personality. Women reportedly scored higher on the Openness to feelings and aesthetics facets than men, whereas men generally score slightly higher than women on Openness to Experience, Modestly higher, Openness to Ideas and Values facets (Costa et al., 2001; Samuel et al., 2015). Women score higher on than males Extraversion, as well as on its facet Warmth, where males in turn are higher on the facet Excitement seeking (d = -.10) (Costa et al., 2001; Feingold, 1994; Samuel et al., 2015). Women also score higher than their male counterparts on Agreeableness as a higher-order construct, but males in turn score higher on the lower-order traits at facet level for Modesty (d = -.02) and Assertiveness (Costa et al., 2001; Feingold, 1994; Samuel et al., 2015). In addition, Agreeableness and Pleasantness significantly (negatively) predicted Counterproductive Workplace behaviour at individual level only for males, whereas Emotional Stability in turn significantly (negatively) predicted Counterproductive Workplace behaviour at individual level only for females (Gonzalez-Mulé, DeGeest, Kiersch, & Mount, 2013). Women tend to score higher on Neuroticism than men, whilst men score higher on the facet Impulsiveness (d = -.18) (Costa et al., 2001; Feingold, 1994; Samuel et al., 2015). Longitudinal research conducted on narcissistic behaviour of an American student cohort revealed that men score higher than women on the Exploitative/Entitlement (d = .04), Leadership/Authority (d = .20) and Grandiose/Exhibitionism (d = .04) Narcissism facets (Grijalva et al., 2015).

It is important to note that of all the studies reported in the aforementioned section, Grijalva et al. (2015) were the only authors who investigated (or rather reported) the possibility of measurement bias and reported that no evidence for measurement bias was found. As a result, they concluded that the reported gender differences should be considered as real group differences. Given the contradicting research results in literature regarding gender differences, and the possible negative impact that psychometrically questionable personality assessments used in employee selection could have on candidates, test developers are obliged to empirically test the permissibility of their claim that observed scores for personality assessments have the same meaning in terms of the underlying latent variables for males and females. Hence, test developers are required to investigate personality assessment's measurement invariance and measurement equivalence, to confirm that the instrument is not biased against any gender group.

In addition to the question whether gender differences exist in the mean standing on specific first- and second-order personality dimensions further pertinent questions to measurement bias in personality assessment are (a) whether the behavioural denotations of specific first- and second-order personality dimensions differ across gender groups and (b) whether genders differ systematically in characteristics that can affect the manner in which they respond to test items that are not related to their standing on the first- or second-order personality dimension? Latent variables like (*inter alia*) language proficiency, stimulus familiarity, performance motive, educational level social desirability proneness and response style (like an acquiescence response style or an extreme response style) could affect the response option chosen even when controlling for the personality dimension measured by an item. If the behavioural denotations of

specific first-and/or second-order personality dimensions differ across genders a given personality inventory could suffer from construct bias as the loading pattern of items on the personality dimensions comprising the personality construct in terms of the constitutive definition could differ and even the number of factors required to satisfactorily account for the observed inter-item covariance matrix. The current study regards gender differences in the behavioural denotations of personality dimensions to be unlikely. The current study regards such differences as more likely across cultures. Gender differences in latent variables like language proficiency, stimulus familiarity, performance motive, educational level social desirability proneness and response style could affect the intercept, slope and error variances of the regression of item responses on the latent personality dimension the items were designated to reflect depending on whether they act as main effect or in interaction with the latent personality dimension.

1.5. South African Personality Inventory

To address the need to conceptualise an indigenous personality construct and to develop a personality instrument that would provide reliable, construct valid and unbiased measures of such a construct across the 11 language groups in South Africa that would make it appropriate to use in the complex South African context, researchers from South Africa and the Netherlands initiated a project to develop the South African Personality Inventory (SAPI)[7]. The research team set out to develop a personality assessment that is not biased, and that answers theoretical questions of indigenous personality in South Africa (Meiring et al., 2005; Nel et al., 2012; Valchev et al., 2011). The SAPI attaches a specific connotative definition of the personality latent variable, with specific latent dimensions conceptualised that are elicited with specific items. The measurement model implied by the designed intention of the test developers and reflected in the SAPI scoring key ensure that each personality dimension is measured in a true and uncontaminated manner (Holtzkamp, 2013)

Mouton (2017) reported close fit for the SAPI first-order measurement model, as well as completely standardised factor loadings above the critical cut-off value of .50. Despite some difficulties experienced in that study, the SAPI was able to discriminate successfully between the various latent personality dimensions' distinct aspects. Mouton was unable to converge the second-order measurement model. In finding close fit for the first-order measurement model, Mouton (2017) recommended that subsequent measurement and structural invariance and equivalence analyses be conducted on the SAPI.

1.6. Research Initiating Question

The current study is initiated by the research initiating question whether the construct-referenced inferences derived from the first-order personality dimension scores obtained on the SAPI are unbiased. More specifically the current study is initiated by the research initiating questions as to (a) whether the multigroup SAPI measurement model implied by the design intention of the test developers and their constitutive definition of the personality construct as reflected in the SAPI scoring key, fits the instrument data from male and female groups at least reasonably well, when a series of increasing constraints are imposed on the multigroup measurement model via a series of multigroup confirmatory factor analyses are conducted on the data and (b) whether the multigroup SAPI measurement models in which the model structure and specific

[7] Chapter 2 will elaborate on the rationale and development process for the SAPI.

parameters are constrained to be equal across genders fits significantly poorer than a multigroup SAPI measurement model in which only the structure is constrained to be equal across genders.

1.7. Research Objectives

The research objectives of this study are as follows.

Determine whether the SAPI demonstrates measurement invariance across male and female groups by investigating whether the multigroup measurement model with:

- Only the structure but no parameters constrained to be equal across the groups,
- The structure and the slope parameters constrained to be equal across the groups
- The structure, the slope and the intercept parameters constrained to be equal across the groups, and
- The structure, the slope, the intercept and the error variance parameters constrained to be equal across the groups

fits the data obtained from the male and female samples.

Determine whether the SAPI demonstrates measurement equivalence across male and female groups by investigating whether the multigroup measurement model in which only the structure but no parameters are constrained to be equal across the gender groups fits the data obtained from the two gender samples significantly poorer than:

- A multigroup measurement model with the structure and the slope parameters constrained to be equal across the groups.
- A multigroup measurement model with the structure, the slope and the intercept parameters constrained to be equal across the groups, and
- A multigroup measurement model with the structure, the slope, the intercept and the error variance parameters constrained to be equal across the groups.

1.8. Brief Chapter Overview

This study is organised in several chapters. Chapter 2 will provide an in-depth literature study into theories of personality, the application and implications of psychological testing in the South African context, as well as concluding with an overview on the development and reported psychometric properties of the SAPI. Chapter 3 elaborates on the differentiation between measurement bias, measurement invariance and measurement equivalence, as well as provide the taxonomy that was applied in this study for measurement invariance and measurement equivalence. Chapter 4 offers a detailed explanation regarding the research methodology that was applied in the study. Chapter 5 covers ethical issues that were considered in the study, and Chapter 6 elaborates on the research results. Chapter 7 concludes with a detailed discussion on the research findings, study limitations and recommendations for future research.

CHAPTER 2: LITERATURE REVIEW ON THE SAPI AGAINS THE BACKDROP OF PERSONALITY ASSESSMENT IN SOUTH AFRICA

2.1. Introduction

Chapter 2 will provide an in-depth discussion of personality assessment, and specifically personality assessment in South Africa, and against that backdrop, discuss why research into the psychometric properties of the SAPI is of such critical importance. The chapter will commence by defining personality and then explore various prominent personality theories. This is followed by an investigation into the South African context for personality assessment. The impact of cultural and gender diversity on personality assessments is examined closely, and several approaches on how personality can be researched cross-culturally is discussed. The different stages of the SAPI's development are elaborated on before the chapter will conclude with a summary of research findings across several studies that investigated the instrument's psychometric properties.

2.2. Theories of Personality

The underlying definition of personality that researchers hold will determine the selection of variables to study (Saucier, 2008). Some of the broader definitions of personality include Allport's (1963) view of personality as the "the dynamic organisation within the individual of those psychophysical systems that determine his characteristic behaviour and thought", and Mischel's (1976) definition of personality as "the distinctive patterns of behaviour (including thoughts and emotions) that characterise each individual's adaptation to the situations of his/her life". Pervin, Cervone, and Johan (2005) admittedly opted to define personality very broad, as "those characteristics of the person that account for consistent patterns of feeling, thinking, and behaving". Other authors hold a narrower stance on personality, like Cattell (1965) who defined personality as "that which people will do, think, or say when placed in a specific or given situation". Bergh (2016) states that despite the existence of several definitions of personality, there is some consensus among researchers that both person characteristics and situational factors should be included to adequately explain the impact of personality on behaviour.

Typically personality researchers have viewed personality as a set of stable, non-malleable characteristics that distinguish one individual from another. These characteristics are assumed to determine behaviour and because they are assumed to hold across time and place, the behaviour of a specific individual (with specific, stable personalities) is expected to be consistent across many different situations. An individual high on Introversion is expected to behave introvertantly consistently in all situations and an individual high on the Neuroticism dimension should act neurotically across a wide variety of situations. This assumption has, however, been difficult to prove empirically (Mischel, 2004). The finding that the same individual will show substantial variation as the situations vary, has since become widely accepted. Still controversial, however, is the question why behaviour varies across situations. This question is moreover of critical importance for the conceptualisation of personality. The conventional position is that situational characteristics exert a causal influence on behaviour but that they do so independent of personality traits. In terms of this line of reasoning situational latent variables then represent nuisance variables that need to be statistically controlled if the influence of personality on behaviour is to be clearly understood (Mischel, 2004). Mischel (1973; 2004)

differs from the conventional position and argues that intra-individual variability in behaviour across situations should form part of the conceptualisation of personality. In the attempt to conceptualise personality and understand how it affects behaviour, situational latent variables should not be regard as nuisance variables that obscure the influence of personality. Situational latent variables, Mischel (1973; 1977; 2004) argues, should rather be seen as necessary and indispensable components of personality theory. How situations are appraised depends on characteristics of the individual. More specifically, Mischel (1973; 1977; 2004) argues that individuals' subjective interpretation of the situation (rather than the objective features of the situation), along with individuals' personality, affect behaviour. Mischel (1973; 1977; 2004) therefore argues that intra-individual behavioural consistency will only occur across situations if the individual with relatively stable characteristic appraises the different situations similarly with regards to one or more subjective situational characteristics. This line of reasoning lead Mischel (2004) to conclude that to explain intra-individual variability in behaviour personality theory needs to make provision for more complex if … then situation-behaviour relationships. Mischel (2004, pp. 4-5) describes his position as follows:

> This approach outlined the underlying psychological processes that might lead people to interpret the meanings of situations in their characteristic ways, and that could link their resulting specific, distinctive patterns of behaviour to particular types of conditions and situations in potentially predictable ways. The focus thus shifted away from broad situation-free trait descriptors with adjectives (e.g., conscientious, sociable) to more situation-qualified characterizations of persons In contexts, making dispositions situationally hedged, conditional, and interactive with the situations in which they were expressed. A main message was then— as it still is 30 years later—that the term "personality psychology" need not be behaviour11 for the study of differences between individuals in their global trait descriptions on trait adjective ratings; it fits equally well for the study of the distinctiveness and stability that characterize the individual's social cognitive and emotional processes as they play out in the social world. In this social cognitive view of personality, if different situations acquire different meanings for the same individual, as they surely do, the kinds of appraisals, expectations and beliefs, affects, goals, and behavioural scripts that are likely to become activated in relation to particular situations will vary. Therefore, there is no theoretical reason to expect the individual to display similar behaviour in relation to different psychological situations unless they are functionally equivalent in meaning. On the contrary, adaptive behaviour should be enhanced by discriminative facility—the ability to make fine-grained distinctions among situations—and undermined by broad response tendencies insensitive to context and the different consequences produced by even subtle differences in behaviour when situations differ in their nuance (Cantor & Kihlstrom, 1987; Cheng, 2001, 2003; Chiu et al., 1995; Mendoza-Denton et al., 2001; Mischel, 1973). In short, the route to finding the invariance in personality requires taking account of the situation and its meaning for the individual, and may be seen in the stable interactions and interplay between them (e.g., Cervone & Shoda, 1999; Higgins 199;, Kunda, 1999; Magnusson & Endler, 1977; Mischel, 1973, Mischel & Shoda, 1995).

Although seemingly not explicitly suggested by Mischel (2004) by extending the foregoing line of reasoning, it could be argued that to explain inter-individual variability in behaviour, personality theory needs to make provision for even more complex if and if … then person-situation-behaviour relationships.

Mischel's (1973; 1977; 2004) argument should not be construed that personality assessment should be abandoned, although many seemed to have interpreted Mischel's position in this way (Mischel, 2013). Rather what Mischel (2013) has in mind is a "constructive reconceptualization of personality" that formally takes the appraisal of the situation into account. Rather than implying abandoning personality assessment, his position seems to imply the need for the assessment of the situation in terms of perceived "situational traits" as well.

Various definitions of personality are rooted in different underlying personality theory postulated by the respective authors. The following section will explore some of the prominent theories on personality, such as psychoanalytic theories, behaviourist or learning theories, humanistic and existential approaches, trait and type theories, and lastly cognitive and social-cognitive theories.

2.2.1. Psychoanalytical Theories

These theories propose that personality is constituted by unconscious forces and while people are mostly unaware of why they behave the way they do in situations, they nonetheless strive for an awareness of the reasons why they act a certain way (Bergh, 2014; Saucier, 2009). Sigmund Freud is widely regarded as the founder of psychodynamic/psychoanalytic theory, whilst other influential contributors to the theory include Adler, Jung, Sullivan and Western (Cloninger, 2009).

The underlying assumption of psychoanalytic theory postulates that personality differences occur in the manner of which three separate, but interdependent psychological forces work together. Freud named these forces the *id* (found in the unconscious and comprises of irrational impulses that are uncontrolled and strive to immediately gratify sexual, physical and emotional needs and through that attain pleasures irrespective of moral or social acceptability); the *superego* (the second part of personality which operates according to the morality principle: values and morals with regard to what is right and wrong) (Phares, 1984). The *superego* consists of two parts, (1) the conscience that uses guilt to punish what is wrong and (2) the ego ideal that is responsible for rewarding what is right and develops during childhood and through socialisation. These act as inhibitors as opposed to oppressors of the pleasure-seeking demands of the *id*. The psychological force that forms the last part of personality as proposed by psychoanalytical theory is the *ego*; the *ego* acts as the balancing agent between the *id* and the *superego*. Thus the *ego* chooses the best manner to gratify the *id*'s needs whilst being socially acceptable and limiting undesirable consequences. However, the bigger the conflict between the impulse and what is morally right, the more difficult this becomes. From this interplay, defence mechanisms are born to maintain a positive self-image whilst solving these unconscious conflicts by satisfying id urges in a manner acceptable to the superego (Phares, 1984; Anderson & Lewis, 1998)

Emphasis is placed on the conflict people experience due to the norms of society, internal biological drivers, past events and unconscious motives. Translated to the work context these theories suggest that people's performance differ from one another as a result of the interaction between unconscious forces (Albertyn, 2003).

Psychodynamic theory assumes that the most important part of personality development occur during early childhood. They further believe that any problems in early life can potentially create disruptive influences

later in (adult) life (Bergh, 2014; Phares, 1984). The theory is sometimes criticised for poor testability of the concepts and the consequential lack of research evidence, as well as the emphasis on sexist ideas (Bergh, 2014).

2.2.2. Behavioural Theories

Behaviourist theorists argue that personality is influenced by the environment and the circumstances that people find themselves in, instead of unconscious forces (Bergh, 2014). According to this theory, personality is characterised by expectations, thoughts and observable behaviour that is continually learned and rewarded to varying degrees in the various environments and circumstances people find themselves (Bergh, 2014). Due to continuous learning throughout the individual's life, personality is regarded as dynamic across time and situations (Bergh, 2014). Authoritative researchers on this theory, such as Michel, Bandura and Skinner, ascribed the individual differences between people as dependent on their environmental influences and information that they previously learned (Bergh, 2014). Later behavioural theorists emphasise self-regulation, in that people can learn through rational thinking. Translated to a work context these theories suggest that people's performance can be influenced through training and motivating employees (Bergh, 2014).

2.2.3. Humanistic, Phenomenological and Existential Theories

Humanistic, phenomenological and existential theories view personality as people's unique qualities, such as their subjective and unique experience of reality, and while striving to find meaning in life (Bergh, 2014). In contrast to the previous two theories the person is now regarded as a free and rational being, and not controlled by unconscious forces or the environment (Weiten, 2011). Influential authors on these theories such as Seligman, Csikszentmihalyi, Rogers, Maslow and Allport, argue that personality development takes place throughout the individual's life, and individual differences can therefore be ascribed to people's unique experiences (Bergh, 2014; Cloninger, 2009). These theories can be applied in the work context, through counselling, positive psychology, and management approaches (Bergh, 2014). However, the lack of empirical support, the lack of clarity on some of the concepts and an overly optimistic view of human nature are some of the critique against these theories (Bergh, 2014).

2.2.4. Cognitive and social-cognitive theories

Bandura and Mischel were important authors contributing to the cognitive and social-cognitive theories that regard personality and behaviour as being shaped by the consequences of learning (Bergh, 2014). These theories emphasise ways that people apply to understand and control the world, as well as their own and others' behaviour (Bergh, 2014). Examples of such manners include self-regulation, self-efficacy, perception, memory, and cognitive schemas and processes (Bergh, 2014). Therefore, from the stance of these theories, personality develops according to the interaction between the environment, situations and the person's self-created cognitive constructs (Bergh, 2014).

In contrast to other personality theories, cognitive and social-cognitive theorists refrain from generalising behaviour patterns, and instead assume that behaviour is unique due to specific psychologically significant situations that have different influences on individuals (Bergh, 2014). These theories posit that individual differences are caused by the unique combination of different constructs that each person has, in contrast to

the other personality theories proposing that people have different levels of the same constructs (Bergh, 2014).

2.2.5. Trait Theories

Trait theorists view personality as "characterised by distinguishable, enduring and consistent attributes and patterns of behaviour" that can be explained through concepts such as traits, factors, dispositions and dimensions (Bergh, 2014, p. 296). Eysenck proposed that traits should be derived through psychometric evidence (e.g. factor analysis); must have an underlying biological foundation; be based on a strong theoretical argument; as well as possess social relevance (e.g. optimising person-job-fit) (Pervin, Cervone, & John, 2005). These propositions lead to several assumptions on which the trait theory is based.

Trait theorists assume that there exists an interaction between the individual and the environment, but that inherited biological factors have the dominant influence on the demonstration of traits (Bergh, 2014). McCrae and Costa (2010) argue that personality should also be universal across all cultures, based on the premises that personality is a function of biology and that all human beings share a common (i.e. universal) genome. Five Factor Model (FFM) theorists therefore emphasise the universality and stability of these traits in individuals and groups, across time and situations (Bergh, 2014). McCrae and Costa maintain that although unique personality factors might exist in certain cultures, one might be able to categorise those factors within the FFM. The trait approach therefore allows practitioners to measure, summarise and compare human behaviour, with minimal descriptions and several taxonomies. The added benefit of traits being measurable with psychometric assessments is that future behaviour can be predicted in the workplace and other contexts, to the extent that the criterion construct to be predicted is at least to some degree systematically related to personality. As a result trait theory personality assessments, such as the SAPI, are often used in personnel selection, career counselling, personal development programmes, and team development (Foxcroft & Roodt, 2010).

Although the trait theory, and in particular the FFM, dominates the field of psychology (Laher, 2008), it is not without critique. Firstly, traits are measurable with psychometric instruments, providing empirical evidence that questions the claim for universality (Bergh, 2014). Contrary to Agreeableness, Extraversion and Conscientiousness, research findings for Neuroticism and Openness to Experience do not replicate consistently across studies (Saucier, as cited in Pervin et al., 2005). Pervin et al. (2005) suggest that these inconsistencies in findings for Neuroticism might be due to cultural variations in perceptions of negative emotions in interpersonal settings, whereas the inconsistent results for Openness to Experience are ascribed to a lack of consensus in literature regarding the construct definition which tends to include both or either intellectual and cultural openness. Furthermore, cross-cultural and age group studies on traits do not always support the notion of trait consistency either (Bergh, 2014; Heaven & Pretorius, 1998; Visser & du Toit, 2004), which lead McCrae and Costa (2010, p. 167) to concede that traits should rather be considered as "relatively stable".

Secondly, the FFM specified certain personality structures in the theoretical framework, but do not explain how the behaviour is caused by the postulated biological and psychological mechanisms/processes (known as "dynamic processes") (Pervin et al., 2005, p. 265). McCrae and Costa (2010) maintain that personality

traits originate from biological functions in the human body, but research findings only support this claim for two of the FFM traits, namely Extraversion and Neuroticism (Jaušovec & Jaušovec, 2007; Pervin et al., 2005).

Thirdly, FFM theorists assume that results from one level of analyses can automatically be applied to other levels of analysis. The FFM is supported by statistical analyses of populations of people. Pervin and colleagues (2005b) emphasise the benefit that such analyses bring is that inividual differences in the larger population can be summarised. They argue that the concern with the differences in the levels of analyses is that conclusions from analyses on population level of analyses does not demonstrate that each individual in the population has all of the five factors. Borsmboom, Mellenbergh and Van Heerden (2003) corroboborate this line of thining:

> "Finally, in a standard measurement model, the causal ingredient of realism can be defended in a between-subjects sense but not in a within-subject sense. The within-subjects causal interpretation may be viewed as a fallacious application of between-subjects results to individuals. To substantiate causal conclusions at the level of the individual, one must investigate patterns of covariation at the individual level, that is, one must fit within-subject latent variable models to repeated measurements in the sense of Cattell and Cross (1952) and Molenaar (1985). On the basis of this line of thinking, the possible relations between within-subjects models and between-subjects models were used as the foundation for a classification of psychological constructs as locally homogeneous, locally heterogeneous, and locally irrelevant. The main implication of this analysis for psychological research is as simple as it is instructive: If one wants to know what happens in a person, one must study that person. This requires representing individual processes where they belong, namely at the level of the individual. On the other hand, if the study of the individual is dismissed as too difficult, too labor intensive, or simply as irrelevant, one cannot expect between subjects analyses to miraculously yield information at this level."

Lastly, trait theory oversimplifies personality by minimising the number of traits. Douglas and Martinko (2001) reported that 62% of variance in self-reported aggression at the workplace was explained by several constructs that were not included in the FFM.

Despite these shortcomings to the stated trait theory assumptions, organisations use personality assessments that are psychometrically sound to select the best candidate for jobs by predicting future behaviour on the job. Ample empirical evidence in literature is available to prove the link between personality traits and job performance. For instance, several meta-analyses have reported that Conscientiousness and Emotional Stability validly predict overall performance to the extent that it explained variance for the motivational component of performance, whereas general mental ability has been shown to affect performance through "can do" capabilities (Barrick & Mount, 2005; Schmidt & Hunter, 1998). In their meta-analysis, Schmidt and Hunter (1998) reported that general mental ability explained 31% of variance in overall job performance, whereas integrity and Conscientiousness respectively explained 14% and 9% of variance in overall job performance. Hence, personality is an important, but not the only predictor that explains variance in behaviour and job performance.

The following section will underpin the rationale as to why the empirical basis (i.e. psychometric soundness) for personality assessments is so critical.

2.3. Psychological Assessment in South Africa

The segregation laws in South Africa prior to 1994 resulted in separate psychological tests being developed for different language and race groups. Owen (1991) points out that the majority of psychological tests were developed for the White racial group. This caused a dire shortage of psychological assessments that were applicable to Black, Coloured and Indian racial groups, and psychologists subsequently opted to use Westernised assessments for the other racial groups despite the fact that those assessments were only standardised for White test takers (Lubbe, 2012).

During the late 1980's, advances in the South socio-political circumstances resulted in a call for so-called culturally fair psychological assessments (Meiring, Van de Vijver, & Rothmann, 2006). Due to the shortage of skilled psychological assessment developers in South Africa, psychologists adapted Westernised psychological tests such as the Fifteen Factor Questionnaire Second Edition (15FQ+) and the Sixteen Personality Factor Questionnaire (16PF), rather than develop new assessments (Foxcroft & Roodt, 2010). Foxcroft (2004) points out that concerns were raised about measurement bias, inequivalence and cultural relevance caused by such test adaptation. For example several studies on the 15FQ+ in the South African context reported low internal consistencies of some of the subscales (i.e. thirteen of the fifteen subscales showed coefficient alpha below the generally accepted Cronbach alpha of .70), as well as the instrument lacking construct equivalence (Meiring et al., 2006; Moyo & Theron, 2011). Abrahams and Mauer (1999) evaluated the 16PF and reported that thirteen of the sixteen factors demonstrated alpha coefficients below .50 for the Black population, and proposed that problematic language proficiency was a probable cause for differences in reliability across the different cultures. These two personality assessments remained widely used despite being deemed as inappropriate for sections of the South African population with inadequate English language proficiency[8] (Abrahams & Mauer, 1999; Meiring et al., 2006; Moyo & Theron, 2011).

The socio-political changes further resulted in the implementation of post-apartheid legislation, and in particular the Employment Equity Act (EEA) (Republic of South Africa (RSA), 1998), which prohibits the use of psychological tests in the workplace unless the test is (a) shown to be valid and reliable; (b) can be applied in a fair manner to all employees; and (c) is not biased against any individual or group. Furthermore, the EEA (RSA, 1998, p.15) specifically prohibits any form of unfair discrimination:

> No person may unfairly discriminate, directly or indirectly against an employee, in any employment policy or practice, on one or more grounds, including race, gender, sex, pregnancy, marital status, family responsibility, ethnic or social origin, colour, sexual orientation, age, disability, religion, HIV status, conscience, belief, political opinion, culture, language and birth.

[8] One should, however, be careful to criticise specific tests if their measures turn out to display low reliability for subpopulations with inadequate English language proficiency that prevents them from properly understanding the test items. If test items are not understood test-takers' response to them become essentially random. The test developers, however, have developed and standardised the test for a specific target population that, if not explicitly stated as such, at least implicitly restricts the use of the test to a population that has the necessary language proficiency to understand the instructions and items. What should be criticised is the test user's decision when deciding to use the test on individuals that do not meet the criteria in terms of which the target population was defined.

> It is not unfair (*direct*) discrimination to take affirmative action measures consistent with the purpose of this Act; or distinguish, exclude or prefer any person on the basis of an inherent requirement of a job (bracketed and italicise text added).

The EEA poses a challenge for organisations in their endeavour to apply psychological assessments in a fair manner, since South Africa consists of groups with different cultural, educational and socio-economic backgrounds (Ramsay et al., as cited in Cohen, 2013; Hill, et al., 2013). Consequently, for organisations to ensure that they comply with the EEA demands despite these stated challenges, the EEA requires organisations to provide evidence that the psychological assessments that they apply are appropriate for the context, i.e. shown to be valid, reliable, unbiased to any employee, and applied fairly within their organisation. The onus therefore rests on test developers to provide test practitioners with psychometric evidence for the personality assessments marketed.

The EEA was promulgated to ensure fair direct and indirect discrimination amongst groups in selection for employment and into development opportunities and to ensure equitable (i.e. proportional representation in) employment and development opportunities. The manner in which the EEA is currently furmulated essentially argues that the extent to which both these objectives will be achieved depends on the psychometric property of the psychological test (or other similar assessment prosedure) used to inform selection decisions. The psychological test may not be used to inform the decision unless it can be shown that the test is valid, reliable, unbiased and that it can be *applied* fairly. This formulation is unfortunate in that the test *per se* is neither valid nor reliable nor biased and in that it is not so much the *application* that should be fair (i.e. consistent) but rather the criterion inferences derived from the test scores. Validity refers to the permissibleness of the inferences that are derived from the scores obtained on the test. Two types of inferences can be derived from test scores. Inferences on the construct being measured and inferences on a (criterion) construct that is not measured but that is systematically related to the construct being measured. Which inference does the EEA have in mind? Seemingly the first? The measures of a test displays the characteristic of reliability whereas bias again refers to a characteristic of the inferences derived from the test scores. Bias in the broadest sense refers to the presence of systematic group-related error in inferences. Taking the previous distinction between the two types of inferences into account this can then refer to either measurement bias (systematic group-related error in the construct-referenced inferences derived from test scores) or predictive bias (systematic group-related error in the criterion-referenced inferences derived from test scores). The requirement that psychological tests may only be used if they are *applied* fairly in essence requires that tests should be standardised. More often than not the EEA is interpreted to mean a psychological test may not be used to inform selection decisions if it cannot be shown that the inferences derived on the to-be-measured construct are (construct) valid, the observed scores are reliable and the construct-referenced inferences are not biased (i.e the construct-referenced inferences do not suffer from construct bias, non-uniform measurement bias, uniform measurement bias or error variance measurement bias). By implication, conversely, a psychological test may be used to inform selection decisions if it can be shown that the inferences derived on the to-be-measured construct are (construct) valid, the observed scores are reliable and the construct-referenced inferences are not biased. This conclusion is, however, false. The EEA prohibits of the use of psychological tests that do not meet these criteria because it believes the extent to which the two objectives of the EEA will be achieved in selection depends on the psychometric

integrity of the tests used to inform the selection decisions. Although the EEA is correct that the use of selection instruments that do not meet these criteria could indirectly unfairly discriminate against members of a specific group (especially in the criterion inferences are derived clinically), it is not correct in its implied conclusion that meeting these criteria will ensure that no member of any constitusionally protected group will not indirectly be unfairly discriminated against. Selection decisions are based on criterion inferences derived clinically or mechanically from test scores. Distinguishing, excluding or prefering any person for a job or development opportunity (Republic of South Africa (RSA), 1998) based on criterion inferences that contain systematic group-related error (i.e. criterion inferences that are predictively biased) will indirectly unfairly disadvantage those individuals for whom the criterion performance is systematically underestimates. The systematic underestimation occurs because the inferences are derived as if the nature of the relationship between the criterion and the predictors is the same across groups when in fact it differs in terms of intercept, slope parameters and/or error variance. The critical point to appreciate is that this can occur even when the predictor provides reliable, construct valid and unbiased measures of the predictor construct. It can also happen when the criterion inferences display predictive validity (although the predictive validity will be lower than it would have been if the predictive bias would have been corrected). It is therefore not possible for the test user to immunise him-/herself against the danger of unfair indirect discrimination by being psychometrically judicious about the nature of the psychological tests that are used as selection instruments. Moreover, it implies naive psychometric thinking if studies aimed at evaluating the reliability, construct validity or measurement bias are motivated in terms of paragraph 8 of the EEA (RSA, 1998).

2.3.1. Culture and personality

Mouton (2017) proposed two aspects to consider when exploring the relevance of culture to personality assessment, namely the manner in which personality is conceptualised and operationalised.

In the first instance, many studies offer empirical support that traits differ cross-culturally depending on the conceptualisation of personality (Marsella, Dubanoski, Hamada, & Morse, 2000; Valchev, Van de Vijver, Nel, Rothmann, & Meiring, 2013). Babbie and Mouton (2001) state that both scientists and the general public (i.e. indigenous cultures) attempt to develop and conceptualise constructs, to communicate their experience and understanding of phenomena in World 1. Their conceptualisations of personality can therefore differ in their connotative meaning. Mouton (2017) argued that differences in the connotative meaning of constructs arise across cultures because of differences in the behavioural events and phenomena that people in different cultures attempt to make sense of and need to explain. Whether or not the industrial psychology fraternity should attempt to incorporate constructs developed by the general public in a specific culture, the question should be asked whether those particular constructs are relevant to the work behaviour that the industrial psychologist intends to explain (Mouton, 2017). Mouton (2017) expressed her concern with the SAPI developers' motivation to include constructs from the indigenous South African public that were previously ignored by scientists, for the sake of providing a comprehensive personality instrument to the workplace (Fetvadjiev, Meiring, Van de Vijver, Nel, & Hill, 2015). She warned that the extension of the traditional Western conceptualisation of personality with personality factors that are currently not reflected in the traditional model will not, merely by being more comprehensive, be able to better explain variance in work performance (Mouton, 2017).

The second aspect highlighted by Mouton (2017) is the manner in which personality constructs are operationalised. She states that the use of self-report questionnaires with sets of stimuli that elicit the recollection of observable behavioural manifestations, which reflects the individual's standing on a particular personality construct, is the most common method for indirectly measuring personality. Despite the assumed universality of the personality structure by trait theory scholars (Beery, Poortinga, Segall, & Dasen, 2002). Mouton (2017) argues that the manner in which traits are expressed in behavioural denotations may differ cross-culturally. When such cultural differences occur between groups, the regression of personality items on the latent personality dimension which the items are intended to reflect might potentially differ between the groups based on the slope, intercept, and/or error variances (Mouton, 2017). This is problematic for industrial psychologists, who are mandated by the EEA to ensure that personality assessments are shown to be unbiased against any group. It is therefore imperative that the industrial psychology fraternity increase its emphasis on measurement invariance and equivalence, so as to identify whether the response of people from different cultures completing a questionnaire are determined by the (same) personality construct of interest (i.e. the absence of construct bias), and if so, whether the nature of the regression of item responses on the latent dimensions of the personality construct are the same across cultural groups in terms of intercept, slope and error variance (i.e. the absence of item bias). Mouton (2017) argues that as long as the items of the questionnaire remain valid reflections of the latent personality dimensions that they were designated to reflect, the problem caused by cultural differences in the manner in which personality traits are expressed in item responses can potentially still be managed without developing equivalent forms of the same test. A possibility in this regard is to use the single-group measurement model parameters of each group to derive latent score estimates. A challenge that still remains here is to ensure that the latent score estimates occur on the same scale and therefore are comparable. However, when the cultural differences become more extreme and what is seen as a behavioural denotation of one personality dimension in one culture is seen as an expression of another trait in another culture, then equivalent tests may become unavoidable.

2.3.2. Approaches to study culture and personality

Three different approaches are used to study personality within and among different cultures. The first is the etic approach, which relies on the assumption that traits are cross-culturally stable (Nel et al., 2012). Etic studies use existing personality inventories that were developed in a culture (e.g. the Five Factor Model in the Western culture) and apply the inventories to other cultures. Although the etic approach assists in identifying commonalities between different cultures, it can unfortunately also cause unique aspects of that are culturally specific, to be underrepresented or even missed (Mouton, 2017; Nel et al., 2012). Evidence for this shortcoming was demonstrated in cross-cultural research indicating that personality assessments developed for the Western culture did not successfully capture non-western cultures' constructs (Mouton, 2017). Hence, the second approach known as the emic approach was developed to address this issue.

The emic approach explores traits in a specific cultural context to ensure that the measurement is appropriate for that culture (e.g. Chinese Personality Assessment Inventory was specifically developed for the Chinese culture) (Cheung et al., 2001; Nel et al., 2012). Some authors list the lack of standardisation on representative norm groups, theoretical challenges and the difficulty to sustain the thorough research

programmes that are necessary to develop valid and reliable instruments, as shortcomings of the emic approach (Cheung, Cheung, Wada, & Zhang, 2003; Mouton, 2017). Nel and colleagues (2012) in turn argue that the benefits and shortcomings of the emic approach are the opposite to that of the etic approach, which lead to the third approach to studying personality across cultures.

The combined etic-emic approach stems from the complementary benefits of both previous approaches. Initial movements toward indigenisation attempted to create context specific and non-western methodologies, constructs and strategies, but failed to integrate their perspectives on personality with that of (assumed) human universals in personality (Mouton, 2017). The combined etic-emic approach offers researchers the opportunity to attain integration, synergy, and balance between both cultural specific and universal aspects of personality (Cheung, Van de Vijver, & Leong, 2011; Morris, Leung, Ames, & Lickel, 1999).

2.3.3. Gender differences in personality assessment

Psychological assessments can only be construct valid and reliable for the groups for which the assessments are developed, validated and standardised (Bedell, Van Eeden, & Van Staden, 1999; Foxcroft & Roodt, 2010; Ramsay, Taylor, De Bruin, & Meiring, 2008; Van de Vijver & Rothmann, 2004). Demonstrating the construct validity of the construct referenced inferences derived from the scores obtained on an instrument for a homogenous group or for a heterogeneous group (by for example demonstrating satisfactory single-group measurement model fit and parameter estimates), raises the question whether the construct validity holds across different (gender, race, language, age, cultural) groups (i.e. whether the measurement model fit holds), and if so, whether the observed scores on the instrument can be interpreted (construct referenced) in the same way across groups (i.e. whether the measurement model parameter estimates hold across groups). Only if the absence of construct bias, strong invariance and scalar equivalence have been shown can observed scores be descriptively compared across groups (Dunbar et al., 2011). Several authors have called for more research on psychological assessment measurement invariance and equivalence for all subgroups listed by the EEA. The measurement invariance and equivalence studies on personality assessments in South Africa tend to focus on investigating differences between races and language groups (Bester, 2008; Chrystal, 2012; Cohen, 2013; Holtzkamp, 2013; Horak, 2012; Kemp, 2013).

The EEA (RSA, 1998) specifies gender as one of the grounds for which unfair discrimination is prohibited. According to research conducted by the World Economic Forum (World Economic Forum, 2016) 42% of respondents globally indicated that gender parity should be enhanced to promote fairness and equality in the workplace. Personality assessments are often part of organisations' selection batteries. A call for invariance and equivalence studies focussing on gender can, however, not be convincingly motivated in terms of the EEA's prohibition of unfair indirect discrimination. Although it is true that a measure contaminated by biased items can systematically underestimate the standing of members of a specific group on specific latent dimensions of a construct and that such bias in construct-referenced inferences can cause predictive bias in criterion-referenced inferences, especially when such inferences are derived clinically, the latter need not unavoidably occur when the former occurs. Measurement bias can, but does not have to, result in predictive bias and unfair indirect discrimination if criterion inferences are derived mechanically via actuarial prediction models. More importantly, however, is that the absence of measurement bias cannot guarantee the absence

of predictive bias in the criterion inferences derived from the (unbiased) predictor and therefore cannot guarantee fair (indirect) discrimination. It is thereby, however, not implied that measurement bias in the predictor should be ignored or condoned. Ensuring a lack of construct bias and a lack of non-uniform, uniform and error variance bias remains important in the interest of proud psychometric workmanship and remains indispensable when measuring the standing of various groups on the construct and its latent dimensions. Hence, assessment practitioners need evidence as to whether or not the applied personality assessment is equivalent across genders. Most bias studies in South Africa context focus on race and language, resulting in scant research that provide such evidence for gender groups in the South African context.

Literature on gender differences seem to be more available in the international front. For instance Vianello, Schnabel, Sriram, and Nosek (2013) reported gender differences between implicit and explicit expressions of the FFM constructs among participants from Western countries. It should be noted that Vianello and colleagues could not substantiate why those differences occurred. In another study Costa, Terracciano and McCrae (2001) reported that gender differences are most prominent among European and American cultures and most attenuated among African and Asian cultures. Costa and colleagues ascribe this finding to the reported gender differences that were also associated with levels of Individualism ($r = .71$, $n = 23$, $p < .01$), as Western individualistic values resulted in higher differences between genders on self-reported traits such as assertiveness, than non-Western collectivistic values. The conclusion that gender groups differ on specific personality dimensions are valid (i.e. permissible) only if prior evidence of lack of construct bias and (non-uniform and uniform) item bias has been shown.

The culturally diverse South African population also consist of both collectivistic and individualistic cultures. It is therefore imperative to investigate not only whether the SAPI is biased against any race (culture), but also against any gender group. To date no gender-based measurement invariance and equivalence analyses on the SAPI have been reported. To substantiate the notion that the SAPI is appropriate for use in both genders, and that the SAPI dimension scores can be interpreted construct-referenced in the same manner across groups, the SAPI developers have to empirically evaluate bias for these groups. Only if a lack of construct bias, as well as a lack of non-uniform, uniform and error variance item bias has been shown, can observed scores be compared across groups and can differences be interpreted in terms of differences in the underlying personality construct. Therefore, only if a lack of construct bias, non-uniform, uniform and error variance item bias has been shown can the development of separate construct-referenced norms for each gender be justified if systematic observed score gender differences would be found. The present study will contribute to this area of research on the SAPI. The ideal would however be to combine both culture (i.e. race) and gender in the research initiating question, but that extended set of criteria is not part of the scope for the current study.

2.4. Overview of the SAPI

The following section will provide a brief overview of the SAPI. The development of the SAPI is elaborated on, specifying the steps taken during the qualitative and quantitative phases. Thereafter, the psychometric properties for the SAPI reported by various studies are discussed in detail. Since this study follows on

Mouton's (2017) research which investigated the SAPI factor structure on the same dataset provided by the SAPI developers, specific emphasis is placed on the psychometric properties that she reported.

2.4.1. Development of SAPI

In 2005 several researchers[9] from South Africa and the Netherlands set out to address the shortage of personality assessments that are appropriate for use on all racial groups (i.e. White, Black, Coloured, Indian, etc.) in South Africa, through a project known as the SAPI, which is an acronym for South African Personality Inventory (Hill, Nel, Vijver, & Meiring, 2013). The project consisted of both qualitative and quantitative phases.

The qualitative phase commenced with the project purpose in mind, namely to develop a personality measure that would be applicable to each of the 11 official languages in the South African population (Hill, et al., 2013). Acknowledging the limitations of the etic and emic approaches, the SAPI developers opted for a combined etic-emic approach to studying personality across different the cultures. The developers also applied the popular psycho-lexical method, which is according to Allport and Odbert (1936) based on the view that individual differences in psychological functioning is grounded in language. This method uses personality descriptions from dictionaries, along with analyses of oneself and other people on those descriptions that result in data that is subsequently subjected to factor analyses (De Raad, et al., 2010; Nel et al., 2012). Since appropriate dictionaries for all of the 11 languages in South Africa were unavailable, the developers decided to adapt the method slightly and instead used free descriptions of personality since English translations were readily available (Mouton, 2017; Nel et al., 2012).

Semi-structured interviews were held with a total of 1217 representatives from each of the respective language groups, who had to describe themselves and at least nine other people in their indigenous language (Nel et al., 2012). The translated interviews resulted in 49818 personality descriptions that were then reduced through several iterative content analyses into 550 sub-facets, then again decreased to 188 narrow facets, followed by 37 midlevel sub-clusters and eventually reduced to only 9 broad representative personality clusters (Fetvadjiev et al., 2015; Mouton, 2017; Nel et al., 2012). These nine clusters were labelled as Conscientiousness, Emotional Stability, Extraversion, Facilitating, Integrity, Intellect, Openness, Relationship-Harmony, and lastly Soft-heartedness (Nel et al., 2012).

The quantitative stage was initiated with the generation and selection of items. A total of 2574 items were generated in English, to reflect the qualitative model and the personality descriptors obtained during the qualitative phase (Fetvadjiev et al., 2015). Fetvadjiev et al. (2015) explained that English was the chosen language because it is the language commonly spoken and understood in the different ethnocultural groups in South Africa; was the common language among the research team; and finally, the research team deemed English to provide the richest vocabulary of personality descriptions among the 11 languages.

[9] The participants in the project are Byron Adams (University of Johannesburg and Tilburg University, The Netherlands), Deon de Bruin (University of Johannesburg at that stage, now University of Stellenbosch), Karina de Bruin (University of the Free State), Carin Hill (University of Johannesburg), Leon Jackson (North-West University), Deon Meiring (University of Pretoria and University of Stellenbosch), Jan Alewyn Nel (North-West University), Ian Rothmann (North-West University), Michael Temane (North-West University), Velichko Valchev (Tilburg University, The Netherlands), and Fons van de Vijver (North-West University and Tilburg University, The Netherlands).

Items were selected based on factor analyses results of data from pilot studies (Fetvadjiev et al., 2015). Fetvadjiev et al. (2015) listed the following psychometric and substantive criteria that were used to iteratively remove items:

- Items with extreme mean values;
- Skewness or kurtosis;
- Items with factor loadings lower than .30 on either or both of the higher and lower level factors, with at least .40 as a cut-off for items that did not achieve the desired .30 factor loadings;
- Items were retained that represented the construct maximally, minimised content overlap between and within the clusters, as well as were in line with the item generation rules of behaviour focus, namely using simple language and being translatable;
- And lastly the measurement model had to be replicable across ethnic groups to minimise the presence of idiosyncratic features and increase the possibility of replicating the factors in future.

Based on the stated criteria, the item pool was reduced to 571 items, which professional translators then translated back into the other 10 languages. The translators offered several recommendations on the linguistic and cultural adequacy of the some items, which resulted in the removal of another 181 items, leaving 250 items in the instrument (Fetvadjiev et al., 2015). The remaining items were subsequently administered to a large, multi-ethnic sample and the stated criteria were applied again on the factor analysed data (Fetvadjiev et al., 2015). The final set of items consisted of 146 items, including the 12 items that is dedicated to the social desirability scale (Deacon, 2016; Fetvadjiev et al., 2015).

The SAPI model corresponds closely to the Big Five constructs, which Fetvadjiev and colleagues (2015) considers strong support for the universality of the Big Five. They do, however, caution against confusing the enforcing of the Big Five and the replication thereof on indigenous cultures, of which the latter was demonstrated with the SAPI project. Whereas the social-relational constructs were highly salient, Openness was the one Big Five factor that replicated the lowest, which corroborates with other studies in non-Western cultures (Fetvadjiev et al., 2015). Although Openness is identified in the SAPI, the construct does not seem to be a very coherent personality concept (Fetvadjiev et al., 2015). Fetvadjiev and colleagues reported a six-factor model for the SAPI, in contrast to the nine-factor model reported by Nel et al. (2012). This has lead Fetvadjiev and colleagues to add their voices to the existing call for an expansion of the Big Five model, to include the social-relational concept that is often found in collectivistic cultures.

2.4.2. Psychometric Properties of the SAPI

There are multiple studies completed and underway to investigate the psychometric soundness of the SAPI. The following section will summarise the literature on the SAPI psychometric properties. Thereafter a brief summary of Mouton's (2017) findings will be provided.

2.4.2.1. SAPI psychometric properties reported by other researchers

The SAPI is still in its development phase. The developers consequently purposely initiated several research studies into the psychometric soundness of the measures of the instrument and the construct-referenced inferences that they intend deriving from the scores obtained on the instrument. Some studies examined the complete instrument (Bruwer, 2016), while other studies focused on the reliability and validity of the

individual first-order factors such as Conscientiousness (Horak, 2012), Emotional Stability (Chrystal, 2012; Cohen, 2013), Extraversion (Geddes, 2012), Intellect (Labuschagne, 2010), Relationship-harmony (French, 2011), and Soft-heartedness (Lubbe, 2012).

Satisfactory internal validity for the SAPI model has been reported through construct and discriminant validity analyses (Bruwer, 2016). Concurrent and predictive validity was also demonstrated for the SAPI through comparisons with instruments measuring constructs such as Cultural Intelligence and Psychological Wellbeing constructs. Despite such reassuring results Bruwer (2016) was unable to replicate the nine factors proposed by Nel et al. (2012), but found statistically significant support for Fetvadjiev et al.'s (2015) suggested six factor model, albeit with a considerably high Chi-square. The very low Normed Chi-square of 1.81, and the satisfactory levels of IFI (.91), TLI (.91), CFI (.91), and RMSEA of .04 that outperformed the five- and nine-factor models, supported the six-factor model fit. It appears that Bruwer (2016) is one of few authors investigating the SAPI's psychometric properties who have applied CFA (in AMOS), in addition to EFA on the SAPI data.

Factor analyses of Emotional Stability demonstrated that the positive and negative aspects of the cluster, measure two separate factors (Chrystal, 2012). The negative facets (i.e. Neuroticism) of Emotional Stability displayed good convergent validity, with Pearson's correlation coefficients showing a strong positive correlation (r = .89) with the Neuroticism scale of the BTI and a moderate positive correlation (r = .66) with the Negative Affect Scales of the PANAS (Chrystal, 2012). Chrystal (2012) found good internal consistency (Cronbach's alpha of .96) for the total Neuroticism scale, as well as for its subscales: Despaired, Anxious, Dependent, Temperamental and Impulsive (with Cronbach's alphas ranging between .86 and .91).
In another study examining the complete Emotional Stability scale, Cohen (2013) used Tucker's congruent coefficients to investigate structural equivalence between three language groups. The reported Tucker's phi for the Emotional Stability factor for each of the participating language groups range between 0.95 and 0.99, exceeding the required minimum level of 0.90 to indicate factorial similarity (Van de Vijver & Leung, as cited in Cohen, 2013). Three items were identified through DIF as problematic and causing a lack of equivalence between the Germanic group and the Nguni and Sotho groups. Cohen (2013) applied DTF and reported significant test bias with v^2 values of 0.20 and 0.26 when comparing the Germanic group with the Nguni and Sotho groups respectively. Cohen concluded that the evidence offered support that the Emotional Stability Scale retained the same meaning across the Germanic, Nguni and Sotho groups.

Although the higher-order Extraversion scale offers promising cross-cultural validity, the overall study findings in turn point toward poor equivalence between the respective groups studied (Geddes, 2012). For instance, construct equivalence on the Extraversion scale revealed that Sociability factors display good construct equivalence, the participating language groups have different perspectives on some of the factors, especially the Talkativeness and Positive Emotionality factors (Geddes, 2012). Geddes recommends that the Talkativeness factor be revised since she found the sub-scale to be psychometrically unreliable and biased.

The Relationship Harmony scale displayed both convergent and predictive validity, by significantly predicting Prosociality and converging with the Agreeableness scale from the Basic Traits Inventory's shortened version (French, 2011).

Horak (2012) recommends that the Conscientiousness scale be revised at sub-factor level, while results from the Schmid-Leiman higher order factor analyses of the Conscientiousness scale offer evidence that a single general factor is measured across the various participating language groups. At sub-scale level items do not consistently measure a particular sub-factor across the various groups, and DTF found between the Germanic and Sotho groups.

The Soft-heartedness scale demonstrated reliability, divergent validity, convergent validity, and predictive validity (Morton, 2011). Method invariance testing was also conducted on this scale and Lubbe (2012) reported that paper-and-pencil testing consistently outperformed the computerised version based on mean scores, skewness, kurtosis, factor loadings, inter-item correlations and reliability. It seems that the participants answered the items in a positive manner, for which she offers several debatable arguments[10] and causes her to recommend the use of a polytomous rating scale rather than the current two-point rating scale.

2.4.2.2. Mouton (2017) SAPI psychometric properties results

Mouton (2017) set out to evaluate the construct validity of the SAPI. She evaluated the construct validity by determining whether the measurement model implied by the internal structure of the personality construct, as constitutively defined by the SAPI developers, along with the design intention of developers in terms of which specific behavioural denotations were assigned to specific latent personality dimensions fitted the item data obtained on the SAPI and whether the measurement model parameter estimates indicated that the items reliably and validly reflected their designated personality factors. To comprehensively evaluate whether the design intention underlying the SAPI succeeded in providing reliable and construct valid measures of the personality construct as constitutively defined, the following analyses were conducted: (i) item analysis; (ii) dimensionality analysis or alternatively referred to as exploratory factor analysis (EFA); and (iii) confirmatory factor analysis (CFA).

The purpose of item analysis is to identify any items that fail to discriminate between different states of the underlying latent variable that the item is supposed to reflect, and/or items that fail to reflect a common underlying latent variable (Mouton, 2017). The items demonstrated a satisfactory ability to discriminate between the different levels of the latent variable as indicated by no outlier items in the standard deviation distributions of the various subscales. with A small number of items that responded too greatly to non-relevant sources of variance that revealed themselves as outlier items in the R^2 and corrected item-total correlation distributions of the various subscales. The items of each of the twenty subscales responded with relative unison to systematic differences for each of the relevant underlying latent variables measured by the subscale items. Eleven of the twenty subscales showed high alpha coefficients (i.e. $.90>\alpha\geq.80$), and the remaining nine subscales returned acceptable alpha coefficients (i.e. $.80>\alpha\geq.70$).

EFA enables the researcher to investigate the extent to which the SAPI reflects the test developers' designed intention to create a questionnaire with twenty uni-dimensional sets of items, which should reflect

[10] For an elaborate explanation of Lubbe's explanation, refer to the study by Lubbe (2012).

variance in each of the twenty personality dimensions and together as a whole comprise personality as conceptualised by the test developers. Mouton (2017) performed principal factor analyses (unrestricted) with oblique rotation on each of the subscales and reported evidence for uni-dimensionality for only seven of the twenty SAPI subscales, and in the remaining thirteen subscales factor fission occurred. The thirteen subscales that failed the test for uni-dimensionality included: Achievement Orientation, Broadmindedness, Conflict Seeking, Deceitfulness, Emotional Balance, Hostile Egoism, Integrity, Intellect, Interpersonal Relatedness, Negative Emotionality, Social Intelligence, Orderliness and Traditionalism-Religiosity. In the case of Integrity, Hostile Egoism and Orderliness an obliquely rotated three-factor solution had to be extracted to obtain a credible explanation of the observed inter-item correlation matrix. The remaining ten subscales delivered a two-factor structure to provide a satisfactory explanation of the observed correlation matrix. Furthermore, in ten of the cases where factor fusion occurred the factor solution was suggested by the eigenvalue-greater-than-unity rule. In three of the cases the extraction of a second factor was indicated by an unacceptably large percentage of large residual correlations associated with the single-factor factor solution (Mouton, 2017). In the thirteen subscales where factor fission occurred, Mouton (2017) could describe the identity of the extracted factors by identifying common themes shared by the items that loaded on each of the extracted factors. In those cases the manner in which Mouton (2017) interpreted the extracted factors revealed the extracted factors to be meaningful facets of the latent first-order personality dimension they were developed to reflect. Although these facets were not at the outset formally acknowledged in the constitutive definition of the latent first-order personality dimensions, the qualitative development history of the SAPI nonetheless implicitly acknowledged that the twenty latent first-order personality dimensions arose as second-order factors out of narrower, more-specific behavioural personality descriptors. The question that Mouton (2017) could not adequately answer was whether the responses to the items comprising each of the thirteen SAPI subscales where factor fission occurred, may permissibly be interpreted as reflecting test takers' standing on the specific latent personality dimension it was earmarked to reflect interpreted as a second-order factor. Mouton (2017) acknowledged that fitting second-order measurement models via confirmatory factor analyses for those subscales where factor fission occurred, would have assisted in arriving at a more definite stance on the matter than the her study was able to achieve. This was acknowledged as a limitation. [11]

The test for exact fit for the SAPI measurement model describing the first-order factor structure indicated that the exact fit null hypothesis (H_{01}:RMSEA=0) had to be rejected, indicating that the SAPI measurement model was unable to reproduce the observed co-variance matrix to a degree of accuracy that could be explained in terms of sampling error alone. The test for close fit indicated that the probability of observing the sample RMSEA value of .0484 (that indicated a good model fit under the close fit null hypothesis) was sufficiently large ($p>.05$) in order for the close fit hypothesis not to be rejected. Mouton (2017) specified that the 90 percent confidence interval for RMSEA (.0479; .0489) further supported the conclusion of a good model fit, as the upper bound of the interval still remained below the critical cut-off of .05. The good fit for the measurement model was further supported with several of the fit indices that exceeded the critical value of

[11] If second-order measurement models would be fitted where factor fission is indicated by an EFA the adequacy of the items of the subscale as indicators of the second-order factor could be determined by calculating (via the PO command in LISREL) the indirect effect of the second-order factor on the item indicator by calculating the product of the loading of item i on first-order factor j and the loading of first-order factor j on the second-order factor ($\lambda_{ij}\gamma_{j1}$).

.90 and the more ambitious critical value of .95. Some of these indices include the normed fit index (NFI = .973), the non-normed fit index (NNFI = .973), the comparative fit index (CFI = .975), the incremental index (IFI = .975) I = and the relative fit index (RFI = .970). Nonetheless, it seems that the measurement model may lack some influential paths since the Akaike information criterion (AIC = 22979.964) for the fitted measurement model could not provide a more parsimonious fit than the saturated model (6320.000), although it did fit more parsimonious than the independent/null model (1051293.007) (Kelloway, 1998).

Mouton (2017) reported highly satisfactory item parcel loadings, with the latent variable that the item parcels were intended to reflect explaining 50% or more of variance in the majority of item parcels. Fifty-five of the seventy two item parcels delivered loadings in excess of .71. Twenty-one item parcels showed loadings between .71 and .60, while only three items parcels loaded between .60 and .50. The lowest completely standardised factor loading (.566) was reported for an item parcel explaining only approximately 32% of the variance in in the variable it was designed to reflect, namely Deceitfulness.

The majority of the latent variable pairs successfully passed the average variance extracted (AVE) criterion[12], which Mouton (2017) regards as a very stringent challenge for instruments measuring comprehensive multi-dimensional constructs comprised of a notable number of latent dimensions. Mouton (2017) concluded that although the SAPI did fail to convincingly discriminate between some of the latent personality dimensions it nonetheless succeeded in doing so on the majority of the latent personality dimensions.

Mouton (2017) originally intended to fit the second-order measurement model that reflects the SAPI's claim that the twenty first-order personality factors load on six second-order personality factors. Mouton (2017) reported that the second-order factor structure unfortunately did not converge, and therefore only the first-order factor structure could be interpreted.

2.5. Conclusion

Chapter two provided an in-depth discussion of the investigation into the complexities of personality assessment in South Africa. Several personality theories were explained in terms of their assumptions, differential characteristics and contribution to the workplace. The trait theory was discussed more extensively, since assessments that are based on trait theory are common in the workplace. South Africa's cultural diversity makes it challenging to use culturally appropriate assessments that are not biased. The country's heritage of disparity is addressed through legislation such as the EEA that aim to solve the crisis by prohibiting unfair discrimination in the workplace and prohibiting psychological instruments that do not comply with psychometric requirements. The danger of naively arguing that the latter will ensure the former was pointed out and motivated. An overview was provided on the different approaches to studying personality across different cultures. The SAPI developers combined the etic and emic techniques to optimise cultural appropriateness and inclusiveness.

[12] The average variance extracted (AVE) reflects the average proportion of variance in the item indicator variables that is accounted for by the latent variable that the indicator variables were designated to reflect (Diamantopoulos & Sigauw, 2000). The AVE calculated for each latent first-order personality dimension should be greater than .50 and should be greater than the squared correlation between the latent variables to indicate discriminant validity (Farrell, 2010)

The SAPI is now in the process of being subjected to the rigorous process of research into the assessment's psychometric soundness. Several research findings were presented, and in particular Mouton's (2017) findings, since this study is a follow-on to her study. The present study aims to contribute to the research regarding the SAPI's measurement invariance and equivalence. The following chapter will elaborate on the theoretical bases and technical understanding and terminology of measurement invariance and equivalence.

CHAPTER 3: MEASUREMENT INVARIANCE AND EQUIVALENCE

3.1. Introduction

This chapter offers an exploration of literature on bias, measurement invariance and measurement equivalence. It will provide an overview on the differentiation, rationale and methodology relating to each of these terms.

3.2. Measurement

Industrial psychologists seek to describe and explain the differences between individuals, groups and organisations by means of latent variables that represent distinguishing attributes, with the purpose of informing human resource-related interventions. Latent variables are however not directly observable. Hence, researchers use measurement instruments whereby numbers are systematically allocated to the latent variables of people, objects or events which the researchers want to investigate (Vandenberg & Lance, 2000). The results obtained from measuring instruments are used in the workplace to predict peoples' behaviour and inform decisions about them. Since the quality of the measurement instruments directly influences the quality of the subsequent intervention[13], industrial psychologists have to ensure that the instruments are of superb quality in order to provide appropriate information.

Historically classical test theory (CTT) has provided the foundation for evaluating the quality of a measurement instrument in terms of true and error scores by means of reliability and validity studies (Vandenberg & Lance, 2000). Although reliability and validity studies provide valuable information, they do not investigate whether the measurement instrument's properties are transportable across populations (Holtzkamp, 2013; Vandenberg & Lance, 2000). In line with Vandenberg and Lance's (2000) argument that the measures derived from measurement instruments should be relatively free from random measurement error (i.e. reliable) and that the construct-referenced inferences derived from such measures should be (construct) valid (i.e. permissible) but that it also should be shown that these construct-referenced inferences are invariant and equivalent across groups, this study will investigate the measurement invariance and equivalence of the SAPI by applying a confirmatory factor analytical (CFA) technique.

3.3. Bias

Test developers strive to achieve the ideal that observed scores are the result of only the latent construct of interest (i.e. that the construct-referenced inferences derived from observed scores are perfectly construct valid). Stated differently, test developers strive to develop samples of stimuli of such a nature that the observed responses of test takers to these stimuli are solely determined by the construct of interest. However, this is a practically unattainable ideal. Observed scores are always to some degree influenced by non-relevant systematic factors and non-relevant random factors that cause differences in the measured scores. Measures are unreliable to the extent that non-relevant random factors produce variance in the observed scores. Measurement bias comprises all systematic factors that could explain unique variance in observed test scores, which cannot be explained in terms of the latent variable of interest (Theron, 2012). Conceptualisations of measurement bias are, however, itself somewhat biased in as far as all sources of

[13] Although reliability, construct validity and lack of measurement bias are necessary conditions to ensure interventions that add value in a fair manner these are not sufficient conditions

non-relevant systematic observed score variance are not of equal interest. Test developers are specifically interested in the question whether (cultural, ethnic, language, age, gender) group membership (i.e. non-relevant sources of systematic variance that correlate with group membership) explain variance in observed scores that cannot be explained in terms of the latent variable of interest. Less interest resides in non-relevant sources of systematic variance that do not correlate with group membership.

Measurement bias is not an inherent property of a measurement instrument, but rather reflects the differences in characteristics of the respondents that cause them to interpret one or more of the items differently (Van de Vijver & Poortinga, 1997) and to respond to the items differently. Several authors (Cheung & Rensvold, 2002; Theron, 2011a; Van de Vijver & Poortinga, 1997) for this reason encourage the investigation into reasons for these differences to contribute to the level of understanding of differences between groups.

These group-related differences in the manner in which test takers respond to the test items in turn produce systematic group-related differences in the relationship between the observed responses and the underlying latent variable. This in turn produces systematic group-related error in the construct-referenced inferences derived from observed test scores if the differences in the relationship between the observed responses and the underlying latent variable are not appropriately acknowledged. The response of test takers from different groups to the items comprising a test could firstly differ in the sense that the construct of interest as constitutively defined does not determine the response to the items in accordance with the design intention of the instrument in all groups. The response of test takers from different groups to the items comprising a test could secondly differ in the sense that although the construct of interest as constitutively defined does determine the response to the items in accordance with the design intention of the instrument in all groups but the regression of the item response on specific latent dimensions of the construct differs across groups. Observed scores may only be regarded as meaningfully interpretable (construct referenced) across different groups when it can empirically be shown that the measures of the latent variables are free from measurement bias (i.e. the same psychological meaning of observed scores) for the respective groups (Vandenberg & Lance, 2000)[14].

Van de Vijver and Poortinga (1997) distinguish between three different forms of measurement bias namely: (a) the construct of interest, (b) the methodological process followed when gathering the data and (c) the item content.

3.3.1. Construct Bias

Construct bias occurs when the scores obtained on specific items reflect different psychological constructs or when the behaviours that denote the measured construct differ across various groups (Holtzkamp, 2013; Meiring, Van de Vijver, Rothmann, & Barrick, 2005). Construct bias is present when differences between groups occur either in terms of a latent variable's internal structure (i.e. the number of latent variables comprising the measurement model), the manner in which the latent dimensions of the construct determine

[14] This claim represents the conventional position regarding measurement bias. The question should, however be considered, specifically in relation to the second form of measurement bias, whether measurement bias cannot potentially be circumvented by formally taking the difference in the regression of item responses on latent dimensions of the construct into account where it occurs?

the response to items (i.e. the loading pattern in the measurement model) or the manner in which the construct is embedded in the larger nomological network (i.e. structural model) (Byrne & Watkins, 2003; Davis, 2014). Davis (2014) argues that the first instance refers to different factor structures that are required to closely reproduce the observed covariance matrix across groups. These factor structures may differ with regard to the number of factors, the correlation between the factors and/or the loading pattern. The latter form of construct bias occurs when different structural models are required across groups to adequately represent the observed inter-item covariance matrix. Test developers are therefore advised to not only investigate instruments for construct bias at the level of the construct's internal structure, but also explore whether the construct functions differently in relation to other constructs in the nomological network.

Van de Vijver and Poortinga (1997) proposed the following causes for construct bias. Firstly, the behaviours that denote a particular construct in one group, might not be appropriate and/or relevant for the same construct in another group. Secondly, the definition attributed to a construct might overlap only partially across different groups. Thirdly, an inadequate number of items might cause insufficient sampling of construct behaviours. Lastly, all the facets of the latent variable might be inadequately covered for different groups.

3.3.2. Item Bias

Item bias only becomes relevant once construct bias has been ruled out. Once the reassurance exists that the test measures the same construct, as constitutively defined, across groups, the question arises whether the test measures the same construct in the same manner? The question is therefore whether a given observed score on an item reflects the same standing on the underlying latent variable across groups? Item bias, also known as differential item functioning (DIF)[15], occurs when respondents with the same standing on the latent variable respond to items differentially across groups, thereby causing the regression relationship between the item response and the underlying latent variable reflected by the item to differ across groups (Byrne & Watkins, 2003; Van de Vijver & Rothmann, 2004).

Theron (2016) offers two perspectives on defining item bias that differ in terms of degrees of stringency. The first is a more lenient item bias definition stating that item bias exists when the expected item score differs across groups, despite them having the same standing on the latent variable. These differences in the expected item scores can be ascribed to differences across group in the regression of the observed responses on the latent variable with regard to intercept and/or slope (Theron, 2016). When the intercept differs across groups, known as uniform bias, the average item score given a specific standing on the latent variable being measured for one group is lower across that whole group's test scores compared to other groups' scores and thereby implies that group membership has a main effect on observed item scores (Theron, 2016; Van de Vijver & Poortinga, 1997). Slope differences across groups, known as non-uniform bias, occur when items have different discriminatory power across groups and thereby imply an interaction effect between group membership and the latent variable (Theron, 2016; Van de Vijver & Poortinga, 1997). The second and more stringent definition states that item bias occurs if the probability of achieving a specific

[15] The item response function in item response theory describes the probability of a given item response as a function of the test takers standing on the latent variable been reflected by the item. The term DIF thus also reflects the fact that at its core item bias lies in differences across groups in the regression of the item response on the latent variable being measured by the item.

observed score on an item, given a specific standing on the latent variable being measured, differs across groups (Theron, 2016). The differences in the probability of achieving a specific observed score can be ascribed to differences across groups in the regression of the observed responses on the latent variable with regard to intercept and/or slope and/or error variance (Theron, 2016). In addition to the slope and/or intercept differences that correspond with the more lenient item bias definition, the stringent definition also includes conditional probability or error variance bias[16], which occurs when error variance on items differs significantly across groups (Holtzkamp, 2013).

Item bias is caused by several sources, of which the first is ambiguous and confusing item content that causes different interpretation of items across different groups (Van de Vijver & Rothmann, 2004). Inadequate item translation due to a lack of knowledge regarding different meanings and/or nuances for certain words across different translations (Van de Vijver & Rothmann, 2004). Item content that is inappropriate and/or unfamiliar to certain respondent groups could also result in differential responses across groups.

The International Test Commission (2012) lists item bias analysis as the first step in analysing a test's psychometric properties[17]. De Beers (2004) agrees and even urges test developers to investigate instruments during test construction for any item bias, since the items can then be corrected or eliminated before publishing the instrument (De Beer, 2004). If the level of measurement bias reduces after the correction or elimination of problematic items, it can be deduced that the difference between groups were due to items being biased and not inherent differences across groups in the latent variable (Van de Vijver & Leung, 1997).

3.3.3. Method Bias

Method bias occurs when group-related characteristics, which are unrelated to the measured psychological construct, cause different groups to respond differently to most or all of the items in an instrument (Theron, 2016; Van de Vijver & Poortinga, 1997). These differential responses often cause groups to have different mean scores that could incorrectly be interpreted as valid differences between the groups, whilst it should be regarded as method bias (Van de Vijver & Leung, 2001). In contrast to item and construct bias, method bias does not describe a unique facet of the relationship between the measured psychological construct and the manner in which test takers respond with observed scores on the indicator variables (Theron, 2016). Theron (2016) proposes that method bias should therefore be regarded as an explanation for why item (and probably construct) bias occurs, rather than a unique type of bias.

Identifying the sources of method bias enables practitioners to avoid the variance caused by it. The three types of method bias include sample bias, administration bias and instrument bias. Sample bias relates to

[16] It is acknowledged that the term conditional probability bias is not a widely used and generally accepted term that in the measurement bias literature. The current study nonetheless considers it an appropriate term in as far as differences in error (or residual) variance in the regression relationship, even when intercepts and slopes coincide, will cause the conditional probability of observing a specific observed score or larger given a specific standing on the latent variable being measured, to differ across groups.

[17] The current study would argue that item bias analysis should form part of the prepublication psychometric analysis that should also include item analysis, dimensionality analysis, confirmatory factor analysis aimed at investigating the construct validity of the (single group) construct-referenced inferences and structural equation modelling aimed at investigating the construct validity of the (single group) construct-referenced inferences. The current study would argue that the latter analyses should precede item bias analysis.

the extent to which samples are not comparable on characteristics that are unrelated to the measured psychological construct such as language proficiency, demographic or biographical characteristics but that affect the response to test items (Byrne & Watkins, 2003; Davis, 2014). Administrative bias occurs when the instrument is administered differently across groups and this causes test-takers to respond differently to the test items[18] (Van de Vijver & Rothmann, 2004). Examples of administrative bias include ambiguous instrument instructions to some groups or providing practice items only to some of the groups, whilst other respondents are not given practice items (Byrne & Watkins, 2003). Instrument bias occurs when general features of the instrument cause groups to respond differently to the test items (Byrne & Watkins, 2003; Van de Vijver & Rothmann, Assessment in multicultural groups: The South African case, 2004). Instrument bias is commonly caused by the following features. Firstly, differential stimuli familiarity occurs when certain groups are unfamiliar with the manner in which the instrument is presented, such as computer-based-testing or Likert-type scaling (Byrne & Watkins, 2003). Secondly, differential response patterns can occur in either response style bias where certain groups consistently select extreme scale points, or in response set bias where certain groups consistently select response options in such a manner as to create a positive impression of themselves (either unconsciously or intentionally) (Byrne & Watkins, 2003).

The foregoing discussion of sample bias, administration bias and instrument bias serves to illustrate that these forms of method bias do not describe an additional aspect of the relationship between the observed responses and the underlying latent dimensions of the construct elicited by the items. Rather these forms of method bias explain why the relationship differs across groups in terms of characteristics that differ across groups or in terms of characteristics that are differentially mobilised across groups (even though they do not differ systematically across groups). This is probably somewhat less apparent in the case of administration bias. If the instrument is allowed to differ across its administration to different groups, specific characteristics become more important in specific groups. If for example, instructions are allowed to become more ambiguous for one group, test performance in that group is likely to be more strongly influenced by characteristics such as abstract thinking capacity than in groups where the instructions are clearer.

3.4. Measurement Invariance and Equivalence

Measurement bias (such as in the case of the current study) and cross-validation studies are closely related concepts, and Little (1997) differentiated between the two concepts by referring to measurement bias as Category 1 invariance and cross-validation as Category 2 invariance. Measurement bias and cross-validation studies differ on the following characteristics. Firstly, while measurement bias studies evaluate multigroup measurement models[19] on samples from different populations (e.g. male and female), cross-validation studies evaluate multigroup measurement models on different samples from the same population (Dunbar et al., 2011). Secondly, measurement bias investigates nuisance factors (that prevent meaningful comparison of scores across groups) that influence the structure of the measurement model and the parameters (Λ^x, τ and Θ_δ) that describe the regression of the observed item responses on the underlying latent variable (Dunbar et al., 2011). Cross-validation studies in turn, investigate the structure of the measurement model and the parameters (Λ^x, τ and Θ_δ), along with the latent variable variances and co-

[18] The instructions, time testing conditions, scoring procedure along with the test items all form part of the test. Strictly speaking administrative bias therefore occurs when the instrument is allowed to differ across groups. Standardisation is a process of controlling these aspects of the test to attempt to ensure that they do not vary across different test occasions.
[19] A multi-group measurement model is defined in equation 2 in Chapter 4.

variances (Φ) (Dunbar et al., 2011). The third difference lies in the examination techniques. Whereas measurement bias is investigated with means and covariance structure (MACS) modelling, covariance structure (CS) is applied in cross-validation studies (Dunbar et al., 2011). CS does not formally model the means of the observed variables, thereby excluding the modelling of the intercepts of the items on latent variables (Dunbar et al., 2011). Dunbar et al. (2011) argue that the exclusion of intercept terms in CS is justified since the purpose of the study is to assess the generalisability of the single-group measurement model (in which the intercepts are not estimated) across multiple groups. The verdict in both a measurement bias and a cross-validation study depends on whether the multigroup measurement model demonstrates invariance and equivalence across the respective groups.

Measurement invariance refers to the examination of the question whether a multigroup measurement model where none, some or all of the parameters are constrained to be equal across groups fits the data from two or more samples. Dunbar et al. (2011) acknowledges that ideally exact fit should be obtained to claim that measurement invariance has been achieved. Due to the low probability of obtaining exact fit in social science research (Browne & Cudeck, 1993), the more lenient stance is generally accepted which states that if a multigroup measurement model demonstrates close-fit it may be concluded that the parameters that were constrained to be equal in the multigroup measurement model are identical across the various groups (Dunbar et al., 2011). Alternatively stated, measurement invariance is demonstrated when the null-hypothesis of close fit cannot be rejected for a multigroup measurement model of which none, some or all parameters are constrained to be equal across groups. Under the conditions that the close fit null hypothesis had not been rejected the position that the constrained measurement model parameters are equal across groups becomes a tenable and permissible position to hold. The position becomes tenable because the parameter estimates that were obtained under these constraints were able to reproduce the observed covariance matrix to such a degree of accuracy that the deviation of the sample RMSEA estimate from .05 could be explained in terms of sampling error only under the close fit null hypothesis.

Measurement equivalence[20] in turn refers to the question whether a particular multigroup measurement model with some of its parameters constrained to be equal across the groups, fits the data significantly poorer than a multigroup measurement model in which fewer parameters are constrained to be equal across the various groups. The extent to which a model fits the data indicates the multigroup measurement model's ability to accurately reproduce group-specific observed covariance matrices. Therefore, if a multigroup measurement model with some of its parameters constrained to be equal fits the data significantly poorer than a multigroup measurement model with no parameter constraints and only the measurement model structure constrained to be the same, it can be concluded that one or more parameters differ significantly across at least two of the various groups (Dunbar et al., 2011).

Testing for measurement invariance can be described as a relatively lenient test for the presence of measurement bias. Finding a lack of invariance therefore constitutes strong evidence of the presence of

[20] It is acknowledged that the measurement bias literature does not general make the conceptual distinction between measurement invariance and measurement equivalence. The measurement bias literature generally uses the two terms interchangeable as synonyms. However, the current study contends that making the conceptual distinction as defined here is useful in clearly separating the two questions as outlined here. The measurement bias literature does, however, generally draw the distinction between these two questions.

measurement bias (specifically item bias). Conversely a finding of measurement invariance constitutes weak evidence of the absence of measurement bias. Testing for measurement equivalence in turn can be described as a more stringent test for measurement bias (specifically item bias). Finding a lack of equivalence therefore indicates weak evidence of a lack of item bias. Conversely, a finding of equivalence constitutes strong evidence of the absence of item bias.

Vandenberg and Lance's (2000) seminal review on measurement invariance indicated that measurement invariance analyses enable test developers to reduce the likelihood of incorrect inferences when instruments are used across different groups. Horn and McArdle (as cited in Vandenberg and Lance, 2000) argue that drawing scientific inferences for psychological instruments are severely lacking and cannot be interpreted unambiguously, without evidence of measurement invariance and equivalence. This line of reasoning underpins the requirement that test developers should prove that construct-referenced inferences can unambiguously be drawn from assessments without systematic group-related error, in order to comply with the EEA (RSA, 1998) that prohibits the use of psychological assessments that are biased. The very real concern, however, exists that the EEA's prohibition of biased "psychological tests and other similar assessments" (RSA, 1998, p. 16) is rooted in the erroneous conviction that the use of biased predictors in selection will unavoidably result in indirect unfair discrimination and conversely that the use of unbiased predictors in selection will ensure that the process of distinguishing, excluding or preferring a person in selection will not constitute indirect unfair discrimination. Hence, from a research integrity perspective advocated by several authors (Dunbar et al., 2011; Mavondo, Gabbott, & Tsarenko, 2003; Vandenberg & Lance, 2000), along with the legislative perspective, assessments should be evaluated to ensure invariance and equivalence across groups. Furthermore, an assessment can only be regarded as invariant if it demonstrates both measurement and structural invariance and equivalence across groups (Vandenberg & Lance, 2000)[21]. Although the scope of this study is limited to the measurement invariance and equivalence of the SAPI, other future research may include other latent variables such as performance and interventions to investigate structural invariance and equivalence.

3.4.1. Evaluating Measurement Invariance and Equivalence

Two types of procedures are used to evaluate measurement invariance and equivalence, which include exploratory factor analysis (EFA) with targeted rotation and confirmatory factor analysis. Exploratory factor analysis (EFA), which Bruwer (2016) and Van de Vijver and Leung (2001) are of the opinion to be the most popular technique, is followed by target rotation and investigating the factorial agreement across the samples. It appears that although the most frequently used measure of factorial agreement is Tuckers' phi, the recommended values to use are disputed (Van de Vijver & Leung, 2001). Therefore, Van de Vijver and Leung (2001) proposes that researchers also inspect the target matrix and rotated source matrix for any differences in loadings to further identify any anomalous items (Van de Vijver & Leung, 2001). Recent studies investigated measurement invariance and equivalence for the SAPI by means of EFA (Cohen, 2013;

[21] The connotative meaning of a construct lies in the internal structure of the construct and in the manner in which the construct is embedded in a larger nomological network of latent variables. To demonstrate that a measure successfully measures a specific construct as constitutively defined, and that construct-references on the construct as conceptualised may permissibly be derived from the observed scores it needs to be shown that the measurement model reflecting the internal structure of the construct (and the design intention of the instrument) as well as the structural model reflecting the manner in which the focal construct is embedded in a larger nomological network, should fit data.

Geddes, 2012; Horak, 2012; Morton, 2011). Several of the authors from these studies recommended that future studies on the measurement invariance and equivalence of the SAPI use CFA (Geddes, 2012; Horak, 2012; Morton, 2011), with Horak (2012) suggesting that EFA makes the process to establish metric and scalar equivalence challenging. Another reason for recommending CFA when investigating measurement invariance and equivalence might be that the EFA hypothesis for testing whether the factor loadings are identical across groups, is only tested once the differences in the eigenvalues of the factors have been corrected (Van de Vijver & Leung, 2001). Van de Vijver and Leung (2001) make a controversial claim that the hypothesis eventually tested in EFA should therefore be regarded as weaker when compared to the hypothesis tested in CFA, which tests the identity of the factor loadings.

As the alternative to EFA, CFA investigates measurement invariance and equivalence by means of a sequence of analyses. Subsequent equivalence analyses are subject to the successful outcome of each increasingly stringent invariance analysis. In other words, an equivalence analysis will only be justifiable if the multigroup measurement model demonstrated close-fit to the data in the preceding invariance analysis (i.e. H_{02}: RMSEA ≤ .05 was not rejected). Measurement equivalence compares the constrained multigroup measurement model with the fully unconstrained measurement model based on various statistics. The ideal would be to use the extremely sensitive chi-square (χ^2) difference test, which is a test that evaluates whether the sample difference between two nested multigroup measurement models may be considered statistically significant. The disadvantage of this sensitivity is that the test also detects trivial difference between the measurement models (Cheung & Rensvold, 2002; Van de Vijver & Leung, 2001). Based on their Monte Carlo study, Cheung and Rensvold (2002) instead recommend reporting the changes in the CFI fit index, the Gamma Hat fit index and the Mcdonald non-centrality index (Mc) as sufficient to conclude whether the compared multigroup measurement models differ practically significantly. Although not agreeing on the usefulness of the Gamma Hat fit index due to its high correlation with the CFI index, Mead, Johnson, and Braddy (2008) recommend that researchers report the χ^2 likelihood ratio test, in addition to the changes in the CFI fit index and the Mcdonald non-centrality index (Mc). In the interest of thoroughness all of the aforementioned CFA statistics will be reported in the current study. Some of the disadvantages of CFA, in addition to the sensitivity of the χ^2 difference test to trivial differences, include researchers being unfamiliar with the technique and the necessity for a priory item classification (Van de Vijver & Leung, 2001).

3.4.2. Taxonomy of Measurement Invariance & Equivalence

This section will provide a taxonomy for the various types of measurement invariance and equivalence that will be analysed in this study. The specific sequence of analyses will be explained in detail in Chapter 4 that elaborates on the methodology to be applied in this study.

Dunbar et al. (2011) recommend that the taxonomy which is displayed in Table 3.1 and initially distinguished by Meredith (1993), should be reserved for evaluating measurement invariance and thereby determining the extent to which multigroup measurement models with increasingly more stringent parameter constraints placed on it, fits the multigroup dataset closely (Davis, 2014). Moving from left to right across Table 3.1 each subsequent invariance analysis is increasingly more stringent than the previous.

Table 3.1

Degrees of measurement invariance

Configural invariance	Weak invariance	Strong invariance	Strict invariance	Complete invariance
A multigroup measurement model in which the structure of the model is constrained to be the same across groups fits multigroup data.	A multigroup measurement model in which the structure of the model is constrained to be the same across groups and in which the factor loading matrix (Λ^x) is constrained to be the same across groups fits multigroup data.	A multigroup measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^x is constrained to be the same across groups and in which the vector of regression intercepts (τ^x) is constrained to be the same across groups fits multigroup data.	A multigroup measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^x is constrained to be the same across groups, in which τ^x is constrained to be the same across groups and in which the measurement error variance-covariance matrix (Θ_δ) is constrained to be the same across groups fits multigroup data.	A multigroup measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^x is constrained to be the same across groups, in which τ^x is constrained to be the same across groups, in which Θ_δ is constrained to be the same across groups and in which the latent variable variance-covariance matrix (Φ) is constrained to be the same across groups fits multigroup data.

(Mels, 2010)

Due to a lack of consistency in literature regarding generally accepted terms to refer to various forms of measurement equivalence, Dunbar et al. (2011) proposed the four hierarchical levels of measurement equivalence displayed in Table 3.2 with their associated definitions. The terms metric and scalar equivalence are generally recognised terms in the literature. Dunbar et al. (2011), however, created the term 'conditional probability equivalence' to provide a term for the scenario in which a multigroup measurement model in which the regression of the indicator variables on the latent variable are constrained to be equal in terms of slope, intercept and error variance across the various groups, does not fit the multigroup data (statistically or practically) poorer than a multigroup model in which only the structure is constrained to be equal across groups. The term therefore refers to strong evidence that the error (or residual) variance of the regression of X_i on ξ_j do not differ across groups. These analyses will assist researchers in determining whether a multigroup measurement model with increasingly more parameter constraints placed on it, fits the dataset (statistically or practically) significantly poorer than a multigroup measurement of which the parameters are freely determined but the structure is constrained to be the same across groups. This explanation in addition implies that a specific form of equivalence will only be tested if configural invariance has been shown and if the constrained multigroup model showed close fit (i.e. weak, strong or strict invariance has been shown).

Table 3.2

Degrees of measurement equivalence

Metric equivalence	Scalar equivalence	Conditional probability equivalence	Full equivalence
A multigroup measurement model in which the structure of the model is constrained to be the same across groups and in which the factor loading matrix (Λ^x) is constrained to be the same across groups does not fit multigroup data poorer[3] than a multigroup measurement model in which the structure of the model is constrained to be the same across groups but all model parameters are freely estimated (i.e., the configural invariant multigroup model).	A multigroup measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^x is constrained to be the same across groups and in which the vector of regression intercepts (τ^x) is constrained to be the same across groups does not fit multigroup data poorer than a multigroup measurement model in which the structure of the model is constrained to be the same across groups but all model parameters are freely estimated.	A multigroup measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^x is constrained to be the same across groups, in which τ^x is constrained to be the same across groups and in which the measurement error variance-covariance matrix (Θ_δ) is constrained to be the same across groups does not fit multigroup data poorer than a multigroup measurement model in which the structure of the model is constrained to be the same across groups but all model parameters are freely estimated.	A multigroup measurement model in which the structure of the model is constrained to be the same across groups, in which Λ^x is constrained to be the same across groups, in which τ^x is constrained to be the same across groups, in which Θ_δ is constrained to be the same across groups and in which the latent variable variance-covariance matrix (Φ) is constrained to be the same across groups does not fit multigroup data poorer than a multigroup measurement model in which the structure of the model is constrained to be the same across groups but all model parameters are freely estimated.

(Dunbar et al., 2011, p. 8)

A finding of configural invariance means a finding of a lack of construct bias. A finding of weak invariance constitutes weak evidence of a lack of non-uniform bias. A finding of a lack of weak invariance constitutes strong evidence of non-uniform bias in one or more items[22]. A finding of strong invariance constitutes weak evidence of a lack of uniform bias. A finding of a lack of weak invariance constitutes strong evidence of uniform bias in one or more items. A finding of strict invariance constitutes weak evidence of a lack of error variance bias. A finding of a lack of strict invariance constitutes strong evidence of error variance bias in one or more items.

A finding of metric equivalence (given a finding of weak or partial weak invariance) constitutes strong evidence of a lack of non-uniform bias. A finding of a lack of metric equivalence constitutes weak evidence of non-uniform bias in one or more items. A finding of scalar equivalence (given a finding of strong or partially strong invariance) constitutes strong evidence of a lack of uniform bias. A finding of a lack of scalar

[22] A finding of a lack of weak invariance requires the items suffering from non-uniform bias to be identified before proceeding with the test of strong invariance (assuming that not all items are suffering from non-uniform bias and that close fit for a multi-group model is obtained in which the slope parameter is freely estimated across groups for a limited number of items. The issue of partial invariance is discussed in the next paragraph.

equivalence constitutes weak evidence of uniform bias in one or more items. A finding of equal probability equivalence (given a finding of strict or partially strict invariance) constitutes strong evidence of a lack of error variance bias. A finding of a lack of equal probability equivalence constitutes weak evidence of error variance bias in one or more items.

3.4.2.3. Partial Measurement Invariance and Partial Measurement Equivalence

Given the more lenient and more stringent definitions of item bias, the ideal will always be to achieve strict invariance and equal probability equivalence. But, such results seldom occur and assessments might be invariant and equivalent in some but not all populations (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Weak, strong and strict invariance and metric, scalar and conditional probability equivalence evaluations only apply to multigroup measurement models that have demonstrated configural invariance. Weak invariance and metric equivalence is sometimes difficult to achieve and consequently several authors recommend that lessening the stringency in favour of partial measurement invariance and partial measurement equivalence on those evaluations' conditions, to permit cross-group comparisons (Byrne et al., 1989; Vandenberg & Lance, 2000). Hence, partial measurement invariance and partial measurement equivalence offers researchers the opportunity to conduct cross-group comparisons that would otherwise not be appropriate. It can be deduced that partial weak invariance therefore only becomes relevant when weak invariance has not been obtained and partial metric equivalence only becomes relevant when the metric equivalence requirement has not been met (but weak invariance, or partial weak invariance, has not been obtained)[23]. If a lack of weak invariance is found, the source of the lack of invariance needs to be identified so as to (a) flag the items suffering from non-uniform bias, to (b) determine which items' parameters should be freed to allow the more stringent test of metric equivalence, and to (c) determine whether it is meaningful to examine whether any of the items suffer from uniform bias. It would only be meaningful to evaluate whether any item suffers from uniform bias, if not all items have been flagged as suffering from non-uniform bias under the more lenient and/or more stringent test of non-uniform bias. This logic extends to findings of a lack of strong invariance and a lack of strict invariance (if a mere stringent interpretation of item bias is embraced).

Partial measurement invariance and partial measurement equivalence requires the researcher to relax equality constraints on the models so as to allow invariant parameters to be freely estimated until close fit is

[23] Partial invariance and partial equivalence brings to the fore the need for an extended taxonomy and terminology that has as yet not been clarified. The problem is that a variety of combinations of findings are possible. Firstly, partial weak invariance can be obtained and when comparing the partial weak invariance to the configural invariance model, no significance difference in fit is obtained (partial weak-metric equivalence). Secondly, weak invariance can be obtained but when subjected to the more stringent metric equivalence one or more biased items are flagged. Once the slope parameters for the biased items are freely estimated across groups partial metric equivalence is obtained (weak invariance-partial metric equivalence). Thirdly, partial weak invariance can be obtained and when comparing the partial weak invariance to the configural invariance model significance difference in fit is obtained (partial weak invariance-partial metric equivalence). Fourthly, weak invariance can be found along with metric equivalence (weak invariance-metric equivalence). The number of permutations quickly increase when strong invariance and scalar equivalence is also brought into play. Looking at strong invariance and scalar equivalence in isolation the same four possible outcomes exist (partial strong invariance scalar equivalence, partial strong invariance partial scalar equivalence, strong invariance partial scalar equivalence and strong invariance scalar equivalence). Each of these four combinations can, however, now combine with any of the four possible combinations that could result from the evaluation of non-uniform bias. Resulting therefore in 16 possible outcomes, when testing for both non-uniform and uniform bias. The problem further aggravates under the more stringent interpretation of item bias that also requires the investigation of strict invariance and equal probability equivalence. In total there are, under the strict interpretation of item bias 64 different outcome combinations. The current study suggests that the only way of differentiating between them is to use a 3 double barrel descriptor (e.g. partial weak invariance-partial metric equivalence; partial strong invariance-scalar equivalence; partial strict invariance-partial conditional probability equivalence).

obtained and the difference in the fit of the partially constrained (H_0) measurement model and the unconstrained (H_a) measurement model is no longer (statistically or practically) significant. The question is how the items should be identified for which the equality constraint on the slope, intercept or error variance parameter should be lifted, and in what order the constraints should be lifted, until close fit is achieved, or until the difference in model fit is no longer (statistically or practically) significant. Although the first recommended process for partial measurement invariance and partial measurement equivalence was proposed almost three decades ago by Byrne, Shavelson and Muthén (1989), there remains a lack of consensus on the optimal process for evaluating partial measurement invariance and partial measurement equivalence (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). For instance, the statistical criteria (e.g. modification indices (MIs), expected parameter changes (EPCs), and goodness-of-fit heuristics) for relaxing invariance constraints are not applied consistently, nor is there consensus in the manner which the criteria should be applied (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). This lack of consensus has resulted in a lack of literature regarding the statistical or conceptual implications of allowing partial invariance and partial equivalence (Putnick & Bornstein, 2016).

The number of invariant items that may be released and still claim partial invariance and partial equivalence is another contentious issue. Steenkamp and Baumgartner (1998) propose releasing constraints for metric invariance up to the point of only retaining the reference indicator and a single other indicator item per latent variable. In contrast, Vandenberg and Lance (2000) insist that releasing the majority of invariant items will result in inappropriate comparisons of mean group differences on measures that are non-equivalent. Vandenberg and Lance maintain that constraints should only be relaxed when: (a) the researcher has a sound theoretical argument to support it; (b) the problematic items are in the minority; and (c) and when the constraints' relaxation viability is supported by cross-validation evidence. The criteria recommended by Vandenberg and Lance will be applied in this study.

The following are processes recommended in literature to identify invariant items. The discussion focusses on the detection of non-uniform bias in specific items. The procedures, however, apply equally well to the identification of uniform bias and error variance bias in specific items. The first process applies to multigroup measurement models that consist of several constructs, whereby any indicator that demonstrates a lack of invariance is identified (Byrne et al., 1989; Cheung & Rensvold, 1999). A separate multigroup measurement model is tested for each construct by constraining all the factor loadings[24] for the relevant construct of interest to be equal across the groups, while all other factor loadings are allowed to freely estimate (Cheung & Rensvold, 1999). Each model's fit to the data is then compared with that of the configural invariance model to determine whether or not it fits significantly poorer (Cheung & Rensvold, 1999)[25]. This comparison between the models can be based on either a statistical or practical significance level (Cheung & Rensvold, 1999). A significant poorer fit is indicative of at least one invariant item in the construct of interest. The researcher then reports any subscale that lacks invariance, before evaluating them to identify the items that lack invariance. Byrne et al. (1989) proposes that the same process is repeated on item level, whereby

[24] The same procedure would apply to intercepts or error variances in the case of uniform and error variance bias.
[25] The literature's use of the term invariance in this context illustrates the point made earlier that typically the terms invariance and equivalence are used interchangeably as synonyms. The current study would regard the suggested here as a procedure aimed at identifying biased items to achieve partial equivalence rather than partial invariance.

separate measurement models are created in which the contribution of each item in the subscale lacking invariance is established, by evaluating whether the decrease in model fit is (practically or statistically) significant, when compared to the configural model's fit. The items can be released in a sequential order either by adding item constraints (forward method) in an iterative manner, or by constraining all the items and sequentially releasing item constraints (backward method) (Putnick & Bornstein, 2016; Yoon & Kim, 2014). The disadvantage of this process is that large measurement models such as that of the SAPI may require a very cumbersome iterative process (Davis, 2014).

The second proposed process to identify items lacking invariance involves investigating the configural model's factor loadings for items that differ greatly across the groups (Cheung & Rensvold, 1999)[26][27]. However, this technique does not directly apply significance tests to identify items lacking invariance. The item that shows the largest parameter difference in the configural invariance model across groups has its factor loading (or intercept or error variance) freely estimated across the groups and the close fit of the partial invariance measurement model is then evaluated. If close fit is not attained the procedure is repeated. The procedure is repeated until close fit is attained. Whether the improvement in fit brought about by the freeing of each item parameter is statistically significant is not typically evaluated.[28]

The third technique in turn investigates whether factor loadings are significant for one group but not the other group (Cheung & Rensvold, 1999). However, the concern then arises in instances where the significance values are almost the same across the groups but it is significant for one group but not for the other (Cheung & Rensvold, 1999; Davis, 2014). Moreover, parameter estimates can still differ significantly even when they are both statistically significant.

The fourth technique involves examining whether a fully constrained model contains any large modification index values (MIs) and expected parameter change values (EPCs), which are indicative of items that lack invariance and that will improve the model fit if released to be freely estimated (Davis, 2014; Cheung & Rensvold, 1999). This technique should be used with caution as it can potentially allow cross-loadings from items on different sub-dimensions (Davis, 2014; Steenkamp & Baumgartner, 1998). Steenkamp and Baumgartner (1989) stress that when MIs and EPCs are used to release invariance constraints, it should only be done when EPCs are substantial and MIs are highly significant in terms of magnitude and in comparison with other items. The number of model modifications should be kept to the minimum, and only be applied to respecifications that will rectify severe problems in model fit.

[26] Typically the comparison would be made based on the common metric completely standardised solutions for each group.
[27] The technique seems to implicitly assume two groups. However, this need not necessarily be the case. Which then raises the question how the technique generalises from two groups to three and more groups? In the case of three groups there are three difference scores that can be calculated for each group and in the case of four groups there are 6 difference scores to be calculated. One possibility is to identify for each item the highest difference score and rank-order these. Equality constraints imposed on item parameters are then released in accordance with this rank-ordered list of items.
[28] There is, however, really no reason why it could not be evaluated by calculating the chi-square difference and evaluating its statistical significance at one degree of freedom.

In a multigroup measurement model analysis each construct contains a referent item (alternatively known as a marker item) for which the slope parameter is set to be equal to unity across the various groups[29]. It is possible that the referent item might suffer from non-uniform bias and be one of the items that cause the subscale to lack invariance. The last two techniques are recommended in literature as ways to investigate whether the referent item lacks invariance. The fifth technique, known as the "triangle heuristic", addresses this problem by allowing the researcher to identify a set of several invariant items. An invariant set of items is a number of items that can all serve as referent items to the other items. Hence, if an item cannot be included in the set of invariant items, (i.e. that all other items cannot serve as referent items to a particular item) it can be deduced that such particular item lacks invariance (Cheung & Rensvold, 1999). The triangle heuristic process entails systematically drawing up a matrix with all referent items (i.e. all items in the set except the one investigated for invariance) in the columns and the items investigated (i.e. arguments) in the rows (Cheung & Rensvold, 1999). The rows and columns can be swapped around with the main aim to create the largest possible closed triangular array of nonsignificant entries below the diagonal (Cheung & Rensvold, 1999). Invariant sets of items are constructed by items that define the rows and columns of the triangle (which includes the diagonal) (Cheung & Rensvold, 1999).

The last technique, known as the factor-ratio test, can also be applied to identify whether the referent item is causing the lack of variance in the construct. The researcher systematically investigates all combinations of the referent item and items tested for invariance, across all the groups, again using the matrix format of arguments and referents that Cheung and Rensvold (1999) suggested (Cheung & Rensvold, 1999; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The difference with triangle heuristic method is that one item serves as the referent item, against which all other items are tested for invariance. The item that cannot serve as a referent item to the rest of the items is identified as problematic items.

Several authors have investigated the superiority in effectiveness of different processes and techniques to evaluate partial invariance and partial equivalence. Sequential releasing of parameters based on the highest modification index offered a smaller Type I (false positive) error rate than the method of releasing all problematic parameters at the same time (Yoon & Kim, 2014). However, when a high proportion of the items appear to lack invariance the sequential release method presents researchers with limitations such as unsuccessful identification of noninvariant items (Yoon & Millsap, 2007). Furthermore, the sequence of releasing parameters is also contested. Jung and Yoon (2016) compared the effectiveness of three types of systematic releasing techniques. They reported that systematic adding of parameters (forward method) was more effective than the systematic releasing of parameters (backward method) except when adjusted criteria are used, in which case both methods works well. The third type of sequential releasing, the factor-ratio method, demonstrated the highest level of error rates.

Once referent items have been identified, and partial weak invariance and metric equivalence is achieved, partial strong invariance and partial scalar equivalence may be investigated (Steenkamp & Baumgartner, 1998) utilising essentially the same procedures. Once items have been released to freely estimate in one

[29] This is done to set the metric of the latent variable equal to the metric of the marker variable. This is preferred over standardisation typically used in single group measurement model analysis (which sets the unit of measurement to the standard deviation) so as to allow the variances of the latent variables to be estimated and not set to 1 and equal across groups.

step of the measurement invariance and equivalence process they remain unconstrained for the subsequent steps.

The current study identified noninvariant items as recommended in the fourth procedure, by calculating and rank-ordering the absolute difference in the common metric completely standardised factor loadings obtained for the configural invariance multigroup model. The items with the highest absolute difference in the respective aspects investigated in each invariance or measurement equivalence, was allowed to be freely estimated in favour of a partial measurement invariance or partial measurement equivalence model.

The question then arises as to what to do with problematic items that have been identified. Cheung and Rensvold (1999) recommend the following options to deal with items that seem to be the cause of the lack of measurement invariance and/or measurement equivalence, namely: (a) delete any items that cause the multigroup measurement model not to display close fit and thus cause the lack of invariance; (b) interpret the data from items that lack invariance as cross-group data in its own right[30]; and (c) apply partial measurement invariance and partial measurement equivalence to retain problematic items. In line with this reasoning, one could argue that problematic items should not be considered when making inferences on responses. However, inferences are drawn based on dimension scores, not on individual item scores. When deciding what to do with an item, the researcher should consider the aggregate impact that noninvariant items have on dimension level. For instance, since some noninvariant items might benefit males, whereas other noninvariant items might benefit females, one needs to investigate the overall impact of these items on dimension level. Alternatively stated, it needs to be investigated whether the noninvariant items cancel each other out, or are consistent in being biased in favour of a particular group on dimension level. Deleting items is not in the scope of the current study, since the intention is to contribute to the body of knowledge regarding the extent to which the SAPI demonstrates measurement invariance and equivalence across gender groups. Hence, the first option was not considered[31]. Guenole and Brown (2014) advocate the releasing of invariant parameter estimates to account for partial invariance of items eliciting a construct, rather than allowing the relations among constructs to be negatively impacted by non-invariant items that are ignored. Hence, the second option proposed by Cheung and Rensvold was also not considered. Therefore the third option of investigating partial measurement invariance and partial measurement equivalence was chosen, in the event that any SAPI items display a lack of measurement invariant and/or equivalence. Partial measurement invariance and partial measurement equivalence offers a solution when using biased measures to evaluate differences in latent means across groups or when investigating structural invariance and equivalence across groups. Partial measurement invariance and partial measurement equivalence, however, does not offer a practical solution when interpreting dimension scores obtained from biased measures across groups. This topic will be further debated in Chapter 7.

3.5. Conclusion

This chapter provided an in-depth study on the different types of bias. It further provided the differences between measurement invariance and equivalence, as well as the measurement invariance and equivalence

[30] In this instance test developers should acknowledge when they leave noninvariant items in a dimension to be interpreted.
[31] The first option can only be considered by the developers of the SAPI. The current study can only flag specific items as displaying non-uniform bias and/or uniform bias and/or error variance bias.

taxonomy wat will be applied in the current study. In the instance that certain items might be found to lack invariance, different methods for evaluating partial invariance and partial equivalence was presented in detail. The following chapter will provide a detailed description on the research methodology that was applied during the investigation of the measurement invariance and equivalence for the SAPI.

CHAPTER 4: RESEARCH METHODOLOGY

4.1. Introduction

This chapter elaborates on the research methodology that was applied in the research study. The substantive research hypotheses are provided which formed the foundational hypotheses for the study. The research design provides the investigative plan through which the substantive hypotheses were explored. To empirically test the permissibility of the substantive hypotheses, they were translated into statistical hypotheses. The desired sample characteristics are explained, after which the series of statistical analyses are also explained in depth. An acknowledgement of limitations to the research study is provided, before the chapter concludes with several ethical matters that were considered.

4.2. Substantive Research Hypothesis

The SAPI was developed to measure personality in the South African population. As it is argued in Chapter 3, to enhance the verdict that (a) inferences on the personality construct as constitutively defined by the SAPI may permissibly be derived from scores obtained on the SAPI for both genders and that (b) the inferences derived about individuals standing on the latent personality dimensions given specific observed dimension scores are the same for both genders, the measurement and structural invariance and equivalence has to be shown to be satisfactory. Therefore, to ensure that the inferences on the personality construct as constitutively defined by the SAPI from scores obtained on the SAPI are permissible for both genders and can be used with confidence in the same manner across genders, the measurement invariance and equivalence were evaluated in this study. The evaluation of the SAPI's structural invariance and equivalence is not in the scope of this research study.

The overarching substantive hypothesis is that the SAPI provides a valid and reliable measure of personality as constitutively defined by the SAPI across both gender groups, and the manner in which it measures personality remains the same across gender groups. This translates into the following specific operational hypotheses:

Measurement invariance

- Operational hypothesis 1:
  A single-group measurement model as implied by the SAPI scoring key closely reproduces the observed covariances among the individual items that comprise the various personality dimensions, in both gender groups separately (Dunbar et al., 2011; Theron, 2016).
- Operational hypothesis 2:
  A multigroup measurement model as implied by the SAPI scoring key, in which all measurement model parameters are set to be freely estimated across the two gender groups except the structure that is constrained to be equal across groups, closely reproduces the observed covariances between the individual items that comprise the various personality dimensions, in the combined sample. Alternatively stated, the configural invariance multigroup measurement model that is implied by the SAPI scoring key demonstrates close fit on the gender groups data (Dunbar et al., 2011; Theron, 2016). By achieving close fit of the configural invariance model operational hypothesis 2 is corroborated and evidence of a

lack of construct bias would be demonstrated (i.e. the number of factors and loading pattern do not differ between the two groups) (Dunbar et al., 2011; Theron, 2016).

- Operational hypotheses 3:

  A multigroup measurement model as implied by the SAPI scoring key, of which the structure and the strength of the factor loadings (lambda matrix) are constrained to be equal across the two gender groups but the remaining measurement model parameters are set to be freely estimated, closely reproduces the observed covariances between the individual items that comprise the various personality dimensions, in the combined sample (Dunbar et al., 2011; Theron, 2016). Alternatively stated, the multigroup weak invariance measurement model that is implied by the SAPI scoring key demonstrates close fit on the gender groups data. By achieving close fit of the weak invariance model, operational hypothesis 3 is corroborated and weak evidence of a lack of non-uniform bias would be demonstrated (i.e. the slopes of the relationships between the item responses and the underlying latent personality dimensions that they represent do not differ between the two groups) (Dunbar et al., 2011; Theron, 2016).

- Operational hypotheses 4:

  A multigroup measurement model as implied by the SAPI scoring key, of which the structure (i.e. number of factors and loading pattern), the strength of the factor loadings (lambda matrix), and the intercepts terms (tau vector) are constrained to be equal across the two gender groups, but the rest of the parameters are left to be freely estimated, closely reproduces the observed covariances between the individual items that comprise the various personality dimensions, in the combined sample (Dunbar et al., 2011; Theron, 2016). Alternatively stated, the strong invariance multigroup measurement model that is implied by the SAPI scoring key demonstrates close fit on the gender groups data. By achieving close fit of the strong invariance model, operational hypothesis 4 is corroborated and weak evidence of a lack of uniform bias would be demonstrated (i.e. the intercepts of the relationships between the item responses and the underlying latent personality dimensions that they represent, do not differ between the two groups). Therefore, due to the hierarchical nature of the various multigroup invariance models, by demonstrating strong invariance, weak evidence of a lack of both uniform and non-uniform bias has been established during the overall process (Dunbar et al., 2011; Theron, 2016). A finding of strong invariance therefore implies weak evidence of a lack of item bias across the items of the SAPI when item bias is more leniently defined (see paragraph 3.3.2).

- Operational hypotheses 5:

  A multigroup measurement model as implied by the SAPI scoring key, of which the structure (i.e. number of factors and loading pattern), the strength of the factor loadings (lambda matrix), the intercepts terms (tau vector) and the measurement error variance terms (theta-delta matrix) are constrained to be equal across the two gender groups, but the phi parameters are left to be freely estimated, closely reproduces the observed covariances between the individual items that comprise the various personality dimensions, in the combined sample (Dunbar et al., 2011; Theron, 2016). Alternatively stated, the strict invariance multigroup measurement model that is implied by the SAPI scoring key demonstrates close fit on the gender groups data. By achieving close fit of the strict invariance model, operational hypothesis 5 is corroborated and weak evidence of a lack of error variance bias would be demonstrated (i.e. the dispersion of observations around the regression of the item responses on the underlying latent personality dimensions that they represent do not differ between the two groups). A finding of strict

invariance therefore implies weak evidence of a lack of item bias across the items of the SAPI when item bias is more stringently defined (see paragraph 3.4.).

Measurement equivalence[32]*:*

Strong evidence of the absence of item bias, stringently defined, in the SAPI items will exist if the multigroup measurement model that is implied by the SAPI scoring key demonstrates metric equivalence, scalar equivalence and conditional probability equivalence across the gender groups. The degree of measurement equivalence is determined by comparing the fit between the configural invariance multigroup measurement model and the weak, strong and strict invariance multigroup measurement models with increasing degrees of equality constraints placed on the model parameters (Dunbar et al., 2011; Theron, 2016). Measurement equivalence would be demonstrated when a "multigroup measurement model with no equality constraints placed on its parameters or with specific equality constraints placed on the parameters, does not fit practically or statistically significantly poorer than a model with more specific equality constraints placed on its parameters" (Theron, 2016, p. 323). The significance of equivalence is evaluated either practically or statistically. For the purposes of this research study the equivalence would be evaluated based on practical significance.

- Operational hypotheses 6:

  The multigroup measurement model that is implied by the SAPI scoring key demonstrates metric equivalence across the gender groups. Metric equivalence is demonstrated when the weak invariance multigroup measurement model in which only the structure and the strength of the factor loadings (lambda matrix) are constrained to be equal across the two gender groups, but the rest of the parameters are left to be freely estimated, does not fit practically significantly poorer on the gender groups data than the configural invariance multigroup measurement model in which only the structure is constrained to be equal across the two gender groups but all model parameters are freely estimated (Dunbar et al., 2011; Theron, 2016).

- Operational hypotheses 7:

  The multigroup measurement model that is implied by the SAPI scoring key demonstrates scalar equivalence across the gender groups. Scalar equivalence is demonstrated when the multigroup measurement model in which only the structure, the strength of the factor loadings (lambda matrix) and the intercepts terms (tau vector) are constrained to be equal across the two gender groups, but the rest of the parameters are left to be freely estimated, does not fit practically significantly poorer on the gender groups data than the configural multigroup measurement model in which the structure is constrained to be equal across the two gender groups but all parameters are freely estimated (Dunbar et al., 2011; Theron, 2016).

[32] A hypothesis on the multi-group measurement model as implied by the SAPI scoring key, of which all parameters (including the freed elements of the phi matrix) are constrained to be equal across the two gender groups, could also have been formulated. Alternatively stated, a hypothesis on the fit of the complete invariance multi-group measurement model that is implied by the SAPI scoring could also have been formulated. Complete invariance, or the lack thereof, does not, however, add any information on item bias or the lack thereof. Complete invariance was therefore not evaluated. It could possibly be argued that evidence on complete invariance could add information on construct bias. The current study, however, choose not to do so.

- Operational hypotheses 8:

  The multigroup measurement model that is implied by the SAPI scoring key demonstrates conditional probability equivalence across the gender groups. Conditional probability equivalence is demonstrated when the multigroup measurement model with the structure, the strength of the factor loadings (lambda matrix), the intercepts terms (tau vector) and the measurement error variance terms (theta-delta matrix) are constrained to be equal across the two gender groups, but the phi parameters are left to be freely estimated, does not fit practically significantly poorer on the gender groups data than the configural invariance multigroup measurement model in which the structure is constrained to be equal across the two gender groups but all parameters are freely estimated (Dunbar et al., 2011; Theron, 2016)[33].

4.3. Research Design

The substantive hypothesis described in the aforementioned paragraphs posits that four specific multigroup SAPI measurement models on which increasingly strict equality constraints are imposed should closely fit the gender groups data and that the imposition of increasingly strict equality constraints on the model parameters should not practically significantly weaken the fit of the multigroup measurement model in relation to the configural invariance multigroup SAPI measurement model. These positions have been captured in the eight operational hypotheses formulated in paragraph 4.2. The purpose of this research study was to determine whether the construct-referenced inferences made from the SAPI are construct valid and gender unbiased and can subsequently be used with confidence across the two gender groups. It is acknowledged that ideally the SAPI would need to be fitted into the larger nomological network in which the SAPI indicator variables, employee performance measures, and other latent variables were embedded in a structural model and tested empirically. However, for the purpose of this study only the exogenous multigroup measurement model was evaluated for measurement invariance and equivalence. It is noted that even if the multigroup measurement model demonstrated a good fit to the data, the evidence would remain insufficient to conclude that the SAPI is indisputably cleared from any bias between the genders. Yet, the inverse is not true. If the multigroup measurement model would not fit the data well, serious concerns would be cast on the inferences derived from the SAPI.

The research design describes the strategy or investigative plan that researchers follow when gathering and investigating evidence in the testing of the substantive and/or operational hypotheses and controlling variance, that will enable them to unambiguously interpret the results when answering the research initiating question (Babbie & Mouton, 2001; Brits, 2011; De Vos, Strydom & Fouche, as cited in Brits, 2011; Kerlinger, 1973; Theron, 2012). Moyo (2009) stated that the purpose of a research design is to evaluate whether the stated operational hypotheses have any merits, by conducting a systematic empirical enquiry in such a manner that the obtained results can be used and interpreted unambiguously either for or against the operational hypotheses. Furthermore, the research design need to make provision for controlling for variance (Kerlinger, 1973) by maximizing systematic variance, minimising error variance and controlling external variance (MAXMINCON) (Kerlinger, 1973; Kerlinger & Lee, 2000)

[33] A hypothesis on the degree to which the complete invariance multi-group measurement models practically significantly fits the gender groups' data poorer than the configural invariance multi-group measurement model. Full equivalence, or the lack thereof, does not, however, add any information on item bias or the lack thereof. Full equivalence was therefore not evaluated.

Traditionally (in an explanatory research study) the operational hypotheses indicates the tentative relational statement which hypothesises a specific relationship in a structural model between one or more independent observed variables (X) and a minimum of one dependent observed variable (Y) (Moyo, 2009). This relationship would be displayed when Y changes in a certain manner due to changes in X. It is therefore critical to choose a research design that can distinguish the variance in Y that is attributable to the independent variable investigated (X), from the variance in Y that is attributable to other non-relevant X variables such as error variance (Kerlinger, 1973; Kerlinger & Lee, 2000; Theron, 2012). Through the variance controlling objective MAXMINCON research designs attempt to clearly distinguish between variance in the dependent variables caused by relevant and non-relevant independent variables respectively. Firstly MAXMINCON requires the research design to maximise systematic variance which increases the likelihood that H_0 will be rejected during statistical testing. Secondly, it requires the research design to minimise error variance which increases the likelihood that the impact of X on Y becomes more discernible from the impact of other non-relevant independent variables on Y and thereby also increases the likelihood that H_0 will be rejected during statistical testing. Thirdly, the principle requires research designs to control for extraneous variance by (*inter alias*) incorporating non-relevant extraneous variables into the design as covariates during the statistical analysis. (Kerlinger & Lee, 2000).

The current study did not investigate the validity of a relational statement which hypotheses a specific relationship in a traditional structural model, but instead investigated only a measurement model, more specifically a multigroup measurement model. A measurement model assumes that indicator variables regress with a positive (or negative) and statistically significant slope on specific latent variables that are represented by the indicator variables (Moyo, 2009). The dependent variables in the measurement model are the observable indicator variables (i.e. the SAPI items) and the independent variables of interest are the personality constructs. Operational hypotheses 2 to 8 constituted hypotheses on the extent to which the regression of item responses on latent personality dimensions are the same across the two gender groups. The current study therefore still investigated relationships between (observed and latent) variables and therefore required a strategy or investigative plan to guide the gathering of empirical evidence in the testing of the tentative claim made regarding the relationships. Since these latent variables are inherently not manipulatable and no structural (i.e. causal) interrelations were hypothesised to exist between them, the study applied a non-experimental research design (Kerlinger, 1973; Kerlinger & Lee, 2000). Therefore, to test the claims made by operational hypotheses 2 to 8 on the nature of the relationship between specific item responses and specific latent personality dimensions in the two gender groups an *ex post facto* correlational research design was most appropriate for this purpose. The *ex post facto* correlational research design as it applies to the current study is shown in Figure 4.1.

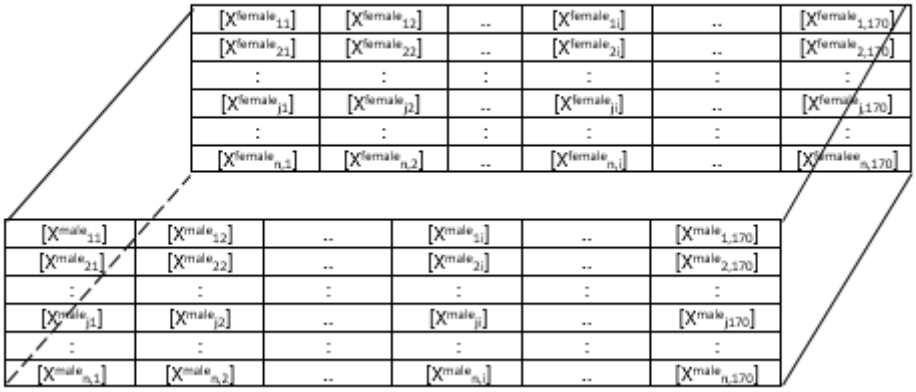

Figure 4.1. A schematic depiction of the *ex post facto* correlational design used to evaluate measurement bias in the SAPI[34]

The logic underpinning an *ex post facto* correlational research design is that it instructs the researcher to individually observe the various indicator variables in the measurement model in each gender group and to determine the extent to which they co-vary in each group (Brits, 2011; Kerlinger & Lee, 2000; Moyo, 2009). Estimates of freed parameters in the multigroup measurement model are obtained in an iterative fashion with the objective of reproducing the observed gender-specific covariance matrices as closely as possible (Diamantopoulos & Siguaw, 2000). If the measurement model failed to accurately reproduce the observed covariance matrices (as judged by the fit statistics calculated for the multigroup model), it would indicate that the specific multigroup SAPI measurement model does not provide an acceptable explanation for the observed covariance matrices (Diamantopoulos & Siguaw, 2000; Kelloway, 1998). Alternatively stated, the failure to accurately reproduce the observed covariance matrices would have indicated that the equality constraints that were hypothesised to characterise the relationships hypothesised by the multigroup measurement model to exist between specific item responses and specific latent personality dimensions, do not provide an accurate portrayal of the psychological process that shapes male and female test takers' performance on the SAPI. Yet, the inverse is not true. If the fitted gender-specific covariance matrices derived from the freed parameters estimates obtained for the multigroup measurement model closely agreed with the observed gender-specific covariance matrices, it could not be concluded that the psychological process, characterised by specific equality constraints across the two gender groups, postulated by the measurement model necessarily produced the observed covariance matrices. A high degree of fit between the observed and estimated covariance matrices would only imply that the psychological process and the equality constraints that characterise it, postulated by the multigroup measurement model provided a plausible or valid (i.e. permissible) explanation for the observed covariance matrices (Brits, 2011; Diamantopoulos & Siguaw, 2000; Kelloway, 1998; Moyo, 2009).

Unfortunately the *ex post facto* correlational research design has its limitations. Kerlinger (1973) articulated three major limitations of *ex post facto* research designs. Firstly, its inability to manipulate independent variables causes the researcher to not have direct control over the variables; compared to experimental

[34] The research design depicts the intention to fit the multi-group SAPI measurement model in which the 20 latent personality dimensions had been operationalised via the170 items of the second experimental version of the SAPI

research designs where independent variables can be manipulated and controlled. Secondly, the lack of power to randomly assign participants to groups or to randomly assign treatments to groups. The participants may however be grouped as a result of self-selection according to their manifested levels of the observed variable (Moyo, 2009; Van Heerden, 2011). However, the danger with this limitation is that the observed relationship between the variables may be due to the manifestation of other unknown variables, rather than the causal relationship as per the designed intent. Since the researcher cannot say with certainty that the variance is only caused by the variables of interest, and not caused by unknown variables, it cannot be stated that the indicator variables provide error-free representations of the latent variables. The measurement model acknowledges this danger and makes provision for measurement error terms (δ_i) that can also cause variance in the indicator variables. Thirdly, there is the risk of improper and erroneous interpretation of results. Because the model represents only one plausible answer and correlations do not imply causation, the research design prevents casual inferences being made from statistically significant path coefficients.

The research problems in social sciences do not lend themselves to experimentation as they are usually not manipulatable, but do require some degree of controlled enquiry (Kerlinger & Lee, 2000). The latent personality dimensions in the current study were also not manipulatable. Hence despite the stated limitations, *ex post facto* correlational research design along with structural equation modelling as analysis technique was applied in the current study and careful consideration was given to those limitations (Moyo, 2009).

4.4. Statistical Hypotheses

As indicated earlier, the test developers' designed intent for the SAPI implies that certain personality constructs are hypothesised to influence test takers' scores on certain assessment items. The score that an individual obtains on specific SAPI items is therefore hypothesised to be due to specific latent personality constructs. These relationships are depicted in the measurement model. It has also been proposed that the permissibility of the inferences is determined by the extent to which the measurement model fits the data, and secondly the strength and statistical significance with which the items load on the latent variables. This aspect has been empirically investigated by Mouton (2017) via a single-group CFA study. Support for the construct validity of the SAPI had been obtained (Mouton, 2017). The current study built on the Mouton (2017) study by investigating whether the measurement model depicting the SAPI's design intention fitted the data of both male and female South African test takers and if so whether the hypothesised regression relationships between specific items and the latent personality dimensions they were designated to reflect differ across the two gender groups in terms of slope, intercept and/or error variance. The research design and the envisaged statistical analyses determine how the statistical hypotheses will be formulated. As indicated, the *ex post facto* correlational design with structural equation modelling was be used for this study. The statistical hypotheses were therefore formulated with the LISREL notational system (Du Toit, Du Toit, & Hawkins, 2001).

The overarching substantive hypothesis stipulated that the SAPI measurement model provided a valid description of the psychological process that determined test takers' responses to the SAPI items for both genders and that the parameters characterising the psychological process were the same across gender

groups. (Hair, Black, Babin, Anderson, & Tatham, 2006) The substantive research hypothesis translated into eight operational hypotheses.

Operational hypothesis 1 was tested by testing the exact fit null hypotheses (H_{01i}; i = 1_{Female}, 2_{Male}) which represented the stance that each of the single-group measurement models provided an accurate account of the manner in which the latent variables influenced the indicator variables in each gender group. The exact fit null hypothesis was formulated as follows, where Σ represented the observed population co-variance matrix and $\Sigma(\theta)$ represented the reproduced co-variance matrix obtained from the fitted model (Kelloway, 1998).

H_{01i}: $\Sigma = \Sigma(\theta)$; i = 1_{Female}, 2_{Male}
H_{a1i}: $\Sigma \neq \Sigma(\theta)$; i = 1_{Female}, 2_{Male}

Browne and Cudeck (1993) propose the following alternative formulation for the exact fit null hypothesis:

H_{01i}: RMSEA = 0; i = 1_{Female}, 2_{Male}
H_{a1i}: RMSEA > 0; i = 1_{Female}, 2_{Male}

Some authors regard the expectation that the measurement model will be able to reproduce the observed co-variance matrix to a degree of accuracy in the sample that can be explained in terms of sampling error only under H_{01} (i.e. a positive but statistically insignificant chi-square statistic) as unrealistic (Browne & Cudeck, 1993). Browne and Cudeck (1993) consequently propose testing the close fit null hypothesis, in which the stance is taken that the measurement model is an approximate account of the manner in which the latent variables influence the indicator variables.

H_{02i}: RMSEA ≤ .05; i = 1_{Female}, 2_{Male}
H_{a2i}: RMSEA > .05; i = 1_{Female}, 2_{Male}

Conditional on whether H_{01i} and/or H_{02i} could not be rejected for both i = 1_{Female}, 2_{Male}, a series of additional operational hypotheses (operational hypotheses 2-8) relating to the slope, intercepts and error variance of the regression for the items on the respective latent personality dimensions were evaluated.

Operational hypothesis 2 was tested by testing the close fit of the multigroup configural invariance model by testing H_{03}.

H_{03}: RMSEA ≤ .05
H_{a3}: RMSEA > .05

Operational hypothesis 3 was tested, conditional on whether H_{03} could not be rejected (i.e. conditional on a finding of configural invariance), by testing the close fit of the multigroup weak invariance model by testing H_{04}.

H_{04}: RMSEA ≤ .05
H_{a4}: RMSEA > .05

Rejection of H_{04} would imply the presence of non-uniform bias in one or more items. These items would identified by calculating and rank-ordering the absolute difference in the common metric completely standardised factor loadings obtained for the configural invariance multigroup model in Microsoft Excel. The item with the highest absolute difference in factor loading would then be allowed to be freely estimated in the partial weak invariance multigroup model. The close fit of the partial weak invariance model was tested by testing H_{041}: RMSEA ≤ .05. If H_{041} was still rejected the procedure would be continued by freeing the factor loading of the item with the second highest absolute difference in factor loading and testing H_{042}. This process would continue until H_{04i}; i < 170 was not rejected.

Operational hypothesis 4 was tested, conditional on whether H_{04} (or H_{04i}) could not be rejected (i.e. conditional on a finding of weak invariance or partial weak invariance), by testing the close fit of the multigroup strong invariance model (or the partial weak-strong invariance model) by testing H_{05}.

H_{05}: RMSEA ≤ .05
H_{a5}: RMSEA > .05

Rejection of H_{05} would imply the presence of uniform bias in one or more items. These items would be identified by calculating and rank-ordering the absolute difference in the unstandardised tau estimates obtained for the configural invariance multigroup model in Microsoft Excel. The item with the highest absolute intercept difference would be allowed to be freely estimated in the partial strong invariance multigroup model. The close fit of the partial strong invariance model would then be tested with H_{051}: RMSEA ≤ .05. If H_{051} was still rejected the intercept of the item with the second highest absolute intercept difference would be freed by testing H_{052}. This process would continue until H_{05i}; i < 170 was not rejected.

Operational hypothesis 5 was tested conditional on whether H_{05} (or H_{05i}) could not be rejected (i.e. conditional on a finding of strong invariance or partial strong invariance), by testing the close fit of the multigroup strict invariance model (or the partial strong-strict invariance model)[35] by testing H_{06}.

H_{06}: RMSEA ≤ .05
H_{a6}: RMSEA > .05

Rejection of H_{06} would imply the presence of error variance bias in one or more items. These items would be identified by calculating and rank-ordering the absolute difference in the common metric completely standardised error variance estimates obtained for the configural invariance multigroup model in Microsoft Excel. The item with the highest absolute difference in error variance would be allowed to be freely estimated in the partial strict invariance multigroup model. The close fit of the partial strict invariance model would be

[35] It is acknowledged that this still allows for some ambiguity since it does not clarify whether weak invariance had been obtained or partial weak invariance. A more unambiguous description would therefore be the weak-partial strong-strict invariance model or the partial weak-partial strong-strict invariance model.

tested by testing H_{061}: RMSEA ≤ .05. If H_{061} was still rejected the measurement error variance of the item with the second highest absolute difference in error variance and testing H_{062}. This process would be continued until H_{06i}; i < 170 was not rejected.

Conditional on whether $H_{0j;}$ j = 4, 5, 6 (or H_{0ij}; i < 170; j = 4, 5, 6) could not be rejected (i.e. conditional on a finding of weak, strong and strict invariance or partial invariance), operational hypotheses 6 – 8 on the SAPI measurement equivalence were evaluated by testing the practical significance of the difference in fit between the various multigroup invariance models and the configural invariance model.

Metric equivalence was tested by evaluating the practical significance of the difference in fit between the multigroup weak invariance model (or partial weak invariance model) and the configural invariance model, by reading off the Comparative Fit Index (CFI) from the standard LISREL fit statistics output and calculating the Gamma Hat Fit Index (Γ_1) and the McDonald Non-Centrality Index (Mc) for the multigroup weak invariance model (or partial weak invariance model) (the H_{04} model) and for the multigroup configural invariance model (the H_{03} model). Cheung and Rensvold (2002) argue that the difference in model fit may be considered practically insignificant[36] and that metric equivalence is demonstrated if a change of less than -0.01 in the CFI fit index, a change of less than -0.001 in the Gamma Hat fit Index (Γ_1) and[37] a change of less than -0.02 in the McDonald Non-Centrality Index (Mc) is found between the partially constrained multigroup weak invariance (H_{04}) model and the unconstrained multigroup configural invariance (H_{03}) model.

H_{07}: The multigroup weak invariance model and multigroup configural invariance model fits equally well in the parameter

H_{a7}: The multigroup weak invariance model fits poorer than the multigroup configural invariance model in the parameter

Scalar equivalence was tested by evaluating the practical significance of the difference in fit between the multigroup strong invariance model (or the partial strong invariance model) and the configural invariance model, by reading off the Comparative Fit Index (CFI) from the standard LISREL fit statistics output and calculating the Gamma Hat Fit Index (Γ_1) and the McDonald Non-Centrality Index (Mc) for the multigroup strong invariance model (or partial strong invariance model) (the H_{05} model) and for the multigroup configural invariance model (the H_{03} model). The Cheung and Rensvold (2002) decision rule was also used to evaluate the practical significance of the difference in model fit brought about by the equality constraints.

H_{08}: The multigroup strong invariance model and multigroup configural invariance model fits equally well in the parameter

[36] The current study argues that in the case of testing for practical significance no statistical hypotheses in the conventional sense are formulated since the decision-rule is not based on the calculation of the probability of observing the sample findings conditional on the parametric assumption made under the null hypothesis. The Cheung and Rensvold (2002) decision rule is rather based on the findings of Monte Carlo simulations when drawing samples from populations where the H_0 and H_a models fit equally well. Nonetheless, when the Cheung and Rensvold (2002) criteria have been met in the sample then the inference was made that in general the two multi-group models do not differ in fit. Hence, null hypotheses were formulated but they are not true statistical hypotheses that posit a value for a parameter.

[37] The current study regarded the difference in fit between the multi-group configural invariance measurement model and the multi-group weak (or strong or strict) measurement model to be practically insignificant if all three criteria proposed by Cheung and Rensvold (2002) have been met.

H_{a8}: The multigroup strong invariance model fits poorer than the multigroup configural invariance model in the parameter

Conditional probability equivalence was tested by evaluating the practical significance of the difference in fit between the multigroup strict invariance model (or the partial strict invariance model) and the configural invariance model, by reading off the Comparative Fit Index (CFI) from the standard LISREL fit statistics output and calculating the Gamma Hat Fit Index (Γ_1) and the McDonald Non-Centrality Index (Mc) for the multigroup strict invariance model (or partial strict invariance model) (the H_{06} model) and for the multigroup configural invariance model (the H_{03} model). The Cheung and Rensvold (2002) decision rule was also used to evaluate the practical significance of the difference in model fit brought about by the equality constraints.

H_{09}: The multigroup strict invariance model and multigroup configural invariance model fits equally well in the parameter

H_{a9}: The multigroup strict invariance model fits poorer than the multigroup configural invariance model in the parameter

If the difference in fit between the weak, strong and strict invariance models (or partial invariance models) on the one hand and the configural invariance model on the other, would be found to be practically significant and H_{07}, H_{08} and/or H_{09} would be rejected, the source of the lack of metric, scalar and/or conditional probability equivalence would be identified in essentially the same manner that was used to identify the source of a lack of weak, strong and/or strict invariance. In the case of measurement equivalence, however, the procedure would continue until the difference would no longer be practically significant as judged by the Cheung and Rensvold (2002) criteria.

Based on the results from these statistical hypotheses, the extent to which the SAPI successfully measures the personality dimensions in an unbiased manner across both gender groups, in accordance with the scoring key, were established.

4.5. Sample

Strydom (2014) insists that whenever gender is a relevant characteristic to research, the male and female groups should be representative of the population. To ensure that a sample group is representative of the population from which it is drawn, the ideal sampling technique is random sampling (Strydom, 2014). Random sampling requires the researcher to have a list (or sampling frame) to randomly select individuals from. Without such a list for the current research study, the researcher applied convenience sampling which is a type of nonprobability sampling technique that refers to the reliance on available subjects to provide data for the purposes of research studies (Tabachnick & Fidell, 2007). The SAPI developers provided institutional permission to use archival SAPI data for this study, causing the sample to be regarded as a non-probability sample of respondents representing both gender groups from the South African population.

Although the large sample size (n = 4254) implies that if the measurement model implied by the design of the SAPI fits the data well, it would be concluded that the study results are relevant yet limited evidence of

whether or not the instrument demonstrates measurement invariance and measurement equivalence. Therefore it is acknowledged that the findings of this study should be cautiously generalised to the general South African population, since convenience sampling does not allow for control over the representativeness of a sample (Babbie, 2010).

4.6. Statistical Analyses

Statistical analyses involve preparing the data and conducting the relevant analyses.

4.6.1. Preparatory Procedures

The preparation of the data included specifying the to-be-fitted measurement models, evaluating the identification of the models, investigating the statistical power of the study, and deciding on the manner in which missing data was handled.

4.6.1.1. Model specification

The measurement model was specified as follows in SEM notation. This specification allowed for an understanding of the model complexity and the identity of parameters to be estimated (Holtzkamp, 2013).

To test the null hypotheses H_{01i};i = 1_{Female}, 2_{Male}, and H_{02i};i = 1_{Female}, 2_{Male}, the single-group measurement model as displayed by Equation 1, was fitted to the data of the two gender groups separately:

$$\mathbf{X} = \boldsymbol{\tau} + \Lambda^{x}\xi_{i} + \boldsymbol{\delta} \quad \text{.....(1)}$$

Where:

- $\mathbf{X}$ = 170 x 1 column vector of observable item scores;
- $\boldsymbol{\tau}$ = 170 x 1 column vector of the intercept terms;
- $\boldsymbol{\Lambda}^{x}_{i}$ = 170 x 20 matrix of factor loadings;
- ξ = 1 x 20 column vector of latent first-order personality dimensions;
- $\boldsymbol{\delta}$ = 170 x 1 column vector of unique or measurement error components consisting of the combined effect on X of the systematic non-relevant influences and random measurement error (Jöreskog & Sörbom, 1993).

Single-group measurement equations for invariance studies require two additional matrices, in comparison to evaluating single-group measurement models in construct validation studies such as the preceding study by Mouton (2017). The first is a symmetrical variance-covariance matrix $\boldsymbol{\Phi}_{i}$ that describes the variance in and correlations between the latent variables. The current study estimated all unique variance and covariance elements in $\boldsymbol{\Phi}$. The main diagonal in $\boldsymbol{\Phi}$ was freed to be estimated by fixing the first factor loading for each of the 20 latent first-order personality dimensions to 1[38]. The second is a diagonal variance-covariance matrix $\boldsymbol{\Theta}_{\delta i}$ which depicts the variance in error terms associated with indicator variables. The diagonal nature of the $\boldsymbol{\Theta}_{\delta i}$ matrix indicates that the error terms δ_{i} are assumed to be uncorrelated across the indicator variables

[38] The appropriateness of freeing the ϕ_{ii} elements in $\boldsymbol{\Phi}$ by fixing the first factor loading for each latent variable should be critically examined. Doing so makes sense if the testing of complete invariance is warranted. Freeing the ϕ_{ij} elements in $\boldsymbol{\Phi}$ then allow for possible differences in ϕ_{ij} across groups. Differences in the ϕ_{ij} elements in $\boldsymbol{\Phi}$ do not, however, hold any implications for measurement bias under the current ((lenient or stringent) definitions of item bias. Fixing the first factor loading for each latent variable creates the danger that possible non-uniform bias in the reference items may go undetected without having any relevance from a bias perspective. The point raised previously under footnote 32 could, however, counter this argument.

(Donnelly, 2009). Freeing the off-diagonal elements of $\Theta_{\delta i}$ would imply that the error terms may be correlated and therefore allow for possible additional common factors that are not reflected in the model as defined by the test developers, but that also causes the response to the indicator variables (Dunbar-Isaacson, 2006; Holtzkamp, 2013). Hence, freeing the off-diagonal elements of $\Theta_{\delta i}$ could not be substantively justified for this study.

To test the null hypotheses H_{03} and H_{0j}; j = 4,5,6,7 the multigroup measurement model as displayed by Equation 2, was fitted to the combined data of both groups:

$$\mathbf{X}^g = \boldsymbol{\tau}^g + \boldsymbol{\Lambda}^{xg}\boldsymbol{\xi}^g + \boldsymbol{\delta}^g \quad \text{.....(2)}$$

Where:

- $\mathbf{X}^g$ = 170 x 1 column vector of observable item scores for group g; g =1 $_{Female}$, 2_{Male}
- $\boldsymbol{\tau}^g$ = vector of the intercept terms; g = 1_{Female}, 2_{Male}
- $\boldsymbol{\Lambda}^{xg}_i$ = 170 x 20 matrix of factor loadings for group g; g = 1_{Female}, 2_{Male}
- $\boldsymbol{\xi}^g$ = 1 x 20 column vector of latent first-order personality dimensions; g = 1_{Female}, 2_{Male}
- $\boldsymbol{\delta}^g$ = 170 x 1 column vector of unique or measurement error components for group g consisting of the combined effect on X of the systematic non-relevant influences and random measurement error; g = 1_{Female}, 2_{Male} (Jöreskog & Sörbom, 1993).

The symmetrical variance-covariance matrices $\boldsymbol{\Phi}^g$ also described the variance in and covariance between the latent variables. The diagonal variance-covariance matrix $\boldsymbol{\Theta}^g_\delta$ depicted the variance in error terms associated with indicator variables. The diagonal nature of the $\boldsymbol{\Theta}^g_\delta$ matrix indicated that the error terms were assumed to be uncorrelated across the indicator variables (Holtzkamp, 2013).

Although this was not explicitly acknowledged in previous measurement invariance and equivalence studies Equations 1 and 2, along with the description of $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}_\delta$ still did not fully specify the measurement model. Measurement models can differ in terms of the assumptions made about the elements of $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}^x$, and $\boldsymbol{\Theta}_\delta$. More specifically measurement models can differ in terms of the extent to which they constrain the elements of $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}^x$, and $\boldsymbol{\Theta}_\delta$ to be equal across items of a subscale. Graham (2006) distinguishes between the following four measurement models:

- The classically parallel model;
- The tau-equivalent model;
- The essentially tau-equivalent model; and
- The congeneric model

The congeneric model allows the elements of $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}^x$ and $\boldsymbol{\Theta}_\delta$ to be freely estimated across the indicators of each latent variable. “The congeneric model assumes that each individual item measures the same latent variable, with possibly different scales, with possibly different degrees of precision, and with possibly different amounts of error (Raykov, 1997a). Whereas the essentially tau-equivalent model allows item true scores to differ by only an additive constant, the congeneric model assumes a linear relationship between item true scores, allowing for both an additive and a multiplicative constant between each pair of item true

scores" (Graham, 2006, p. 935). The congeneric measurement model assumes that the regression of X_i on ξ_j differs in terms of intercept, slope and error variance across the indicators of the same (unidimensional) latent variable. Both the two single group measurement models, as well as the four multigroup invariance measurement models have been fitted as congeneric measurement models.

4.6.1.2. Model Identification

Model identification allows researchers to investigate whether sufficient information is available to attain a unique solution for the freed parameters to be estimated in the model, prior to fitting the model to sample data (Diamantopoulos & Siguaw, 2000). Two important model specification requirements are proposed by Diamantopoulos and Siguaw (2000), along with MacCallum (1995). Firstly, each latent variable should be allocated a definite scale. Secondly, the number of model parameters to be estimated may not exceed the number of unique variance/covariance terms in the sample observed covariance matrix or matrices in the case of multigroup models and indicator variable means (Diamantopoulos & Siguaw, 2000; MacCallum, 1995). This implies that the model should have positive degrees of freedom.

The measurement model that is depicted in both Equation 1 (single-group measurement model) and Equation 2 (multigroup measurement model) satisfies both these requirements. Addressing the first requirement, a definite scale will be allocated to each latent variable by fixing the factor loading of the first indicator variable of each latent variable to unity. The latent variable scale will therefore be set to be equal to that of the first indicator variable of each subscale. The latter requirement is met with positive degrees of freedom for each measurement model to be tested, as depicted in Table 4.1

The number of unique variance/covariance terms in the sample observed covariance matrix (or matrices) and indicator variable means, will remain more than the number of model parameters to be estimated for each measurement model to be tested.

Table 4.1
Degrees of Freedom for the Single-Group and Multigroup measurement Invariance Models

Hypothesis	Lambda	Tau	Theta-Delta	Phi's	Total Parameters To Be Estimated	Indicator Variables	Groups	Unique Information Pieces	*df*
Single Group: Female	150[39]	170	170	210[40]	700	170	1	14705[41]	14005
Single Group: Male	150	170	170	210	700	170	1	14705	14005
Configural Invariance [H_{03}]	300	340	340	420	1412	170	2	29410	28010
Weak Invariance [H_{04}]	150	340	340	420	1250	170	2	29410	28160
Strong Invariance [H_{05}]	150	170	340	420	1080	170	2	29410	28330
Strict Invariance [H_{06}]	150	170	170	420	910	170	2	29410	28500

[39] The SAPI measures 20 latent first-order personality dimensions via 20 subscales of items. The factor loading of the first item is not estimated but fixed to 1 to set the scale of the latent personality dimension (i.e. 170 – 20 = 150).
[40] Both the latent variable variances and covariances are estimated.
[41] Since the measurement models are fitted via a means and covariance structure (MACS) analysis the 170 indicator means along with the 170 indicator variances and the 14365 covariances constitute the available unique pieces of information from which the parameter estimates have to be derived.

4.6.1.3. Statistical Power

Statistical power refers to the probability of rejecting an incorrect SEM model, i.e. rejecting the null hypothesis given that it is false, as in the case of the exact fit null hypothesis P(Reject H_{01}: RMSEA = 0|H_{01} false). Statistical power plays an important role in the ensuring that the outcome of testing the null hypothesis (i.e. rejecting the null hypothesis or not) can be unambiguously interpreted. For example, if the null hypothesis could not be rejected, the cause thereof may be rooted in either the reflection of the actual situation in the sampled population, or due to the lack of statistical power. Diamantopoulos and Siguaw (2000) pointed out that sample size plays a critical role in the level of statistical power and consequently in the decisions regarding the outcome of the tested models. Statistical power decreases along with a reduction in sample sizes. Small samples cause low statistical power, and raises the question as to whether the decision not to reject the model (i.e. not to reject H_0) is due to the model being accurate or whether the test is too insensitive to detect specification errors in the model. On the other hand, large sample sizes are not necessarily the answer to this predicament, since the resulting high level of power may cause the null hypothesis to be rejected due to minor specification errors in the model.

Tabachnick and Fidell (2007) advocate that statistical power level is determined prior to conducting any statistical analyses. They recommend a desired level of statistical power of at least .80, which indicates that the researcher has a probability of at least 80% to achieve a significant result if an effect exists (i.e. if the model fits poorly). To obtain power estimates for the test of close fit of the single-group measurement models[42], the following values were captured into the syntax that Preacher and Coffman (2006) (Preacher & Coffman, 2006) developed in R (available at http://www.quantpsy.org/rmsea/rmsea.htm): the sample size of 2420 female responses and 1834 male responses[43] and the degrees of freedom for each of the measurement models to be tested, as presented in Table 4.1

A RMSEA of .05 was used for H_{02} and RMSEA values of .08 and .06 for $H_{a2.}$ The Preacher and Coffman (2006) syntax returned a power value of unity, which was considered by Mouton (2017) to be very high, for each of the respective measurement models to be tested. As a result, a finding of invariance (i.e. an outcome of not rejecting H_0) for the tested measurement models could be confidently and unambiguously interpreted as evidence of a lack of bias. Conversely, however, a finding of a lack of invariance for any of the tested models would raise the concern that such an outcome could be due to excessive statistical power.

Table 4.2
Statistical power for the Single-Group Measurement Invariance Models

MODEL/HYPOTHESIS	Alpha	RMSEA H_0	RMSEA H_a	N	*df*	Power
Single Group: Male [H_{01}]	.05	.05	.08	1834	14005	1
Single Group: Female [H_{02}]	.05	.05	.08	2420	14005	1
Single Group: Male [H_{01}]	.05	.05	.06	1834	14005	1
Single Group: Female [H_{02}]	.05	.05	.06	2420	14005	1

[42] The Preacher and Coffman software does not provide the option to evaluate the power of multi-group measurement models.
[43] The current study was fortunate enough not to have to go out and collect SAPI data from scratch but was given access to an archival SAPI data base. The power calculation was therefore not performed with the sample size as the unknown and desired power fixed on .80.

4.6.1.4. Treatment of missing values

The SAPI data that was used in the current study was collected electronically via an online questionnaire. The electronic version does no offer the option of "unable to respond" and requires that all items are completed before allowing the test taker to move to a subsequent test page. The SAPI data set therefore contained no missing values.

4.6.2. Evaluation of the SAPI Measurement Model

Evaluating the various single- and multigroup measurement models referred to in paragraph 4.2 and paragraph 4.4 required clarifying the variable type, exploring the degree to which the model fitted the data, and evaluating the measurement invariance and equivalence through a sequence of analyses.

4.6.2.1. Variable Type

The SAPI uses a five-point Likert-type response scale, for respondents to indicate their level of agreement towards each item. This type of response scale produces ordinal data. Strictly speaking this would require that the confirmatory factor analysis (CFA) on both gender samples had to be performed by analysing the matrix of polychoric correlation coefficients using Diagonally Weighted Least Squares estimation (DWLS) (Jöreskog & Sörbom 1996). Maximum Likelihood estimation (ML) rather than DWLS provides for a more powerful analysis. ML estimation, however, requires the data to be continuous and the analysis of the covariance matrix. This benefit of ML estimation and the popularity of Likert-type response scales, lead Muthén and Kaplan (1985) to permit the specification of ordinal data obtained from Likert scales that use five or more scale points as (approximating) continuous data.

An alternative method to convert the SAPI data to continuous data is the use of item parcelling instead of the single items. Item parcelling offers several advantages when compared to using single items (Little, Cunningham, Shahar, & Widaman, 2002). Firstly, item parcels address several data-related problems such as non-normality, insufficient sample sizes, inadequate ratio for the sample size as compared to the variables, and instability of parameter estimates (Bandalos, Gerke, & Finney, 2001). Secondly, Dunbar-Isaacson (2006) states that item parcels' composite score tends to be more reliable than single item scores. This is mainly due to the lower levels of skewness and kurtosis, and higher validity for item parcels (especially items with two or more items) when compared to single items (Marsh, Hau, Balla, & Grayston, 1998; Mavondo, Gabbott, & Tsarenko, 2003). As a third advantage, Bandolos (2002) states that an increase in the number of unidimensional items per item parcel, leads to an improvement of the model fit indices, particularly root mean squared error of approximation (RMSEA), compared fit index (CFI), and the chi-square test, that are all relevant to measurement invariance and equivalence analyses.

Despite these advantages, item parcelling also poses several disadvantages. When measuring multidimensional constructs item parcels consequently tend to be multidimensional as well, which causes problems with interpreting the item parcel results (Dunbar-Isaacson, 2006). The improved model fit when using item parcels as compared to single items, is likely caused by random and systematic error being cancelled out due to the aggregation of these errors. The resulting problem is that item parcelling can potentially conceal model misspecifications (Meade & Kroustalis, 2006), thereby reducing the probability of

detecting such misspecifications and consequently increasing the probability of failing to reject models that should have been rejected (Type II errors) (Little et al., 2002). More importantly, from a measurement bias perspective parcelling was regarded as an unattractive option in the current study because:

- It creates the possibility that biased items may hide in item parcels (i.e. the item parcel is not flagged as a biased measure) because their effect is washed out/diluted in the composite item parcel;
- It creates the possibility that biased items may hide in item parcels (i.e. the item parcel is not flagged as a biased measure) because their effects tend to cancel each other out in the composite item parcel score;
- It (therefore) creates ambiguity and uncertainty when item parcels are not flagged as biased;
- It prevents individual biased items from being identified when item parcels are flagged as biased.

Agreeing with Mouton (2017), the methodological ideal would be to use single items when fitting a measurement model with the aim of evaluating the construct validity of an instrument. The same position applied to studies aimed at evaluating construct and item bias via multigroup SEM. However, the estimated model parameters for the single-group measurement model in which the latent personality dimensions were operationalised via individual items were 530, consisting of 170 factor loadings, 170 measurement error variances and 190 unique covariance terms. The requirement that the number of observations within each gender group should have preferably been 5-10 times more than the number of freed parameters in the model (i.e. 2650 – 5300) placed several demands on the research. Firstly, large sample sizes were required. Secondly, a measurement model with such a large number of freed model parameters required a large amount of computer memory for the LISREL 8.8 software to run successfully even when the syntax was run in batch mode from the disk operating system. The amount of memory is dramatically increased when the data fails to satisfy the multivariate normality assumption and robust maximum likelihood estimation (rather than ML estimation) is required. According to Mels LISREL 8.8 unfortunately assigns processing memory in a very inefficient manner (Gerhard Mels, personal communication, 7 August 2018). The problem is that LISREL 8.8 does not use a dynamic allocation of processing memory so that memory that had been assigned to some calculation at an earlier point in time cannot be freed up for any subsequent calculations. The consequence then typically is (as was the case in the current study) that there is insufficient memory capacity to calculate the inverse of the asymptotic covariance matric and the Satorra-Bentler chi-square statistic that needs to be calculated under robust maximum likelihood estimation (but not ML estimation) (Gerhard Mels, personal communication, 13 September 2018).

Mouton (2017) admitted that although the ideological ideal remained using individual items to represent the latent variables when fitting a measurement model, the resulting constraints that would be placed on computer memory and processing time, forced the use of item parcels. The same constraints persisted in this current study. This left the current research study with three possible solutions to circumvent the problem. All of them, however, required a compromise on the ideal to evaluate the fit of the SAPI single- and multigroup measurement models in which the 20 latent first-order personality dimensions have been operationalised via the individual items of the SAPI via robust maximum likelihood (RML) estimation[44].

[44] It was considered extremely unlikely that the multivariate null hypothesis would not be rejected for the male and female SAPI data sets.

The first, probably most conventional, option was to reduce the number of indicator variables by forming item parcels. The number of parcels that had to be formed to solve the memory capacity problem would have to be determined via trial and error by staring of with the largest number of smallest possible parcels and gradually reducing the number of parcels by increasing the number of items per parcel. This would gradually reduce the dimensions of the asymptotic covariance matric and at some point allow the successful inversion of the asymptotic covariance matrix and the calculation of the Satorra-Bentler chi-square statistic. The disadvantage of this option was that it does not allow the evaluation of item bias on the level of individual items, although this is the primary interest of a measurement bias study. The evaluation of item bias on the level of item parcels was previously criticised when it was considered as a method to ensure that the CFA was performed on continuous data.

The second option was to reduce the demand for computer memory by simplifying the required calculations by eliminating the inversion of the asymptotic covariance matrix and the calculation of the Satorra-Bentler chi-square statistic from the analysis procedure. The second option was therefore to not use RML estimation to fit the single- and multigroup SAPI measurement models in which the individual items are used as indicators despite knowing that the multivariate assumption is not satisfied, but rather ML estimation to derive estimates[45]. The advantage of this approach was that it allows the fitting of measurement models with individual items. The disadvantage was that one knowingly and intentionally uses an inappropriate estimation technique to derive the parameter estimates for the freed measurement model parameters when the individual item distribution does not follow a multivariate normal distribution in the parameter. The inappropriate use of ML estimation can produce bias in the normal theory chi-square fit statistic estimate and the standard error estimates used to evaluate the statistical significance of parameter estimates. Mîndrilă (2010, p. 61) presents the following argument regarding the use of ML estimation on non-normal data.

> With non-normal continuous data, ML produces relatively accurate parameter estimates, but the bias in chi-square and standard errors increases with nonnormality [Bollen, 1989]. Even when the model is correctly specified, the use of ML in conditions of multivariate non-normality results in inflated chi-squares, particularly when the data have a leptokurtic distribution [Browne, 1984]. Consequently, fit indices such as the Tucker-Lewis Index (TLI), the root-mean square error approximation (RMSEA), and the comparative fit index (CFI), which are functions of chi-square, are also biased. Although ML produces accurate parameter estimates with non-normal continuous data, the standard errors are underestimated, especially when data are leptokurtic. [Hoogland & Boomsma, 1998]. Due to the discrete nature of categorical data, some authors consider it to be inherently non-normal [Muthen & Kaplan, 1985]. However, when ordinal data have a large number of categories and are approximately normal, ML does not produce severely biased results. Bias tends to increase as the number of response categories decreases, and multivariate non-normality increases. Because ML computational procedures are based on Pearson product-moment (PPM) correlational techniques, when the number of response categories is small, the fit indices, parameter estimates, and standard errors can be biased. [Finney & DiStefano, 2006] When data are both ordinal and non-normal, using ML inflates the chi-square and the root mean square residual (RMR), and underestimates the non-normed fit index (NNFI), and

[45] The use of Diagonally Weighted Least Squares estimation (DWLS) would not have offered a solution to the memory capacity problem since DWLS also requires the calculation of the asymptotic covariance matrix.

> the goodness of fit index (GFI). Furthermore, the bias in parameter estimates and standard errors increases when data are skewed or kurtotic, when there are few response categories, the sample is small, or the relationships between factors and indicators are weak [Babakus, Ferguson & Jöreskog, 1987]. Because of the assumption of multivariate normal distribution, it is generally recommended to use ML only when the violations of multivariate normality are only slight. Additionally, ML can be used with ordinal data only if variables can take at least 5 different values, and they are treated as continuous when computing the correlation or covariance matrix [Schumacker & Beyerlein, 2000].

The SAPI utilises a 5-point Likert scale. Strictly speaking therefore the SAPI data should be regarded as ordinal data. Muthen and Kaplan (1985), like Mîndrilă (2010) suggest that ordinal Likert scale data may be treated as approximating continuous data when the number of response options is 5 or more[46]. The danger of using ML estimation on non-normal data depends on the extent to which the item distributions excessively deviate from symmetrical and mesokurtic distributions. The descriptive statistics for the SAPI items for the female and male samples are shown in Appendix A. Table 4.3 provides a summary of the trends presented in Appendix A.

Table 4.3

Summary of the symmetry and kurtosis of the SAPI items

		Male	Female
Symmetry	Negatively skewed (p<.05)	129 (75.88%)	131 (77.06%)
	Symmetric	8 (4.71%)	4 (2.35%)
	Positively skewed (p<.05)	67 (39.41%)	35 (20.59%)
	Highly skewed (positively or negatively)	12 (7.06%)	19 (11.18%)
Kurtosis	Platikurtic (p<.05)	31 (18.24%)	108 (63.53%)
	Mesokurtic	17 (10%)	24 (14.12%)
	Leptokurtic (p<.05)	122 (71.77%)	38 (22.35%)

Table 4.3 indicates that the majority of the item distributions were statistically significantly (p<.05) skewed in both the male (162 (95.29%)) and female (166 (97.65%)) samples. However, only a small percentage of the total number of the item distributions was highly skewed in the male (7.06%) and female (11.18%) samples. Table 4.3 indicates that the majority of the item distributions statistically significantly (p<.05) deviated from a mesokurtic distribution in both the male (153 (90.0%)) and female (146 (85.88%)) samples. Given the item statistics summarised in Table 4.3 it would be naïve to deny that there is danger in using ML estimation. Table 4.3 nonetheless does not depict a situation in which the item distributions excessively deviate from symmetrical and mesokurtic distributions.

The third option was to reduce the size of the indicator variable data set whilst retaining the advantage of operationalising the latent first-order personality dimensions via the individual SAPI items. This could be

[46] In a subsequent paper Muthen and Kaplan (1992) expanded on their (1985) study by examining the impact of non-normal Likert variables on testing and estimation in CFA for models of various sizes. They compared normal theory GLS and the ADF estimator for six cases of non-normality, two sample sizes, and four models of increasing size in a Monte Carlo framework with a large number of replications. Results showed that GLS and ADF chi-square tests are increasingly sensitive to non-normality when the size of the model increased. They did not include the ML estimator in this study but presumably model size has a similarly negative effect on the ML estimator.

achieved by dividing the SAPI first-order measurement model comprising 20 latent variables and 170 item indicators into 5 separate measurements models. Each measurement model would then map the latent first-order personality dimensions (that load on one of the Big Five second-order personality factors) onto the individual SAPI items designated to reflect these first-order factors. The disadvantage of this option was that it increased the number of single and multigroup measurement models that had to be fitted fivefold. More importantly, it would dissect the SAPI into five different measurement models. This in turn had the disadvantage that no single verdict on construct bias would be attained for the SAPI. More importantly the measurement hypothesis made by the SAPI lies in the freed elements of the 170 x 20 Λ^{x} matrix, the 170 x 170 Θ_{δ} matrix and the 20 x 20 Φ matrix. Testing 5 separate SAPI measurement models is not the same and will not render the same results as testing the full SAPI measurement model.

The discussion of the three options clearly illustrate that the current study was left with a major dilemma. The current study chose to use the second option described in the aforementioned section. The current study concedes that this choice brought with it non-ignorable methodological limitations.

4.6.2.2. Measurement Model Fit

The single-group SAPI measurement model represents the design intention of the SAPI developers to have specific indicator variables reflect the specific latent personality dimensions comprising personality as conceptualised by the SAPI. Alternatively stated, the single-group SAPI measurement model reflects the scoring key of the SAPI. The various multigroup SAPI measurement models reflect the design intention of the SAPI developers to develop an unbiased measure of personality. The extent to which these design intentions has succeeded is reflected in the ability of the various single- and multigroup models to reproduce the observed covariance matrix/matrices. When the reproduced covariance matrix/matrices approximate the observed covariance matrix/matrices, the conclusion can be made that the model fits well.

The range of goodness of fit indices, as provided by LISREL (Diamantopoulos & Siguaw, 2000), were interpreted to examine the single-group measurement model fit. The magnitude and distribution of the standardized residuals and the magnitude of the model modification indices that were calculated for Λ_{x}^{g} and Θ_{δ}^{g}, were examined to assess and comment on the quality of the model fit. The modification indices were examined for statistically significant ($p<.01$) values, as these model parameters indicate an improvement in model fit if set free to be estimated. The number and significance of large modification indices were also examined, as these characteristics reflect negatively on model fit, in that they suggest multiple possibilities to improve the model. The multigroup measurement model fit was evaluated by testing the close fit null hypothesis H_{0j}; j = 3, 4, 5, 6.

To meet the objective of this study, to investigate the measurement invariance and measurement equivalence of the SAPI, LISREL 8.8 (Du Toit, Du Toit, & Hawkins, 2001; Jöreskog & Sörbom, 1993) was used to determine the fit of: (i) the single-group measurement model on both gender groups and (ii) the four multigroup measurement models when fitted in a series of multigroup analyses.

4.6.2.3. Testing for measurement invariance and measurement equivalence

This study applied a specific sequence of measurement invariance and equivalence tests as stipulated by Dunbar et al. (2011), to answer a series of research questions that examine the extent to which the

measurement model may be considered measurement invariant and equivalent or not, as well as identify the source of the variance should it exist (Vandenberg & Lance, 2000). The sequence and logic thereof, for the series of tests that Dunbar et al. (2011) recommend when investigating measurement invariance and equivalence is explained as follows.

Step 1: Establish whether the single-group measurement model displays acceptable fit when it is fitted independently to each gender sample.

Dunbar et al. (2011) regard a reasonable fit for the measurement model to each sample group as a prerequisite for evaluating the measurement model in a multigroup analysis. Diamantopoulos and Siguaw (2000) recommend that the Root Mean Square Error of Approximation (RMSEA) is used to evaluate the extent to which the measurement model fits the sample data. RMSEA values below .05 indicate good model fit, whilst RMSEA values above .05 but less than .08 indicate reasonable fit, values above .08 but less than .10 indicate mediocre fit, and lastly values above .10 indicate poor model fit.

Therefore, before the multigroup measurement models could be investigated for measurement invariance and equivalence, the measurement model first needed to be fitted independently to both the female and the male group. Rejecting the null hypothesis of close fit (H_{02i} RMSEA $\leq$ .05;i = 1_{Female}, 2_{Male}) would indicate that the measurement model did not demonstrate adequate fit for either or both of the male or female groups in the parameter. Such an outcome would cause any further examination of the measurement invariance and equivalence to be questionable and resulting in the process to be terminated (Dunbar et al., 2011).

Step 2: Establish whether the multigroup measurement model in which the model structure is constrained to be the same across groups with no freed parameters that are constrained, displays acceptable fit when fitted to the samples in a multigroup analysis. Alternatively stated, establish whether the multigroup configural invariance model displays reasonable fit.

Configural invariance places the emphasis on the theoretical structure of an instrument (Holtzkamp, 2013). Demonstrating configural invariance would permit the stance that the SAPI measures the same underlying construct for males and females. Configural invariance would be indicated if the close fit null hypothesis (H_{03}: RMSEA $\leq$.05) was not rejected. Configural invariance would therefore indicate the absence of construct bias. However, finding a lack of configural invariance would indicate that the SAPI measures different personality constructs across the two groups. The sequential nature of the analyses made the demonstration of configural invariance a prerequisite for the subsequent measurement invariance and equivalence tests (Dunbar et al., 2011). It makes logical sense to entertain the question whether an instrument measures a specific construct in the same manner across gender groups only if the instrument measures the same construct in both groups. Hence, with a lack of configural invariance, all subsequent measurement invariance and equivalence tests would be unnecessary, since this particular model forms the basis against which the other models would be compared (Vandenberg & Lance, 2000). Possible reasons for not achieving configural invariance include data collection problems, translation errors, when the constructs are seemingly culture specific, or when the various groups hold different frames of reference when responding to the items (Cheung & Rensvold, 2002; Holtzkamp, 2013)

Step 3a: Establish whether the multigroup measurement model in which the model structure is constrained to be the same across groups and in which all parameters are estimated freely across the samples, except for the slope of the regression of the indicator variables on the latent variables which is constrained to be equal, displays acceptable fit when fitted to the samples in a multigroup analysis. Alternatively stated, establish whether the multigroup weak invariance model displays reasonable fit

After demonstrating configural invariance, the multigroup weak invariance model may be fitted to the data. This analysis allowed the researcher the opportunity to investigate the similarity of factor loadings of the items on the latent variables. Weak invariance involves testing the null hypotheses of close fit (H_{04}: RMSEA ≤ .05), which investigates whether the slope of the regression of items on the latent variables that they represent, differ between the male and female groups. Whereas the configural invariance model explores similarity in theoretical structure between sample groups, the weak invariance model refers to whether the different samples perceive and interpret item content in the same manner (Byrne & Watkins, 2003).

Demonstrating weak invariance indicates that the multigroup weak invariance model is able to closely reproduce the observed covariance matrices (i.e. H_{04} is not rejected). Dunbar et al. (2011) insist that finding weak invariance supports the test developers' claim that the factor loadings are the same across different samples (i.e. that the items do not display non-uniform bias). As pointed out previously, it is not unreasonable to expect that differences in the slope of the regression of specific items on specific latent dimensions exist between different sample groups. If H_{04}: RMSEA ≤ .05 was rejected due to a few factor loadings with significant differences, partial weak invariance would be investigated. The items suffering from non-uniform bias would be identified via the previously described procedure (see paragraph 3.4.2.3). A finding of a lack of weak invariance (but a finding of partial weak invariance) would constitute strong evidence of non-uniform bias in specific items. A finding of weak invariance would however result in weak evidence of the absence of non-uniform bias.

Although proving weak invariance adds to evidence that the claim for similar factor loadings is a tenable position, it does not mean that differences in some factor loadings between the sample groups, is not a more tenable position (Holtzkamp, 2013). Once the outcome of weak invariance or partial weak invariance was found, metric equivalence could be tested.

Step 3b: Establish whether the multigroup measurement model in which the model structure is constrained to be the same across groups and in which all parameters are estimated freely across the samples, except for the slope of the regression of the indicator variables on the latent variables which is constrained to be equal, fits the multigroup data practically significantly poorer than a multigroup measurement model in which only the structure of the model is constrained to be the same across groups but all the parameters are estimating freely. Alternatively stated establish whether the multigroup measurement model displays metric equivalence by investigating whether the multigroup weak invariance model (or the multigroup partial weak invariance model) displays a practically significantly poorer fit than the multigroup configural invariance model.

Metric equivalence is demonstrated if the multigroup weak invariance model does not fit practically significantly poorer than the multigroup configural invariance model. In other words, metric equivalence is indicated when the change[47] from the configural invariance model to the weak invariance model reflects the following practical significance results[48] (Cheung & Rensvold, 2002):

- A change of -.01 or less in the CFI fit index;
- A change of -.001 or less in the Gamma Hat fit index (Γ_1); and
- A change of -.02 or less in the McDonald Non-centrality index.

Again it is not unreasonable to expect that (more subtle) differences in the slope of the regression of specific items on specific latent dimensions exist between different sample groups. If the three Cheung and Rensvold (2002) criteria have not been met due to a few factor loadings with significant differences, partial metric equivalence would be investigated by freeing (further) factor loadings in group 2 until the three Cheung and Rensvold (2002) criteria were met. A finding of partial metric equivalence would demonstrate weak evidence of non-uniform bias in specific items. A finding of metric equivalence in turn would deliver strong evidence of the absence of (further) non-uniform bias.

The Satorra-Bentler scaled difference test statistic's sensitivity to sample sizes may cause the value to be statistically significant despite the differences in model fit between the groups might be minor. Hence, it was decided to base the verdict of the measurement model equivalence on the practical significance of the difference in multigroup model fit, while still reporting on the statistical significance value.

Step 4a: Establish whether the multigroup measurement model in which the model structure is constrained to be the same across groups and in which all parameters are estimated freely across the samples, except for the factor loadings and the vector of regression of intercepts, displays acceptable fit when fitted to the samples in a multigroup analysis. Alternatively stated, establish whether the strong invariance model displays reasonable fit

The test for strong invariance involves testing the null hypotheses of close fit (H_{05}: RMSEA ≤ .05), which investigated whether the regression slopes and/or intercepts are different between the male and female groups. Finding support for the strong invariance model would provide support for the SAPI developers' claim that the items operate in the same manner, irrespective of the sample groups (Dunbar et al., 2011). A finding of strong invariance would indicate the absence of items that suffer from uniform bias.

If H_{05}: RMSEA ≤ .05 was rejected due to a few intercept terms with significant differences, partial strong invariance would be investigated. The items suffering from uniform bias would be identified via the previously described procedure (see paragraph 3.4.2.3). A finding of a lack of strong invariance (but a finding of partial strong invariance) would provide strong evidence of uniform bias in specific items. Finding strong invariance in turn would constitute weak evidence of the absence of uniform bias.

[47] There is no consensus about the best fit indices that is appropriate across all conditions (Putnick & Bornstein, 2016), leaving the criteria up to researchers. Therefor the decision was taken to use the fit indices proposed by Cheung and Rensvold (2002).
[48] The difference in CFA, Γ_1 and Mc are calculated by subtracting the configural invariance (H_{03}) fits statistics from the weak invariance (H_{04}) fit statistics. All three these fit statistics increase towards 1 as fit improves. Since the constraints in the H_{04} model is expected to reduce the fit, or at best leave the fit unaffected, the calculated change in CFI, Γ_1 and Mc is expected to be negative.

Demonstrating strong invariance does however not mean that differences in any some intercept terms between the sample groups, would not be a more tenable position (Holtzkamp, 2013). Once the outcome of strong invariance (or partial strong invariance) was found, scalar equivalence could be tested.

Step 4b: Establish of the multigroup measurement model in which the model structure is constrained to be the same across groups and in which all parameters are estimated freely across the samples, except for the factor loadings and the vector of regression of intercepts, fits the multigroup data practically significantly poorer than a multigroup measurement model in which only the structure of the model is constrained to be the same across groups but all the parameters are estimating freely. Alternatively stated, establish whether the multigroup measurement model displays scalar equivalence by investigating whether the multigroup strong invariance model displays a practically significantly poorer fit than the multigroup configural invariance model.

Scalar equivalence is demonstrated if the multigroup strong invariance model does not fit practically significantly poorer than the multigroup configural invariance model. In other words, scalar equivalence is indicated when the change from the configural invariance model to the strong invariance model reflects the following practical significance results (Cheung & Rensvold, 2002):

- A change of -.01 or less in the CFI fit index;
- A change of -.001 or less in the Gamma Hat fit index (Γ_1);
- A change of -.02 or less in the McDonald Non-centrality index.

For reasons already stated, the verdict of the measurement model equivalence fit was based on the practical significance outcome, while the statistical significance value would still be reported on.

A finding that (more subtle) differences in the intercept of the regression of specific items on specific latent dimensions exist between different sample groups is not an unreasonable outcome. If the three Cheung and Rensvold (2002) criteria have not been met due to a few intercepts with significant differences, partial scalar equivalence would be investigated by freeing (further) intercepts in group 2 until the three Cheung and Rensvold (2002) criteria were met. A finding of partial scalar equivalence would indicate weak evidence of uniform bias in specific items. A finding of scalar equivalence in turn would constitute strong evidence of the absence of (further) uniform bias.

Step 5a: Establish whether the multigroup measurement model in which the model structure is constrained to be the same across groups and in which all parameters are estimated freely across the samples, except for the factor loadings and the vector of regression intercepts and the measurement error variances of the indicator variables, displays acceptable fit when fitted to the samples in a multigroup analysis. Alternatively stated, establish whether the multigroup strict invariance model displays reasonable fit.

The test for strict invariance involved testing the null hypotheses of close fit (H_{06}: RMSEA ≤ .05), which investigated whether the regression slope, intercept and error variances were different between the male and female groups. Demonstrating strict invariance would illustrate that male and female respondents responded to the items in such a way that no significant variance existed between samples with regard to

error terms associated with the indicator variables (Dunbar et al. 2011). Finding support for the strict invariance model would provide support for the SAPI developers' claim that the items operate in the same manner, irrespective of the sample groups (Dunbar et al., 2011). A finding of strict invariance would imply the absence of items that suffer from error variance bias.

If H_{06}: RMSEA ≤ .05 was rejected due to a few error variance terms with significant differences, partial strict invariance would be investigated. The items suffering from error variance bias would be identified via the previously described procedure (see paragraph 3.4.2.3). A finding of a lack of strict invariance (but a finding of partial strict invariance) would provide strong evidence of error variance bias in specific items, whereas finding of strict invariance would offer weak evidence of the absence of error variance bias.

Demonstrating strict invariance does however not mean that differences in any error variance terms between the sample groups, was not a more tenable position (Holtzkamp, 2013). Once the outcome of strong invariance (or partial strict invariance) was found, conditional probability equivalence could be tested.

Step 5b: Establish whether the multigroup measurement model in which the model structure is constrained to be the same across groups and in which all parameters are estimated freely across the samples, except for the factor loadings, the vector of regression of intercepts and the measurement error variances of the indicator variables, fits the multigroup data practically significantly poorer than a multigroup measurement model in which only the structure of the model is constrained to be the same across groups but all the parameters are estimating freely. Alternatively stated, establish whether the multigroup measurement model displays conditional probability equivalence by investigating whether the multigroup strict invariance model displays practically significantly poorer fit than the multigroup configural invariance model.

Conditional probability equivalence is demonstrated if the multigroup strict invariance model does not fit practically significantly poorer than the multigroup configural invariance model. In other words, conditional probability equivalence is indicated when the change from the configural invariance model to the strict invariance model reflects the following practical significance results (Cheung & Rensvold, 2002):

- A change of -.01 or less in the CFI fit index;
- A change of -.001 or less in the Gamma Hat fit index (Γ_1);
- A change of -.02 or less in the McDonald Non-centrality index.

For reasons already stated, the verdict of the measurement model equivalence fit was based on the practical significance outcome, while the statistical significance value was still reported on.

A finding that (more subtle) differences in the error variances (or standard error of estimates squared) of the regression of specific items on specific latent dimensions exist between different sample groups is not an unreasonable outcome. If the three Cheung and Rensvold (2002) criteria have not been met due to a few error variances with significant differences, partial conditional probability equivalence would be investigated by freeing (further) error variances in group 2 until the three Cheung and Rensvold (2002) criteria were met. A finding of partial conditional probability equivalence would provide weak evidence of error variance bias in specific items, whereas a finding of conditional probability equivalence would constitute strong evidence of the absence of (further) error variance bias.

CHAPTER 5: ETHICAL CONSIDERATIONS

The purpose of this research study has been outlined in Chapter 1. An important part in the process of empirical behavioural research is to evaluate the potential ethical risks associated with the research. This chapter will therefore explore potential ethical risks associated with this study, as well as the guiding principles that informed the researcher.

The objective of reflecting on ethical risks associated with the current study was to ensure that research participants' dignity, rights, safety and well-being remained protected. Behavioural research requires either active or passive involvement from participants, which creates the unfortunate chance that participants' dignity, rights, safety and well-being might be compromised. It has been argued that the purpose of this study is to investigate whether the SAPI might be biased (albeit unintentionally) against one of the gender groups. This benevolent research aim therefore has the broader community of employees' dignity, rights and well-being at heart. To determine whether the potential compromise that participants might have experienced could be justified, the cost that they could have incurred had to be balanced with the potential benefit that the research offers society (Stellenbosch University, 2013).

In Annexure 12 of the Ethical Rules of Conduct for Practitioners Registered under the Health Professions Act [Act no. 56 of 1974] it is stated that a psychologist who performs research is required to enter into an agreement with participants on the nature of the research, as well as the rights and responsibilities of the participants and the researcher. This agreement should comply with the following requirements as stipulated by Annexure 12 (Republic of South Africa (RSA), 2006, p. 42)

> 89. (1) A psychologist shall use language that is reasonably understandable to the research participant concerned in obtaining his or her informed consent.
> (2) Informed consent referred to in sub-rule (1) shall be appropriately documented, and in obtaining such consent the psychologist shall –
> (a) inform the participant of the nature of the research;
> (b) inform the participant that he or she is free to participate or decline to participate in or to withdraw from the research;
> (c) explain the foreseeable consequences of declining or withdrawing;
> (d) inform the participant of significant factors that may be expected to influence his or her willingness to participate (such as risks, discomfort, adverse effects or exceptions to the requirement of confidentiality);
> (e) explain any other matters about which the participant enquires;
> (f) when conducting research with a research participant such as a student or subordinate, take special care to protect such participant from the adverse consequences of declining or withdrawing from participation;
> (g) when research participation is a course requirement or opportunity for extra credit, give a participant the choice of equitable alternative activities; and
> (h) in the case of a person who is legally incapable of giving informed consent, nevertheless –
> (i) provide an appropriate explanation;
> (ii) obtain the participants assent; and
> (iii) obtain appropriate permission from a person legally authorized to give such permission.

One such participant right is to voluntarily make an informed decision on whether or not he/she wishes to participate in the research. Informed consent requires that the participant is informed of the following: (*i*) the research objective and purpose; (*ii*) what the research will involve; (*iii*) the manner in which the research results will be distributed and used; (*iv*) identify the researchers and their respective affiliations; (*v*) clarify where participants can make further inquiries about the research if they wish to so; and (*vi*) specify the participants' research rights and where they can obtain more information relating to their research rights (Stellenbosch University, 2013). On the other hand, it is the researcher's responsibility to ensure that the information is provided to the participants in a dialect that they understand. The test developers obtained informed consent from the research participants for research purposes when the data was initially collected. Since the researcher used archival data provided by the test developers, the researcher acknowledges that she was not in the position to oversee that the information provided to the participants were explained in an understandable manner.

In Annexure 12 of the Ethical Rules of Conduct for Practitioners Registered under the Health Professions Act [Act no. 56 of 1974] (Republic of South Africa (RSA), 2006, p. 41) it is also stipulated that researchers are required to obtain permission from the institution or organisation from which the research participants (or data in this case) is be solicited.

> A psychologist shall –
> - obtain written approval from the host institution or organisation concerned prior to conducting research;
> - provide the host institution or organisation with accurate information about his or her research proposals; and
> - conduct the research in accordance with the research protocol approved by the institution or organisation concerned.

Institutional permission to use the archival dataset on the final version of the SAPI was obtained from the test developers. A copy of the research proposal accompanied the application for institutional permission. This agreement for institutional permission upheld Stellenbosch University's fundamental principal of research ethics and scientific integrity (Stellenbosch University, 2013, p. 3), which obliges researchers "to report research results accurately and transparently in the public domain….and should not allow funders or other stakeholders to influence research publications". Institutional permission therefore allowed the researcher to document the study results with scientific integrity in the form of a maters' thesis, as well as publish the results in an academic article without any influence from the test publishers, irrespective of whether the research results were aligned to the expected results.

The researcher was further responsible for ensuring that the data remained confidential. The dataset provided by the test developers will be treated as anonymous, thereby ensuring confidentiality of participants' information. The focus of this study was not to describe participants' level on the various SAPI constructs, but rather to determine whether the SAPI presents measurement invariance and equivalence between the genders. The results were therefore only presented in an aggregate form. Feedback on the (aggregated) study results will be provided to the SAPI test developers in the form of the thesis document. The researcher was further bound by Annexure 12 of the Ethical Rules of Conduct for Practitioners

Registered under the Health Professions Act [Act no. 56 of 1974] (Republic of South Africa (RSA), 2006, p. 41) to disclose confidential information under the following conditions:

> A psychologist may disclose confidential information –
> - (a) only with the permission of the client concerned;
> - (b) when permitted by law to do so for a legitimate purpose, such as providing a client with the professional services required;
> - (c) to appropriate professionals and then for strictly professional purposes only;
> - (d) to protect a client or other persons from harm; or
> - (e) to obtain payment for a psychological service, in which instance disclosure is limited to the minimum necessary to achieve that purpose.

The SAPI is currently under review to be classified by the Psychometrics Committee of the Professional Board for Psychology (Health Professions Council of South Africa - HPCSA). Hence, the SAPI is regarded as a psychological test under development and not a psychological test as defined by the Health Professions Act [Act no. 56 of 1974] (Republic of South Africa (RSA), 2006). The SAPI was administered by psychometrists and psychologists registered by the HPCSA when the archival data was initially obtained.

The researcher acknowledges that she received financial assistance from the National Research Foundation (NRF). The opinions expressed and conclusions derived at from the research remain that of the researcher and is not attributed to the National Research Foundation (NRF).

Lastly, an application for ethical clearance of the proposed study was submitted to and approved by the Research Ethics Committee Human Research (Humanities) of Stellenbosch University.

CHAPTER 6: RESULTS

6.1. Introduction

The design intention with the SAPI was to measure personality in the diverse South African population. It has been argued in the previous chapters that in order for the SAPI developers to confidently claim that the SAPI measures the personality constructs as constitutively defined across various sample groups, the assessment should also be subjected to measurement invariance and measurement equivalence analyses. Hence, to ensure that the inferences on the personality construct as constitutively defined by the SAPI from scores obtained on the SAPI are permissible for both male and female sample groups in the South African population and can be used with confidence in the same manner across genders, the measurement invariance and equivalence were evaluated in this study.

Chapters 2 to 4 explained in detail the rationale and methodology for using the measurement invariance and measurement equivalence procedure as proposed by Dunbar et al (2011), to evaluate whether the SAPI items are interpreted in the same manner across the two gender groups. This chapter will discuss in detail the research results, the decisions taken on the statistical hypotheses and the subsequent implications thereof.

6.2. Missing values

The SAPI data that was used in the current study was collected electronically via an online questionnaire. The electronic version does no offer the option of “unable to respond” and insists that all items should be completed before allowing the test taker to move to a subsequent test page. The SAPI data set therefore contained no missing values.

6.3. Sampling

This section elaborates on the two sample groups that were used for the present study. The SAPI developers gave institutional permission to use archival SAPI data for this study. The sample should therefore the regarded as a non-probability sample of respondents representing both gender groups from the South African population. One of the disadvantages of using non-probability sampling is that the findings of this study should cautiously be generalised to the general South African population. The age, gender and age x gender frequency distributions are shown in Table 6.1.

Table 6.1

Sample group age, gender and age x gender frequency distributions

Sample	Younger than 20	20-29	30-39	40-49	50-59	60 and older	Total
Female	127	1180	578	381	125	29	2420
Male	79	604	593	340	167	51	1834
Total	**206**	**1784**	**1171**	**721**	**292**	**80**	**4254**

The total sample consisted of 4245 respondents of which 2420 (57%) were female and 1834 (43%) were male. The majority (70%) of respondents are aged either between 20 and 29 years old (42%) or 30 and 39 years old (28%). It is therefore acknowledged that the sample group is slightly skewed towards the female

sample, and that the South African population might be under represented for the age groups of 40 years and older. Table 6.1 indicates no apparent difference in the age distributions of male and female respondents, except perhaps for the age groups 20 to 29 years and older than 59 years. The home language, gender and home language x gender frequency distribution is shown in Table 6.2.

Table 6.2

Sample group home language, gender and home language x gender frequency distributions

Home language	Female	Male	Total
Afrikaans	549	351	**900**
English	615	425	**1040**
isiNdebele	10	2	**12**
isiXhosa	107	57	**164**
isiZulu	176	91	**267**
Unspecified	637	677	**1314**
Other	24	20	**44**
Sepedi	104	61	**165**
Sesotho	46	44	**90**
Setswana	102	62	**164**
Siswati	10	12	**22**
Tshivenda	18	8	**26**
Xitsonga	22	24	**46**
Grand Total	**2420**	**1834**	**4254**

Table 6.2 indicates that most respondents (30.89%) unfortunately did not specify their home language. The majority of respondents (24.45%) selected English as their home language, followed by Afrikaans (21.20%). There is no apparent difference in the language distribution across the two genders. Table 6.3 depicts the race, gender and race x gender frequency distribution.

Table 6.3

Sample group race, gender and race x gender frequency distributions

Race	Female	Male	Total
African	117	66	**183**
Asian	3	2	**5**
Black	543	337	**880**
Coloured	133	99	**232**
Indian	96	88	**184**
Unspecified	627	673	**1300**
Other	51	24	**75**
White	850	545	**1395**
Total	**2420**	**1834**	**4254**

Table 6.3 indicates that a large number of respondents choose not to specify their race (30.56%). White was the race category specified by most respondents (32.79%), followed by Black/African (24.99%)[49]. Table 6.3 indicates no apparent difference in the race distributions of male and female respondents.

6.4. Evaluation of SAPI Measurement Model

Evaluating the various single- and multigroup measurement models required clarifying the variable type and estimation techniques applied in preparation for the subsequent sequential analyses.

As indicated in Chapter 4, the default technique to obtain estimates for freed model parameters when using LISREL is maximum likelihood estimation (ML). The assumption for ML that the indicator variables are normally distributed ensures correct calculations of standard errors and chi-square estimates (Mels, 2003). Robust maximum likelihood estimation (RML) is generally recommended as the preferred estimation technique for dealing with data that is not normally distributed. However, RML requires the researcher to create the asymptotic covariance matrix (ACM), which demands large amounts of computer memory. Mouton (2017) insists that the I ideal remains to evaluation of the SAPI measurement models with individual items rather than with item parcels. The study aimed to evaluate the measurement model with as much accuracy as possible. The concern that item parcels could potentially cause under reporting of biased items not being detected due to averaging of the observed scores contained in the item parcel, resulted in the decision to rather use individual items. Therefore, the benefits of evaluating the individual items with the less appropriate ML technique, irrespective of satisfying the assumption of normal distribution prevailed over the benefits of analysing item parcels with RML.

Although the verdict of whether or not the data is normally distributed would not influence the decision to use ML, the results for the multivariate normality evaluation for both gender groups are nonetheless provided in Table 6.4.

Table 6.4

Test of multivariate normality for continuous variables

Gender Group	Skewness			Kurtosis			Skewness and Kurtosis	
	Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
Female	3525.185	377.477	0.000	34235.664	102.829	0.000	153062.767	0.000
Male	4595.070	368.073	0.000	34447.574	90.599	0.000	143685.858	0.000

The multivariate normality of the indicator variables were evaluated, using PRELIS. The null hypothesis of multivariate normality of the item indicator variables had to be rejected ($p < .05$) for both the male and the female samples.

[49] The biographical data formed part of the archival data supplied by the SAPI test developers. It is assumed that the somewhat unusual collection of response options (e.g. African versus Black; Unspecified versus Other) was due to changes in the response options that were offered to test-takers over time.

6.5. Evaluating the SAPI Single-group Measurement Model Fit (H_{01} & H_{02}) Via Confirmatory Factor Analysis

As a prerequisite for testing the SAPI multigroup measurement models for measurement invariance and measurement equivalence, the SAPI single-group measurement models had to demonstrate good fit. Hence the single-group measurement models were subjected to confirmatory factor analysis.

6.5.1. Measurement Model Fit Indices

LISREL 8.8 was used to apply confirmatory factor analysis in testing the null hypotheses of exact fit (H_{01i}: RMSEA = 0; i = 1_{Female}, 2_{Male}), and close fit (H_{02i}: RMSEA ≤ .05; i = 1_{Female}, 2_{Male}) for both single-group measurement models. A summary of the goodness of fit statistics are provided in Table 6.5. The degrees of freedom shown in Table 6.5 correspond to those calculated in Table 4.1. The complete fit statistics for the female and male sample groups' measurement models are listed Appendix B and Appendix C respectively.

Table 6.5

Summary of goodness fit statistics for the single-group measurement models

CFA Model tested	Single Group: Female	Single Group: Male
Hypotheses Tested	H_{01} & H_{02}	H_{01} & H_{02}
Degrees of Freedom	14005	14005
RMSEA	.0475	.0492
90 Percent Confidence Interval for RMSEA	(.0472 ; .0478)	(.0489 ; .0496)
P-Value for Test of Close Fit (RMSEA < 0.05)	1.000	1.000
Normal Theory Weighted Least Squares Chi-Square	90548.77	76246.177
P-Value for Test of Exact Fit (RMSEA = 0)	0	0
Population Discrepancy Function Value (F0)	31.640	33.956
Normed Fit Index (NFI)	.95	.951
Parsimony Normed Fit Index (PNFI)	.926	.927
Comparative Fit Index (CFI)	.96	.963
Root Mean Square Residual (RMR)	.0401	.0454
Goodness of Fit Index (GFI)	.694	.671
Adjusted Goodness of Fit Index (AGFI)	.68	.659

The exact fit null hypotheses (H_{01i}: RMSEA = 0; i = 1_{Female},2_{Male}) had to be rejected in favour of the alternative hypotheses (H_{a1i}: RMSEA > 0; i = 1_{Female},2_{Male}; p < .05) for both gender groups. The close fit null hypotheses (H_{02i}: RMSEA ≤ .05; i = 1_{Female},2_{Male}) could not be rejected (p < .05) for either of the gender groups. Both single-group measurement models in the sample demonstrated good model fit with RMSEA values below .05 (p < .05). Not rejecting H_{02i}: RMSEA ≤ .05; i = 1_{Female}, 2_{Male} for both groups meant that it was permissible to hold the position that both the male and the female single-group SAPI measurement model fitted closely in the parameter.

Comparative fit indices such as normed fit index (NFI), parsimony normed fit index (PNFI), and comparative fit index (CFI) scores use a benchmark and independence model to contrast a measurement model's ability to reproduce the observed covariance matrix. The critical threshold .for comparative fit indices of .90 and above indicates good fit (Spangenberg & Theron, 2005). The closer these values are to unity, the better the model fit. The NFI, PNFI and CFI values for both sample groups exceeded the critical value of .90.

The root mean square residual (RMR) provides the average value of the residual matrix. Whilst RMR values of 0 indicate perfect fit, scores of .05 and less indicate good fit (Diamantopoulos & Siguaw, 2000). The satisfactory RMR levels for both the female group (.0401) and male group (.0454) serve as further support for the models' good fit.

The goodness of fit index (GFI) reflects the extent to which a model was able to perfectly reproduce the sample covariance matrix (Diamantopoulos & Siguaw, 2000). The adjusted goodness of fit index (AGFI) adjusts the GFI to accommodate for the degrees of freedom in the model and ranges between 0 and 1.0, with AGFI values that exceed .90 indicating good model fit (Jöreskog & Sörbom, 1993; Kelloway, 1998). Both sample groups' GFI and AGFI levels range between .659 and .694, thereby unfortunately not achieving the desired cut-off of .90. However, Kelloway (1998) stresses that this cut-off level is only based on experience and should be applied with caution.

6.5.2. Measurement Model Residuals

Residuals represent the differences in corresponding cells for the observed and fitted covariance matrices (Diamantopoulos & Siguaw, 2000). Residuals that are divided by their estimated standard errors are known as standardised residuals (Jöreskog & Sörbom, 1993). Residuals, standardised residuals in particular, provide invaluable diagnostic information regarding the degree to which a model lacks fit (Kelloway, 1998). Residuals should ideally be distributed symmetrical around zero. The relationships between indicator variables that the model fails to explain are reflected in the number of large (i.e. exceeding the |2.58| cut-off) positive or negative residuals that have absolute values greater than zero (Diamantopoulos & Siguaw, 2000). Large positive residuals point to underestimation and therefore a need for additional paths between latent variables and indicator variables, whereas large negative residuals point to overestimation and therefore a need to remove some paths that are connected to those identified indicator variables (Diamantopoulos & Siguaw, 2000). An excess of large residuals on either side of zero would therefore indicate a systematic over- or underestimation of the variance and covariance terms.

Standardised residuals are evaluated based on both the stem-and-leaf plot and Q-plot. The female measurement model stem-and-leaf plot shown in Figure 6.1., depicts a somewhat positively skewed distribution. This indicates that a few very large positive residuals were obtained that were not observed on the negative side of the distribution. The number of negative (3200) and positive (3199) large standardised residuals were for practical purposes the same. The female single-group measurement model therefore tended to underestimate the observed covariance terms to the same degree than it overestimated covariance terms. The percentage of large variance-covariance residuals (44.03%) suggests a mediocre fitting model.

```
-12|41
-10|88662200987754332000
- 8|99988777666555433322222221111000000099999988888777777766665555544443333 2+26
- 6|99999998888888888777777777777777776666666555555555554444444444444444444 4+95
- 4|99999999999999999999999999999999888888888888888888888888888888888777777 7+90
- 2|99999999999999999999999999999999999999999999999999999999999999999999999 9+90
- 0|99999999999999999999999999999999999999999999999999999999999999999999999 9+93
  0|11111111111111111111111111111111111111111111111111111111111111111111111 1+91
  2|00000000000000000000000000000000000000000000000000000000000000000000000 0+90
  4|00000000000000000000000000000000000000000000000000000000000000001111111 1+96
  6|00000000000000000000000000000000000000001111111111111111111111111111111 1+96
  8|00000000000000000000000011111111111112222222222233333333333444444444444 4+20
 10|00000011111222333344445555555566666666777777889990001122333344455666666 6777
 12|012222333345556666900223345566 9
 14|000045789123456666679
 16|7702224
 18|1555577
 20|0257
 22|6
 24|5
```

Figure 6.1. **Stem-and-leaf plot of standardised residuals for the female sample measurement model**

The Q-plot provides another graphical display of standardised residuals against the quantiles of the normal distribution (Diamantopoulos & Siguaw, 2000). Good model fit is indicated by the extent to which the data points correspond with the 45-degree reference line (Jöreskog & Sörbom, 1993). The Q-plot for the female sample measurement model is displayed in Figure 6.2. The deviation from the 45-degree reference line in the upper and lower regions of the X-axis, indicate that the female sample measurement model showed some degree of problematic fit. The evaluation that emerged from the examination of the standardised residuals was not consistent with the reassuring picture that emerged from the fit statistics.

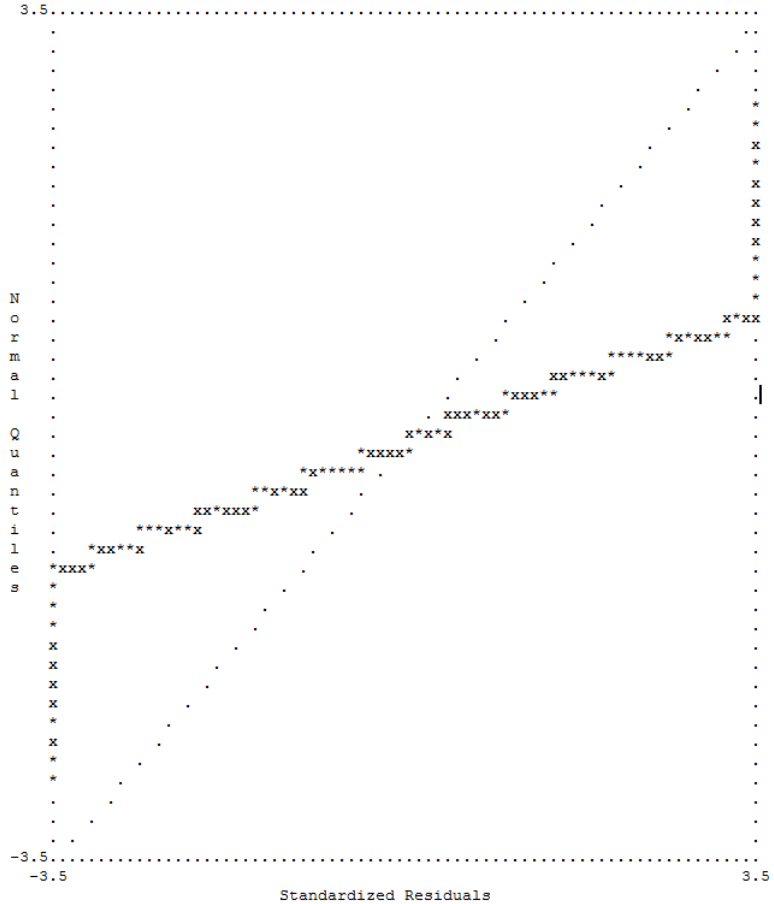

Figure 6.2. **Q-plot of standardised residuals for the female sample measurement model**

The male sample measurement model stem-and-leaf shown Figure 6.3 also depicts a positively skewed distribution, thereby indicating that large positive residuals tend to dominate (albeit only marginally). There were 5845 variance and covariance terms in the 14535-term observed covariance matrix that were poorly estimated (40.21%) with 2840 large negative standardised residuals and 3005 large positive standardised residuals. The male sample measurement model therefore tended to underestimate the observed variance and covariance terms slightly more than it overestimated it.

```
-10|72097753220
- 8|876655544332100099988887666644433333222111111110000
- 6|99999888877777777666666665555555554444444443333333333322222211111111100+96
- 4|99999999999999999999999988888888888888888888888887777777777777777777+96
- 2|99999999999999999999999999999999999999999999999999999999999999999999+91
- 0|99999999999999999999999999999999999999999999999999999999999999999999+90
  0|11111111111111111111111111111111111111111111111111111111111111111111+97
  2|00000000000000000000000000000000000000000000000000000000000000000000+98
  4|00000000000000000000000000000000000000000000000000000000000000000000+92
  6|00000000000000000000000011111111111111111111111111111122222222222222+96
  8|00000000001111111111111112222222222222222223333333444444444555555555556+72
 10|01112223333333345566777788899900001111222223334446678 88
 12|11222333446799012233 99
 14|344697
 16|01278468
 18|428
 20|2662
```

Figure 6.3. **Stem-and-leaf plot of standardised residuals for the male sample measurement model**

Similar to the female measurement model, the Q-plot for the male sample measurement model depicted in Figure 6.4. demonstrated a deviation from the 45-degree reference line in the upper and lower regions of the X-axis.

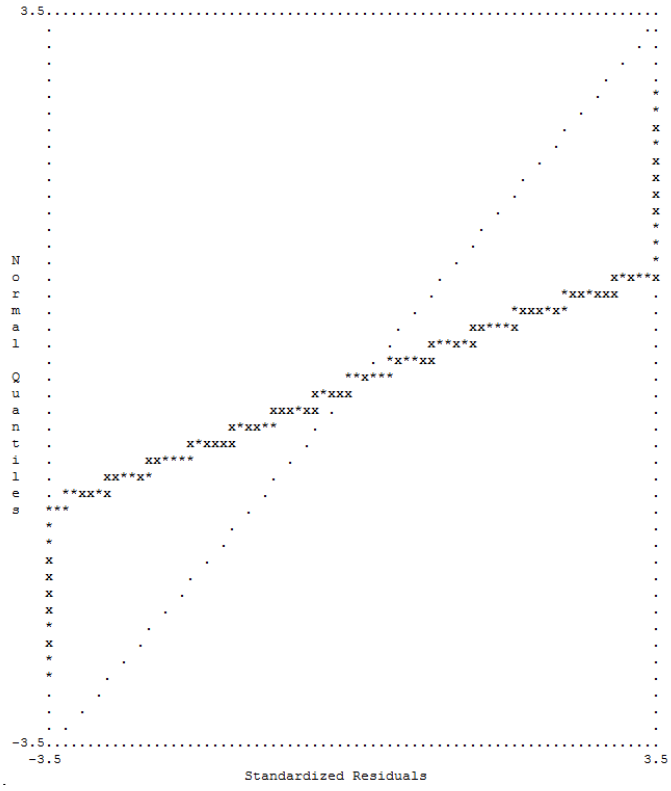

Figure 6.4. **Q-plot of standardised residuals for the male sample measurement model**

Therefore the male sample measurement model also showed some degree of problematic fit. The evaluation that emerged from the examination of the standardised residuals was also inconsistent with the reassuring picture that emerged from the fit statistics.

In summary, considering the female and male sample measurement models' fit statistics holistically it can be concluded that both the single-group measurement models demonstrated sufficiently good model fit to continue with the measurement bias analysis (Dunbar et al., 2011). Consequently, the multigroup measurement models were evaluated for measurement invariance and measurement equivalence.

6.6. Evaluating the SAPI Multigroup measurement Invariance and Equivalence

The sequence of measurement invariance and measurement equivalence tests were performed as described in Chapter 4. A summary of the results are provided in Table 6.6

Table 6.6.

Summary of goodness fit statistics for the multigroup measurement models

	Configural invariance	Weak invariance	Strong invariance	Partial strong invariance	Strict invariance	Partial strict invariance
Hypotheses Tested	H_{03}	H_{04}	H_{05}	H_{0585}	H_{06}	H_{06102}
Degrees of Freedom	28010	28160	28330	28245	28415	28313
RMSEA	.0483	.0483	.0494	.0482	.0484	.0482
90 Percent Confidence Interval for RMSEA	(.0 ; .0)[50]	(.0 ; .0)	(.0 ; .0)	(.0 ; .0)	(.0 ; .0)	(.0 ; .0)
P-Value for Test of Close Fit (H_{0i}: RMSEA < 0.05)	1.000	1.000	1.000	1.000	1.000	1.000
Normal Theory Weighted Least Squares Chi-Square	166794.946	167549.691	175326.865	168032.538	169832.984	168119.255
P-Value for Test of Exact Fit (H_{0i}: RMSEA = 0)	0	0	0	0	0	0
Population Discrepancy Function Value (F0)	32.640	32.782	34.571	32.876	33.259	32.880
Normed Fit Index (NFI)	.931	.931	.928	.931	.93	.931
Parsimony Normed Fit Index (PNFI)	.908	.913	.915	.915	.92	.917
Comparative Fit Index (CFI)	.942	.942	.939	.942	.941	.942
Root Mean Square Residual (RMR)	.0454	.046	.049	.0461	.0462	.0461
Goodness of Fit Index (GFI)	.671	.670	.654	.669	.668	.669

6.6.1. Configural Invariance (H_{03})

The test for configural invariance evaluates whether the structure of the measurement model is invariant across groups by establishing whether the multigroup measurement model of which the structure is constrained to be equal, but all other model parameters are freely estimating across the two gender groups, displays close fit when fitted to the two sample groups simultaneously in a multigroup analysis. Failure to reject the null hypothesis of close fit would indicate that the structure of the SAPI measurement model is invariant across the two gender groups. This in turn would indicate the absence of construct bias. Finding

[50] These 90% confidence intervals do no make logical sense. The upper and lower bounds should be positioned around the obtained point estimate of RMSEA.

support for configural invariance would in turn serve as a prerequisite for evaluating any further aspects of measurement invariance and measurement equivalence, as well as a reference model to evaluate subsequent nested models against in measurement equivalence calculations.

The configural invariance null hypothesis H_{03}: RMSEA ≤ .05 was tested by fitting the multigroup configural invariance SAPI measurement model across the male (N = 1834) and female (N = 2420) samples. A few of the fit statistics for the SAPI configural invariance multigroup measurement model are displayed in Table 6.6. The degrees of freedom shown in Table 6.6 for the multigroup configural invariance measurement model correspond to the degrees of freedom calculated in Table 4.1. The complete fit statistics for the multigroup configural invariance measurement model are provided in Appendix D.

The RMSEA value of .0483 indicated good model fit in the sample. The reported p-value for the test of close fit achieved a value of 1.00. The probability of observing the RMSEA value of .0483 in a sample drawn from a population where the multigroup model fits closely was therefore sufficiently large not to question the assumption of close fit in the parameter. The NFI, PNFI and CFI all exceeded the critical cut-off of .90 and the RMR value remained under the .050 cut-off.

These fit indicators revealed that the configural invariance multigroup measurement model showed good fit. Therefore, the SAPI demonstrated configural invariance indicating that the SAPI successfully measured the same personality construct across the male and female groups. Consequently, the assumption that the SAPI lacks construct bias is permissible.

6.6.2. Weak Invariance (H_{04})

The reasonable fit of the single group measurement models and multigroup configural invariance measurement model allowed for the test of weak invariance. The test for weak invariance investigated whether the factor loadings of items on the latent variables lack non-uniform bias across different samples, by establishing whether the multigroup measurement model in which both the model structure and the slope of the regression of the indicator variables on the latent variables are constrained to be equal across the gender groups, but all other model parameters were freely estimated across the groups, displayed reasonable fit when simultaneously fitted to the two sample groups in a multigroup analysis.

Failure to reject the null hypothesis of close fit would indicate that the factor loadings of the SAPI measurement model are invariant across the two gender groups. Alternatively stated, a lack of weak invariance would imply a difference across the two sample groups in the slope of the regression of one or more of the SAPI's items on the latent variables they represent. Finding support for the weak invariance model would indicate that the male and female groups perceive and interpret the item content in a similar manner (Byrne & Watkins, 2003).

The weak invariance null hypothesis H_{04}: RMSEA ≤ .05 was tested by fitting the multigroup weak invariance SAPI measurement model across the male (N = 1834) and female (N = 2420) samples. A few of the fit statistics for the multigroup SAPI weak invariance measurement model are displayed in Table 6.6. The degrees of freedom shown in Table 6.6 for the multigroup weak invariance measurement model correspond

to the degrees of freedom calculated in Table 4.1. The complete fit statistics for the weak invariance multigroup measurement model are provided in Appendix E.

The RMSEA value of .0483 indicated good model fit in the sample. These fit indicators, along with the reported p-value for the test of close fit of 1.00, revealed that it is permissible to hold the position that the multigroup weak invariance measurement model showed close fit in the parameter.

The normed fit index (NFI = .931), the parsimony normed fit index (PNFI = .913) and the comparative fit index (CFI = .942) all exceeded the critical cut-off of .90. The RMR in turn also remained below the threshold of .050. These fit statistics further supported the permissibility of the claim that the weak invariance multigroup measurement model showed good fit in the sample.

The SAPI demonstrated weak invariance indicating that the position that the SAPI successfully measured the same (personality) construct across the two samples groups, and that both male and female groups perceived, interpreted and responded to the item content in the same manner was a tenable position[51]. The stance that the SAPI lacks non-uniform bias is therefore a tenable position. However, permissibility of this position would be more convincing if it could be demonstrated that the multigroup weak invariance measurement model in which the slopes of the regression of the items on the latent variables that they represent, were the same across both sample groups did not fit practically significantly poorer than a multigroup measurement model in which they are freely estimated within each group (i.e. the multigroup configural invariance measurement model). Hence, metric equivalence was investigated.

6.6.3. Metric Equivalence (H_{07})

The support for weak invariance allowed for the investigation of metric equivalence of the SAPI. Metric equivalence (H_{07}) evaluated whether the weak invariance model (multigroup measurement model in which both the model structure and the slope of the regression of the indicator variables on the latent variables were constrained to be equal across the two sample groups) fitted practically significantly or statistically significantly poorer than the configural invariance model (multigroup measurement model in which only the model structure was constrained to be equal across the two sample groups). Although the verdict of the measurement model equivalence was based on whether or not practical significance is obtained, the results for the statistical significance evaluation are nonetheless provided.

If the probability of observing the normal theory chi-square difference in a multigroup sample under the null hypothesis of no difference in fit between the configural and weak invariance models in the parameters was smaller than or equal to .05, the null hypothesis was rejected in favour of the hypothesis that the fit of the multigroup configural invariance and multigroup weak invariance models differ in the parameter. As indicated in Table 6.7, the null hypothesis of no difference in model fit in the parameter was rejected, indicating a lack of metric equivalence (i.e. a lack of equivalence of factor loadings) across the two gender groups when using statistical significance as the benchmark.

[51] A definite verdict on whether the two gender groups perceived, interpreted and responded to the items in the same manner cannot be reached before non-uniform bias, uniform bias and error variance bias had been convincingly ruled out (via the appropriate tests of measurement equivalence). The position that the two gender groups perceived, interpreted and responded to the items in the same manner remain a permissible position to hold as long as this position is not falsified.

Table 6.7

Statistical significance of the scaled chi-square difference statistic: a test of metric equivalence

Hypotheses		Normal Theory Chi-Square	Df	Prob Normal Theory Chi-Square Diff	Statistical Significance
Configural invariance model	H_{03}	166794.946	28010		
Weak invariance model	H_{04}	167549.691	28160		
Difference (H_{04} - H_{03}) Metric Equivalence	H_{07}	754.745		2.317936E-81	Statistical Significance

Practical significance would be achieved when the change[52] from the configural invariance model to the weak invariance model was -.01 or less in the CFI fit index, -.001 or less in the Gamma Hat fit index (Γ_1) and -.02 or less in the McDonald Non-centrality index (Cheung & Rensvold, 2002). Metric equivalence was established, by achieving the satisfactory levels of change in CFI, Gamma Hat and MacDonald difference as indicated by Table 6.8.

Table 6.8

Practical significance of the CFI, Gamma Hat and MacDonald difference statistics: a test of metric equivalence

Model	Hypotheses	CFI	Γ_1	Mc	Decision
Configural invariance model	H_{03}	.942	.838926174	8.17173E-08	
Weak invariance model	H_{04}	.942	.838338709	7.61166E-08	
Difference [H_{04}-H_{03}; Test Of Metric Equivalence]	H_{07}	.0000	-.0006	-.0000	Metric Equivalence

The SAPI demonstrated metric equivalence indicating that the SAPI weak invariance measurement model (in which both the structure and the slopes of the regression of the items on the latent variables that they represent) does not fit practically significantly poorer than the configural invariance measurement model (in which only the structure is constrained to be equal across the two groups). This means that the position that the SAPI successfully measured the same construct across the two samples groups, and that both male and female groups perceived, interpreted and responded to the item content in the same manner was still a justifiable position even when evaluated against the more stringent yardstick of metric equivalence. Based on these analyses, the stance that the SAPI lacks non-uniform bias is therefore a convincing position. Finding metric equivalence allowed for evaluating the more stringent strong invariance test.

6.6.4. Strong Invariance (H_{05})

The test for strong invariance investigated whether the intercepts of the regression of the items on the latent variables lack invariance across different samples. This test was conducted by establishing whether the multigroup measurement model of which both the model structure, the slope and intercepts of the regression of the indicator variables on the latent variables were constrained to be equal across groups, but all other

[52] There is no consensus about the best fit indices that is appropriate across all conditions (Putnick & Bornstein, 2016), leaving the criteria up to researchers. Therefor the decision was taken to use the fit indices proposed by Cheung and Rensvold (2002).

model parameters were freely estimated across groups, displayed reasonable fit when simultaneously fitted to the two sample groups in a multigroup analysis.

Finding support for the strong invariance model would provide more support for the SAPI developers' claims that the items operated in the same manner (i.e. that the male and female sample groups perceive, interpret and respond to the item content in a similar manner) and that no items suffered from uniform bias (Byrne & Watkins, 2003). Failure to reject the null hypothesis of close fit would indicate that the intercepts of the SAPI measurement model were invariant across the two gender groups. A lack of strong invariance would conversely imply a difference across the two sample groups in the intercepts of the regression of one or more of the SAPI's items on the latent variables they represent.

The strong invariance null hypothesis H_{05}: RMSEA ≤ .05 was tested by fitting the multigroup strong invariance SAPI measurement model across the male (N = 1834) and female (N = 2420) samples. A few of the fit statistics for the SAPI strong invariance multigroup measurement model are displayed in Table 6.6. The degrees of freedom shown in Table 6.6 for the multigroup strong invariance measurement model corresponded to the degrees of freedom calculated in Table 4.1. The complete fit statistics for the strong invariance multigroup measurement model are provided in Appendix F.

The RMSEA value of .0494 indicated good model fit in the sample. These fit indicators, along with the reported p-value for the test of close fit of 1.00, revealed that the strong invariance multigroup measurement model showed close fit in the parameter.

The normed fit index (NFI = .928), the Parsimony Normed Fit Index (PNFI = .915) and the comparative fit index (CFI = .939) all exceeded the critical cut-off of .90. The RMR in turn also remained below the threshold of .050. These results further supported the permissibility of the claim that the strong invariance multigroup measurement model showed good fit in the sample.

The SAPI demonstrated strong invariance indicating that the position that the SAPI successfully measured the same constructs across the two samples groups, and that both male and female groups perceived, interpreted and responded to the item content in the same manner was a justifiable position. The stance that the SAPI lacks uniform bias was therefore a tenable position. However, permissibility of this position will be more convincing if it could be demonstrated that a multigroup measurement model in which the intercepts of the regression of the items on the latent variables that they represented, were the same across both sample groups did not fit practically significantly poorer than a multigroup measurement model in which the intercepts were estimated freely within each group (i.e. the configural invariance model). Hence, scalar equivalence was investigated.

6.6.5. Scalar Equivalence (H_{08})

The support for strong invariance allowed for the testing of scalar equivalence (H_{08}) of the SAPI. Scalar equivalence evaluated whether the strong invariance model (multigroup measurement model in which both the model structure, the slopes and intercepts of the regression of the indicator variables on the latent variables were constrained to be equal across the two sample groups) fitted practically significantly (or

statistically significantly) poorer than the configural invariance model (multigroup measurement model in which only the model structure was constrained to be equal across the two sample groups). Although the verdict of the measurement model equivalence was based on whether or not practical significance was obtained, the results for the statistical significance evaluation are nonetheless provided.

If the probability of observing the normal theory chi-square difference in a multigroup sample under the null hypothesis of no difference in fit between the configural and strong invariance models in the parameters was smaller than or equal to .05, the null hypothesis was rejected in favour of the hypothesis that the fit of the multigroup configural invariance and multigroup strong invariance models differ in the parameter. As indicated in Table 6.9, the null hypothesis of no difference in model fit in the parameter was rejected, indicating a lack of strong equivalence (i.e. a lack of equivalence of intercepts) across the two gender groups when using statistical significance as the benchmark.

Table 6.9

Statistical significance of the scaled chi-square difference statistic: a test of scalar equivalence

Hypotheses		Normal theory chi-square	Df	Prob normal theory chi-square diff	Statistical significance
Configural invariance model	H_{03}	166794.946	28010		
Strong invariance model	H_{05}	175326.865	28330		
Difference (H_{05}-H_{03}) scalar equivalence	H_{08}	8531.919		.000000	Statistical significance

The practical significance of the strong equivalence test was based on the cut-off levels of change in CFI, Gamma Hat and MacDonald difference as indicated by Table 6.10.

Table 6.10

Practical significance of the CFI, Gamma Hat and MacDonald difference statistics: a test of scalar equivalence

Model	Hypotheses	CFI	Γ_1	Mc	Decision
Configural Invariance model	H_{03}	.942	.838926174	8.17173e-08	
Strong Invariance model	H_{05}	.939	.831007328	3.11174e-08	
Difference [H_{05}-H_{03}; Test Of Scalar Equivalence]	H_{08}	-0.0030	-0.0079	-0.0000	Lack Of Scalar Equivalence

Based on the practical significance benchmark, scalar equivalence was not demonstrated with the strong invariance measurement model (multigroup measurement model in which both the model structure, the slopes and intercepts of the regression of the indicator variables on the latent variables were constrained to be equal across the two sample groups). This indicated that some items suffered from uniform bias and that the regression intercepts of those items differed between the male and female groups. The SAPI demonstrated a lack of scalar equivalence indicating that although the position that the SAPI successfully measured the same construct across the two samples groups, the position that both male and female groups perceived, interpreted and responded to the item content in the same manner was no longer a justifiable position when evaluated against the more stringent yardstick of scalar equivalence.

6.6.6. Strong Invariance-partial scalar equivalence model (H_{05i} & H_{08i})[53]

Due to finding a lack of scalar equivalence, an iterative process was followed to identify problematic items by ranking the items from the largest difference between the female and male sample groups based on the configural invariance model's unstandardised tau estimates. Based on this rank-order the equality constraint was released on a total of 85 items by having the tau estimates freely estimated in group 2, before partial scalar equivalence could be demonstrated (i.e. before the difference in fit between the multigroup partial strong invariance measurement model and the multigroup configural invariance model was no longer practically significant as judged by the Cheung and Rensvold (2002) criteria). By relaxing the constraints on the multigroup measurement model the measurement model became a partial strong invariance measurement model. Appendix G provides the list of problematic items for which the tau parameters were freed to be estimated (i.e. items flagged as suffering from uniform bias).

The partial strong invariance null hypothesis H_{0585}: RMSEA ≤ .05 was tested by fitting the multigroup SAPI measurement model across the male (N = 1834) and female (N = 2420) samples. A few of the fit statistics for the SAPI partial strong invariance multigroup measurement model are displayed in Table 6.6. Appendix H provides the complete fit statistics for the partial strong invariance multigroup measurement model.

The RMSEA value of .0482 indicated good model fit in the sample. These fit indicators, along with the reported p-value for the test of close fit of 1.00, revealed that the partial strong invariance multigroup measurement model showed close fit in the parameter. Since H_{05}: RMSEA ≤ .05 was not rejected ($p > .05$), there was no doubt that any of the multigroup partial strong invariance measurement models would demonstrate close fit[54].

The normed fit index (NFI = .931), the parsimony normed fit index (PNFI = .915) and the comparative fit index (CFI = .942) all exceeded the critical cut-off of .90. The RMR in turn also remained below the threshold of .050. These results further supported the permissibility of the claim that the partial strong invariance multigroup measurement model showed good fit in the parameters.

If the probability of observing the normal theory chi-square difference in a multigroup sample under the null hypothesis of no difference in fit between the configural and partial strong invariance models in the parameters was smaller than or equal to .05, the null hypothesis was rejected in favour of the hypothesis that the fit of the multigroup configural invariance and multigroup partial strong invariance models differ in the parameter. As indicated in Table 6.11, the null hypothesis of no difference in model fit in the parameter was rejected for all the investigated partial strong measurement models, indicating a lack of partial scalar equivalence (i.e. a lack of equivalence of intercepts) across the two gender groups when using statistical significance as the benchmark.

[53] A clear, unambiguous terminological convention is still lacking. The current study would argue that the model in question here should be termed a strong invariance – partial scalar equivalence model since strong invariance was achieved with all tau estimates constrained to be equal across the gender groups but scalar equivalence was not achieved before specific equality constraints on the tau vector was lifted. If strong invariance was not initially achieved, and specific equality constraints on the tau vector had to be lifted before strong invariance was obtained and additional equality constraints had to be lifted before scalar equivalence was achieved this would have been referred to as a partial strong invariance -partial scalar equivalence model.

[54] The fit of the partial strong invariance is in and by itself not an issue here since strong invariance was already achieved even when the complete tau vector was constrained to be equal across gender groups. The partial strong invariance fit statistics are of importance only to evaluate the practical significance of the difference in fit between the configural invariance and partial strong invariance models.

Table 6.11

Statistical significance of the scaled chi-square difference statistic: a test of partial scalar equivalence

Hypotheses		Normal theory chi-square	Df	Prob normal theory chi-square diff	Statistical significance
Configural invariance model	H_{03}	166794.946	28010		
Strong invariance model	H_{05}	175326.865	28330		
Partial strong invariance model	H_{0584}	168059.168	28246		
Partial strong invariance model	H_{0585}	168032.538	28245		
Difference (H_{05}-H_{03}) scalar equivalence	H_{08}	8531.919		.000000	Statistical significance
Difference ($H_{05,84}$-H_{03}) partial scalar equivalence	H_{0884}	1264.222		.000000	Statistical significance
Difference ($H_{05,85}$-H_{03}) partial scalar equivalence	H_{0885}	1237.592		.000000	Statistical significance

The iterative process of releasing tau parameters of problematic items to be freely estimated in both gender groups continued until practical significance was demonstrated. Table 6.12 provides the changes in CFI, Gamma Hat and MacDonald difference which were used as indicators for practical significance.

Table 6.12

Practical significance of the CFI, Gamma Hat and MacDonald difference statistics: a test of partial scalar equivalence

Model	Hypotheses	CFI	Γ1	Mc	Decision
Configural Invariance model	H_{03}	.942	.838926174	8.17173E-08	
Strong Invariance model	H_{05}	.939	.831007328	3.11174E-08	
Partial strong invariance model	H_{0584}	.942	.837925494	7.24043E-08	
Partial strong invariance model	H_{0585}	.942	.837950275	7.26218E-08	
Difference [H_{05}-H_{03}; Test Of Scalar Equivalence]	H_{08}	-0.0030	-0.0079	-0.0000	Lack Of Scalar Equivalence
Difference [$H_{05,84}$ –H_{03}; Test Of Partial Scalar Equivalence]	H_{0884}	0.0000	-0.001001	-0.0000	Lack Of Partial Scalar Equivalence
Difference [$H_{05,85}$ – H_{03}; Test Of Partial Scalar Equivalence]	H_{0885}	0.0000	-0.000976	-0.0000	Scalar Partial Equivalence

The SAPI demonstrated strong invariance-partial scalar equivalence indicating that the position that the SAPI measured the same construct across the two samples groups, and the position that both the male and female groups perceived, interpreted and responded to the item content for 85 of the items in the same manner were justified positions. The stance that the SAPI lacks uniform bias was therefore not permissible, as 85 of the 170 items were shown to demonstrate uniform bias (i.e. the intercept of the regression of 85 items on their designated latent first-order personality dimension differed practically significantly across the two gender groups). Table 6.13 indicated that 12 of the 20 constructs were reported to have between 50% and 100% of items that contain biased intercepts. The intercepts for all of the items in *Arrogance, Empathy and Social Intelligence* were biased towards one of the sample groups. All the items in these constructs

therefore had to be released to be freely estimated before strong invariance-partial scalar equivalence was demonstrated. Constructs with more than half of the items containing practically significantly different intercepts included *Deceitfulness*, *Traditionalism–Religiosity, Hostility–Egoism*, *Warm-Heartedness*, *Negative Emotionality, Orderliness, Sociability, Emotional Balance and Playfulness*.

Table 6.13

Identifying which constructs were implicated by biased intercepts

Dimension	Code	% Items in construct with biased intercepts	D value = Female average less Male average Female = Females scored higher on average Male = Males scored higher on average
Achievement Orientation	ACH	36%	Female
Arrogance	ARR	100%	Male
Broad-Mindedness	BRO	33%	Male
Conflict-Seeking	CON	43%	Male
Deceitfulness	DEC	86%	Male
Emotional Balance	EMO	50%	Male
Empathy	EMP	100%	Female
Epistemic Curiosity	EPI	0%	Male
Facilitating	FAC	10%	Female
Hostility–Egoism	HOS	71%	Male
Integrity	INTEG	36%	Female
Intellect	INT	23%	Male
Interpersonal Relatedness	INTER	0%	Female
Negative Emotionality	NEG	60%	Female
Orderliness	ORD	62%	Female
Playfulness	PLA	50%	Female
Sociability	SOC	57%	Female
Social Intelligence	SOCIN	100%	Female

Table 6.13

Identifying which constructs were implicated by biased intercepts (continued)

Traditionalism–Religiosity	TRA	75%	Female
Warm-Heartedness	WAR	64%	Female

The 85 intercepts that remained constrained to be equal across the two sample groups did, by implication, not differ practically significantly across the two gender groups. These remaining 85 items therefore displayed a lack of uniform bias. Since strong invariance and partial scalar equivalence[55] was demonstrated, the more stringent test of strict invariance could be evaluated.

6.6.7. Strict Invariance (H_{06})

The strict invariance analysis explored whether male and female respondents responded to the items in such a way that no practically significant difference existed between samples with regard to the error variance associated with the indicator variables (Dunbar et al. 2011). This test was conducted by establishing whether the multigroup measurement model of which the model structure, the slopes, some intercepts of the regression (i.e. the intercepts that remained constrained from the previous analysis) and the error variances of the indicator variables on the latent variables were constrained to be equal, but all other model parameters were freely estimated across the gender groups (i.e. the diagonal and off-diagonal elements of the 20 x 20 Φ and the 85 elements in τ that were freed in the $H_{05,85}$ multigroup partial strong invariance measurement model), displayed reasonable fit when simultaneously fitted to the two sample groups in a multigroup analysis.

Finding support for the multigroup strict invariance model would provide support for the SAPI developers' claim that the items operate in the same manner, irrespective of the sample groups (Dunbar et al., 2011). A finding of strict invariance would imply the absence of items that suffer from error variance bias.

The strict invariance hypothesis H_{06}: RMSEA ≤ .05 was tested by fitting the multigroup SAPI measurement model across the male (N = 1834) and female (N = 2420) samples. A few of the fit statistics for the SAPI strict invariance multigroup measurement model are displayed in Table 6.6. The degrees of freedom shown in Table 6.6 for the multigroup strict invariance measurement model no longer correspond to the degrees of freedom calculated in Table 4.1 because of the unforeseen 85 elements in the tau vector that had to be freely estimated (28500 – 85 = 28415). The complete fit statistics for the strict invariance multigroup measurement model are provided in Appendix I.

The RMSEA value of .0484 indicated good model fit in the sample. These fit indicators, along with the reported p-value for the test of close fit of 1.00, revealed that the strict invariance multigroup measurement model showed good fit in the parameter.

[55] The literature on multi-group CFA lacks an appropriately nuanced terminology to differentiate between partial strong invariance that was obtained after weak invariance was obtained versus between partial strong invariance that was obtained after partial weak invariance was obtained. The finding in the current study could, more comprehensively, be termed weak - strong invariance. The same dilemma exists with regards to equivalence findings. The finding in the current study could, more comprehensively, be termed metric - partial scalar equivalence.

The normed fit index (NFI = .930), the Parsimony Normed Fit Index (PNFI = .920) and the comparative fit index (CFI = .941) all exceeded the critical cut-off of .90. The RMR in turn also remained below the threshold of .050. These results further supported the permissibility of the claim that the strict invariance multigroup measurement model showed good fit in the sample.

The SAPI demonstrated strict invariance[56] indicating that the SAPI successfully demonstrated that male and female respondents responded to the items in such a way that no practically significant difference existed between samples with regard to error variances associated with the indicator variables (Dunbar et al. 2011). The stance that the SAPI lacks error variance bias was therefore a tenable position. However, permissibility of this position would be more convincing if it could be demonstrated that a multigroup measurement model in which the error variance of items on the latent variables that they represent, were constrained to be the same across both sample groups did not fit the data practically significantly poorer than a model in which the error variances were not constrained to be equal across gender groups (i.e. than the multigroup configural invariance measurement model). Hence, conditional probability equivalence was investigated.

6.6.8. Conditional Probability Equivalence (H_{09})

The support for strict invariance allowed for investigating conditional probability equivalence (H_{09}) of the SAPI. Conditional probability equivalence evaluated whether the strict invariance model (multigroup measurement model in which the model structure, the slope and some of the intercepts of the regression of the indicator variables on the latent variables, and error variance of the indicator variables were constrained to be equal across the two sample groups) fitted practically significantly (or statistically significantly) poorer than the configural invariance model (multigroup measurement model in which only the model structure was constrained to be equal across the two sample groups).

Although the verdict of the measurement model equivalence was based on whether or not practical significance was obtained, the results for the statistical significance evaluation are nonetheless provided in Table 6.14.

Table 6.14

Statistical significance of the scaled chi-square difference statistic: a test of conditional probability equivalence

Hypotheses		Normal theory chi-square	Df	Prob normal theory chi-square diff	Statistical significance
Configural invariance model	H_{03}	166794.946	28010		
Strict invariance model	H_{06}	169832.984	28415		
Difference (H_{06}-H_{03}) conditional probability equivalence	H_{09}	3038.038		.000000	Statistical significance

If the probability of observing the normal theory chi-square difference in a multigroup sample under the null hypothesis of no difference in fit between the configural and strict invariance models in the parameters was smaller than or equal to .05, the null hypothesis was rejected in favour of the hypothesis that the fit of the

[56] The finding in the current study could, more comprehensively, be termed weak - partial strong – strict invariance.

multigroup configural invariance and multigroup strict invariance models differ in the parameter. The null hypothesis of no difference in model fit in the parameter was rejected, indicating a lack of conditional probability equivalence (i.e. a lack of equivalence of error variances) across the two gender groups when using statistical significance as the benchmark.

The practical significance of the conditional probability equivalence test was based on the cut-off levels of change in CFI, Gamma Hat and MacDonald difference as indicated by Table 6.15.

Table 6.15

Practical significance of the CFI, Gamma Hat and MacDonald difference statistics: a test of conditional probability equivalence

Model	Hypotheses	CFI	Γ1	Mc	Decision
Configural invariance model	H_{03}	.942	.838926174	8.17173E-08	
Strict invariance model	H_{06}	.941	.836371329	5.99653E-08	
Difference [H_{03}-H_{06}; Test Of Conditional Probability Equivalence]	H_{09}	-0.0010	-0.0026	-0.0000	Lack Of Conditional Probability Equivalence

The practical significance benchmark revealed conditional probability equivalence was not demonstrated with the strict invariance measurement model (multigroup measurement model in which both the model structure, the slopes and some of the intercepts of the regression of the indicator variables, and error variance of the indicator variables were constrained to be equal across the two sample groups). This indicated that some items suffered from error variance bias, causing practically significant error variance differences between the male and female groups for those items.

6.6.9. Strict Invariance-Partial Conditional Probability Equivalence (H_{06i} & H_{09i})

Due to finding a lack of conditional probability equivalence, an iterative process was followed to identify problematic items by ranking items from the largest difference between the female and male sample groups based on configural invariance model's theta-delta common metric completely standardised solution. A total of 102 items were released to be freely estimated, before partial conditional probability equivalence could be demonstrated. By relaxing the measurement error variance equality constraints on the multigroup measurement model the measurement model became a partial strict invariance measurement model. The list of problematic items for which the theta-delta parameters were freed to be estimated is provided in Appendix J.

The partial strict invariance hypothesis H_{06102}: RMSEA ≤ .05 was tested by fitting the multigroup SAPI measurement model across the male (N = 1834) and female (N = 2420) samples. A few of the fit statistics for the SAPI partial strict invariance multigroup measurement model are displayed in Table 6.6. The complete fit statistics for the multigroup partial strict invariance measurement model[57] are provided in Appendix K.

[57] This model could, more comprehensively, be termed a multigroup weak - partial strong – partial strict invariance measurement model.

The RMSEA value of .0482 indicated good model fit in the sample. These fit indicators, along with the reported p-value for the test of close fit of 1.00, revealed that the partial strict invariance multigroup measurement model showed good fit in the parameter. Again it needs to be conceded that there was never any doubt that any of the multigroup partial strict invariance measurement models would demonstrate close fit, due to H_{06}: RMSEA ≤ .05 not being rejected (p > .05). A series of multigroup partial strict invariance measurement models were fitted not because of a lack of strict invariance but because of a lack of conditional probability equivalence.

The normed fit index (NFI = .931), the Parsimony Normed Fit Index (PNFI = .917) and the comparative fit index (CFI = .942) all exceeded the critical cut-off of .90. The RMR in turn also remained below the threshold of .050. These results further supported the permissibility of the claim that the strong invariance multigroup measurement model showed good fit in the sample.

If the probability of observing the normal theory chi-square difference in a multigroup sample under the null hypothesis of no difference in fit between the configural and partial strict invariance models in the parameters was smaller than or equal to .05, the null hypothesis was rejected in favour of the hypothesis that the fit of the multigroup configural invariance and multigroup strict invariance models differ in the parameter. As indicated in Table 6.16, the null hypothesis of no difference in model fit in the parameter was still rejected for all the investigated partial strict measurement models, indicating a lack of partial conditional probability equivalence (i.e. a lack of equivalence of error variances) across the two gender groups when using statistical significance as the benchmark.

Table 6.16

Statistical significance of the scaled chi-square difference statistic: a test of partial conditional probability equivalence per item

Hypotheses		Normal theory chi-square	Df	Prob normal theory chi-square diff	Statistical significance
Configural invariance model	H_{03}	166794.946	28010		
Strict invariance model	H_{06}	169832.984	28415		
Partial strict invariance model	H_{06101}	168126.921	28314		
Partial strict invariance model	H_{06102}	168119.255	28313		
Difference (H_{06}-H_{03}) conditional probability equivalence	H_{09}	3038.038		.000000	Statistical significance
Difference (H_{06101}-H_{03}) partial conditional probability equivalence	H_{09101}	1331.975		.000000	Statistical significance
Difference (H_{06102}-H_{03}) partial conditional probability equivalence	H_{09102}	1324.309		.000000	Statistical significance

The iterative process of releasing theta-delta parameters of problematic items to be freely estimated continued until practical significance was demonstrated. Again, the cut-off levels of change in CFI, Gamma Hat and MacDonald difference were used as indicators for practical significance and provided in Table 6.17.

Table 6.17

Practical significance of the CFI, Gamma Hat and MacDonald difference statistics: a test of partial conditional probability equivalence

Model	Hypotheses	CFI	Γ1	Mc	Decision
Configural Invariance	H_{03}	.942	.838926174	8.17173E-08	
Strict invariance model	H_{06}	0.941	0.836371329	5.99653E-08	
Partial strict invariance model	H_{06101}	0.942	0.837925494	7.24043E-08	
Partial strict invariance model	H_{06102}	0.942	0.837933754	7.24768E-08	
Difference [H_{06}-H_{03}; Test Of Conditional Probability Equivalence]	H_{09}	-0.0010	-0.0026	-0.0000	Lack Of Conditional Probability Equivalence
Difference [H_{06101} –H_{03}; Test Of Partial Conditional Probability Equivalence]	H_{09101}	0.0000	-0.001001	-0.0000	Lack Of Conditional Probability Equivalence
Difference [H_{06102} –H_{03}; Test Of Partial Conditional Probability Equivalence]	H_{09102}	0.0000	-0.00099	-0.0000	Conditional Probability Equivalence

The SAPI demonstrated strict invariance-partial conditional probability equivalence indicating that the position that the SAPI measured the same construct across the two samples groups, and the position that both the male and female groups perceived and interpreted the item content for most but not all of the items in the same manner were justified positions. The stance that the SAPI lacks error variance bias was therefore not permissible, as 102 of the 170 items were shown to demonstrate error variance. Table 6.18 indicates that 16 of the 20 constructs were reported to have between 50% and 100% of items that contained biased error variances. The error variances for all of the items in *Social Intelligence* are biased towards one of the sample groups and all those items had to be released to freely estimate before metric – partial scalar - partial conditional probability equivalence was demonstrated. Constructs with more than half of the items containing biased error variances include *Achievement Orientation, Arrogance, Broad-Mindedness, Conflict-Seeking, Deceitfulness, Emotional Balance, Empathy, Epistemic Curiosity, Hostility–Egoism, Integrity, Intellect, Negative Emotionality, Playfulness, Social Intelligence, Traditionalism–Religiosity* and *Warm-Heartedness.*

Table 6.18

Identifying which constructs were implicated by biased error variances

Dimension	Code	% Items in construct with biased error variances	D value = Female average less Male average Female = Females scored higher on average Male = Males scored higher on average
Achievement Orientation	ACH	55%	Male
Arrogance	ARR	50%	Male
Broad-Mindedness	BRO	50%	Male
Conflict-Seeking	CON	57%	Male
Deceitfulness	DEC	86%	Male
Emotional Balance	EMO	50%	Female
Empathy	EMP	86%	Male
Epistemic Curiosity	EPI	83%	Male
Facilitating	FAC	30%	Male
Hostility–Egoism	HOS	64%	Male
Integrity	INTEG	73%	Male
Intellect	INT	69%	Female
Interpersonal Relatedness	INTER	44%	Male
Negative Emotionality	NEG	50%	Male
Orderliness	ORD	38%	Male
Playfulness	PLA	67%	Male
Sociability	SOC	43%	Female
Social Intelligence	SOCIN	100%	Male
Traditionalism–Religiosity	TRA	75%	Male
Warm-Heartedness	WAR	73%	Male

The 68 items that remained constrained to be equal across the two sample groups did however display a lack of error variance bias. Therefore, metric – partial scalar - partial conditional probability equivalence was demonstrated.

6.7. Conclusion

The results for the sequence of SAPI measurement invariance and measurement equivalence tests have been discussed in detail. SAPI demonstrated a lack of construct bias and a lack of non-uniform bias. However, the results revealed that a total of 85 of the SAPI items suffered from uniform bias, and 102 items

suffered from error variance bias. Therefore it is concluded that the SAPI measured the same construct for across the two samples groups, but the item content of the some items were perceived and interpreted differently between the two gender groups. Therefore, metric – partial scalar - partial conditional probability equivalence was demonstrated.

CHAPTER 7: DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

7.1. Introduction

Organisations strive to increase long term sustainability through wise utilisation of profit, planet, (i.e. the environment in which the companies function) and people (i.e. human resources). Since the work performance of human resources play a pivotal role in the achievement of these objectives, HR departments set out to support these organisational objectives through various strategies, including selection and talent management strategies. By selecting and developing candidates that best fit the job and organisation, selection and talent management strategies aim to positively impact work-related behaviours and organisational outcomes such as increased task performance, group success, organisational citizenship behaviour, job satisfaction and leadership effectiveness, as well decreased counterproductive behaviour, turnover, absenteeism, tardiness (Barrick & Mount, 2005; Hough, 2003). Personality constructs have been successfully shown to predict these work-related behaviours and organisational outcomes, making personality assessments sought after instruments in selection and talent management HR strategies.

To ensure that predicted job performance inferred from personality assessments are permissible[58], such instruments (and the manner in which they are used) have to comply with relevant legislation (such as the EEA) that governs the use of psychological testing. Regulatory changes have placed test developers under pressure to subject personality assessments, and the manner in which they are used to inform decision-making, to sophisticated scientific analyses by assessing the psychometric appropriateness and relevance of assessments and the inferences derived from assessments to the South African context. For instance, test developers are required to empirically test the permissibility of their claim that observed scores from personality assessment have the same meaning in terms of the underlying latent variables across different groups and that the instrument is not biased against any group, by investigating the assessment's measurement invariance and measurement equivalence.

Chapter 2 provided a detailed literature review on the development of the SAPI. This locally developed personality assessment is classified by the HPCSA as a personality assessment still under development. All research on the SAPI, including the findings from the current study, therefore contribute to the arsenal of sophisticated scientific analyses that the assessment has to be subjected to in order provide evidence for whether or not the inferences derived from the SAPI are appropriate and generalizable across different sample groups. The current study is the first to apply the taxonomy recommended by Dunbar et al. (2011), to investigate the measurement invariance and equivalence of the SAPI.

The purpose of the current study was twofold. Firstly, the study set out to investigate the SAPI's measurement invariance which analyses whether the multigroup SAPI measurement model implied by the design intention of the test developers and their constitutive definition of the personality construct as

[58] The term *permissible* is used here in a wider sense than predictive validity. It is used here to refer to the extent to which the selection decision-making that is based on (clinical or mechanical) job performance inferences can be justified to all stakeholders involved (the developers of the selection procedure, its organisational users, management, the applicants that are affected by it, organised labour and the state).

reflected in the SAPI scoring key, fitted the archival instrument data from male and female groups at least reasonably well, when a series of increasing constraints were imposed on the multigroup measurement model via a series of multigroup confirmatory factor analyses conducted on the data. The second objective of the study was to investigate the SAPI's measurement equivalence. This was done by exploring whether the multigroup SAPI measurement models in which the model structure and specific parameters were constrained to be equal across genders, fitted significantly poorer than a multigroup SAPI measurement model in which only the structure was constrained to be equal across genders.

This chapter aims to discuss the research findings. The implications for various stakeholders such as the SAPI test developers, the broader research community, and HR practitioners are elaborated on. The chapter concludes with several limitations and recommendations for future research.

7.2. Findings

Both single-group measurement models in the sample demonstrated good model fit with RMSEA values below .05 (RMSEA for the female group = 0.048; RMSEA for the female group = 0.0492). The close fit null hypothesis was not rejected for both the male and female samples ($p > .05$). However, the inconsistency between the standardised residuals and the fit statistics suggested mediocre fitting single group models. Nevertheless, not rejecting the null hypothesis of close fit for both the male and female samples permitted the proceeding to the testing of the multigroup measurement models.

The multigroup configural invariance model demonstrated good fit in the (combined) sample with a RMSEA value of .0483. The close fit null hypothesis was not rejected ($p > .05$). This provided evidence supporting the permissibility of the test developers' claim that the SAPI successfully measures the same personality construct across the male and female groups. The SAPI therefore does not suffer from gender-based construct bias. Finding evidence for similar structures across the two sample groups allowed for the testing of the weak invariance multigroup measurement model.

The SAPI multigroup weak invariance measurement model demonstrated good model fit (RMSEA = .0483) in the sample. The close fit null hypothesis was not rejected ($p > .05$) for the multigroup weak invariance measurement model. The multigroup weak invariance measurement model did not fit practically significantly poorer than the multigroup configural invariance model. Metric equivalence was consequently concluded. Hence, the SAPI measured the same construct across the two samples groups, and strong evidence existed that gender did not moderate the effect of the latent first-order personality dimensions on the item responses. Based on these analyses, the stance that the SAPI lacks non-uniform (i.e. slope) item bias is therefore a convincing position. Finding metric equivalence allowed for evaluating the multigroup SAPI measurement model in terms of the more stringent strong invariance test.

The SAPI demonstrated strong invariance with good model fit (RMSEA = .0494) in the sample. The close fit null hypothesis was also not rejected ($p > .05$). The finding strong invariance provided weak evidence of the absence of uniform (i.e. intercept) bias. This suggested that the SAPI measured the same construct across the two samples groups, and that gender did not exert a main effect on the item responses when statistically controlling for the latent first-order personality dimensions. Strong invariance, however, provided only weak

evidence of the absence of a gender main effect on any of the SAPI items. Testing for strong invariance presented the multigroup SAPI measurement model with a relatively lenient test. Yet, when putting the permissibility of the stance that the SAPI lacks uniform bias to a more stringent test, scalar equivalence could not be achieved. The multigroup strong invariance measurement model fitted practically significantly poorer than the configural invariance measurement model. This indicated that some items suffered from uniform bias. A subsequent iterative process identified problematic items by ranking the items based on the largest difference between the female and male sample groups for the configural invariance model's unstandardised tau estimates. Partial strong invariance-partial scalar equivalence was achieved once the equality constraint was released on 85 items to allow those tau estimates to be freely estimated in group 2. The stance that the SAPI lacks uniform bias is therefore not permissible, as 85 of the 170 items demonstrated uniform bias. The remaining 85 items therefore did display a lack of uniform bias, by not differing practically significantly in terms of intercept across the two gender groups. Since strong invariance and partial scalar equivalence was demonstrated, the more stringent test of strict invariance could be evaluated.

The SAPI demonstrated strict invariance. The RMSEA value of .0484 indicated good model fit in the sample. The multigroup strict invariance measurement model showed close fit ($p > .05$). The finding of strict invariance provided weak evidence of the absence of error variance bias. This implied that the item measurement error variances did not differ across gender groups. The probability of a specific item response (or larger), given a specific standing on a latent first-order personality dimension, did not differ across gender groups. Strict invariance, however, provided weak evidence of the absence of error variance differences across gender groups on any of the SAPI items. Testing for strict invariance presented the multigroup SAPI measurement model with a relatively lenient test. When putting the position that the SAPI lacks error variance bias to the more stringent test, a lack of conditional probability equivalence was revealed. The multigroup strict invariance measurement model fitted practically significantly poorer than the multigroup configural invariance measurement model. Finding a lack of conditional probability equivalence implies that under closer inspection some of the SAPI items displayed error variance bias. Once again an iterative process identified the problematic items by ranking items from the largest difference between the female and male sample groups, based on configural invariance model's theta-delta common metric completely standardised solution. For a total of 102 items the measurement error variance term were released to be freely estimated, before partial conditional probability equivalence could be demonstrated. Strict invariance - partial conditional probability equivalence was consequently demonstrated by relaxing the measurement error variance equality constraints on the multigroup measurement model. Therefore, the position that the SAPI measured the same construct across the two sample groups and the position that both the male and female groups perceived, interpreted and responded to the item content for some, but not all of the items in the same manner were tenable positions. The SAPI lacks error variance bias in 102 of the 170 items, whilst only 68 items displayed a lack of error variance bias. Therefore, metric – partial scalar - partial conditional probability equivalence was demonstrated.

7.4. Implications and Recommendations for Future Research

The current study used multigroup confirmatory factor analysis to fit a single multidimensional (multigroup) measurement model containing all 20 first-order personality dimensions. This raises the question to what extent the results that were obtained were dependent on the specific methodology that was used. More

specifically it raises the question whether essentially the same items would be flagged as biased if a different analysis technique would be used (like multigroup item response theory for example). It secondly raises the question whether the use of multigroup confirmatory factor analysis to fit multiple unidimensional (multigroup) measurement models each containing a single first-order personality dimension would render essentially the same results. These questions should be examined in future research studies that utilise the same data that was used in the current study[59].

Finding a lack of scalar and conditional probability equivalence in many of the SAPI items leaves the question as to what the practical implications are for practitioners and researchers. The test developers are left with various options to address the biased SAPI items identified in the current study. The first option is to delete the problematic items. The first option is, however, unattractive since it would result in too few or even no items left to measure some of the dimensions. Another option is to adjust the item content or the wording of the item and evaluate the new version of the SAPI once again with recommended analyses such as item analyses, exploratory factor analysis, validation studies, as well as measurement invariance and measurement equivalence using confirmatory factor analysis. This option is, however, also unattractive because there are no guarantees that the revised items will be free from bias. Adjusting the item content or the wording of the item is difficult because the revision of the items is not rooted in a clear diagnosis as to why the original items displayed uniform and/or error variance bias. Adjusting the item content or the wording of the item without a clear root cause analysis as to why uniform and/or error variance bias permeates the SAPI can therefore be regarded as essentially nothing more than an optimistic and naïve attempt to address the bias.

The latent first-order personality dimensions whose measurement were impacted most by the items that suffered from uniform bias and/or error variance bias were *Arrogance*, *Deceitfulness, Empathy, Social Intelligence* and *Traditionalism-Religiosity*. *Social intelligence* was the only latent first-order personality dimension of which all items comprising its subscale suffered from both uniform bias and error variance bias. Therefore, to achieve metric – partial scalar - partial conditional probability equivalence the intercepts and error variance in all the items that were designated to reflect the latent first-order *Social Intelligence* personality dimension had to be released to be freely estimated. All of the items designated to reflect the other 4 latent first-order personality dimensions (*Arrogance*, *Deceitfulness, Empathy* and *Traditionalism-Religiosity*) suffered from either uniform bias and/or error variance bias (but not always both). Nonetheless the net effect was that all the items comprising these four subscales eventually had to have the intercepts and/or error variances freely estimated to achieve metric - partial scalar- partial conditional probability equivalence.

Appendix L provides a combination of the information presented in Table 6.13 and Table 6.18. Each table also indicates which gender group was favoured by the biased items in the various subscales designed to reflect test-takers standing on the latent first-order personality dimensions. The latter indication was provided to acknowledge that inferences or decisions are not based on the SAPI individual items, but rather on dimension scores calculated by summing (as per the scoring key) the individual item scores over the items of

[59] The contributions of the internal examiner is gratefully acknowledged.

the respective subscales. The question is therefore whether the bias in the item scores accumulate over items to create bias in the dimension score (this would imply that the bias on the item level tends to consistently favour a specific group) or whether the bias in the item scores cancel each other out over items to create no bias in the dimension score (this would imply that the bias on the item level tends to favour one group as much as the other). To determine whether the item bias tended to accumulate across the items of the subscales or rather tended to cancel each other out, the average[60] (algebraic) difference in tau and theta-delta was calculated across the items of each subscale.

This, however, then begs the question how the differences in tau (intercept) and theta-delta (error variance) accumulate to affect the regression of the dimension score on the latent trait. Regarding the differences in intercept the current study argued that the difference in intercept of the regression of the dimension score (as the sum of the item scores) on the latent trait measured by the subscale in question will be (approximately) equal to the sum of the differences in the intercepts of the regression of the item scores on the latent trait. If a specific group therefore tended to be advantaged on the item level in that $E[X_{ij}|\xi_j; \gamma]$ was higher across i in the j^{th} subscale and j^{th} latent first-order personality dimension the same will be true on the dimension score in that $E[X_{tot\_j}|\xi_j; \gamma]$ will be higher. In the case of the error variances it can be assumed that if $S^2_{Xi|\xi;\,\gamma}$ tends to be higher for a specific group across all i, the same will be true for $S^2_{X\_tot|\xi;\,\gamma}$[61]. The question is how the larger error variance on the dimension score level will affect the probability of achieving a specific dimension score X_c (or larger) given a specific standing on the latent first-order personality dimension? Assuming that the items of subscale j do not suffer from non-uniform or uniform bias but only from error variance bias and it is assumed that the error variance for the female group is consistently larger than for the male group across all i, then it can be assumed that the regression of the dimension score on the latent first-order personality dimension coincides. That then implies that the mean of the conditional dimension score distributions coincide for male and female, but that $S^2_{X\_tot|\xi;\,Female}$ is larger than $S^2_{X\_tot|\xi;\,Male}$. This in turn would imply that whatever X_c is it will translate to a more extreme positive or negative z score (assuming that the conditional dimension score distribution follows a normal distribution) for the group with the smaller error variance (the male group in the case of this argument). Given that the probability in question is the probability whether a specific dimension score Xc (*or larger*) given a specific standing on the latent first-order personality dimension will be obtained the group with the larger error variance will be advantaged (and the group with the smaller error variance disadvantaged) with a larger probability if X_c transforms to a positive z-score (i.e., X_c is larger than $E[X_{tot}|\xi]$). But the group with the larger error variance will be disadvantaged (and the group with the smaller error variance advantaged) if Xc transforms to a negative z-sore (i.e., X_c is smaller than E[Xtot|ξ]). When Xc transforms to a z-score of zero (i.e., X_c is equal to E[Xtot|ξ]) no group is advantaged or disadvantaged. This is illustrated in Figure 7.1.

[60] The researchers did consider the calculation of the sum of the algebraic D scores across the items of each subscale.
[61] It is thereby not implied that the error variance of the dimension scores will simply be equal to the sum of the item error variances.

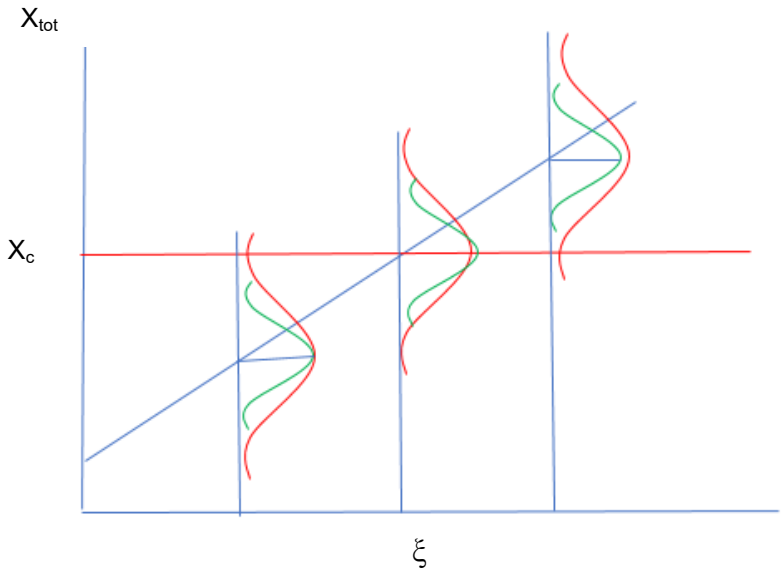

Figure 7.1. Illustrating the effect of error variance bias on the dimension score level (assuming the absence of uniform and non-uniform bias)

The effect of item bias on dimension score bias and the manner in which these different forms of bias on the item level combine to affect bias on the dimension score level and how this affects $E[X_{tot\_j}|\xi_j; \gamma]$ but especially $P(X_{tot} \geq X_c|\xi; \gamma)$, is clearly substantially more intricate when considering the combined effect of the different forms of bias on the dimension score level.

The aforementioned line of reasoning argue that error variance bias in first-order dimensions that measure positive traits advantage the group whose average score on the dimension was smaller whilst the error variance bias in first-order dimensions that measure negative traits advantage the group whose average score on the dimension was larger. Therefore the first-order dimensions that advantaged the female group with both uniform bias and error variance included *Achievement Orientation, Empathy, Facilitating, Integrity, Interpersonal Relatedness, Orderliness, Playfulness, Social Intelligence, Traditionalism–Religiosity* and *Warm-Heartedness.* The first-order dimensions that in turn advantaged the male group with both uniform bias and error variance included *Arrogance, Conflict-Seeking, Deceitfulness, Emotional Balance, Hostility-Egoism* and *Intellect.* The only first-order dimensions that did not advantage a single group with both uniform bias and error variance are *Broad-mindedness, Epistemic Curiosity, Negative Emotionality* and *Sociability.*

Future research could investigate why item bias seems to occur in a particular direction across the items of subscales developed to measure specific latent first-order personality dimensions. In refection on this question it needs to be reiterated at the outset that bias on both the item level and the dimension score level means that either: (i) the expected item score (or dimension score), conditional on the test-takers standing on the latent first-order personality dimension, and/or (ii) the probability of achieving a specific observed item (or dimension) score conditional on the test-takers standing on the latent first-order personality dimension, differs for male and female test-takers when their standing on the latent first-order personality dimension is controlled. Stated more simply, uniform and/or non-uniform bias on the dimension score level means that male and female test-takers with the same standing on the latent first-order personality dimension in question typically (i.e., on average) do not obtain the same observed dimension scores. Likewise uniform

and/or non-uniform bias and/or error variance bias on the dimension score level means that male and female test-takers with the same standing on the latent first-order personality dimension in question do not have the same probability of achieving a specific dimension score or higher. Nothing is implied about group differences on the latent first-order personality dimension in question. Whether a dimension score is biased or not says nothing about the location of the group-specific latent first-order personality dimension distributions. Uniform and/or non-uniform bias on the dimension level would mean that there is variance in the observed dimension scores that cannot be explained in terms of variance in the latent first-order personality dimension. The group-specific dimension score distributions then differ more than they can be expected to differ given the difference in the group-specific latent first-order personality dimension distributions. Rather than measurement bias suggesting anything about group differences on the latent first-order personality dimensions, the absence of measurement bias on the dimension level is a necessary prerequisite for making any definite conclusions about group differences on the latent first-order personality dimensions from statistically significant differences ($p<.05$) in the mean observed first-order dimensions scores. The justification for the current study lies in part in this fact.

The foregoing line of reasoning would suggest that a trend for the items of any subscale to suffer from uniform, non-uniform and/or error variance bias and to favour a specific group cannot be explained in terms of differences in the latent first-order personality dimension being measured by the subscale in question. To reflect on possible explanations for measurement bias in the items of specific subscales that favour a specific group, the definitions of the various forms of item bias need to be revisited. Uniform bias exists when gender, as a main effect, explains variance in the observed scores when controlling for the latent first-order personality dimension in question. Non-uniform bias exists when gender, in interaction with the latent first-order personality dimension, explains variance in the observed scores when controlling for the latent first-order personality dimension in question. Error variance bias exists when the variance in the observed scores that are not explained by the latent first-order personality dimension in question (i.e., the residual variance in the observed scores), differs across gender groups. The residual variance is the result of non-relevant systematic sources of variance and random error not related to the latent first-order personality dimension in question. When the error variance is larger for one gender group than for another, this implies that these of non-relevant systematic sources of variance and random error influences operate more aggressively for one gender group than for another. This in turn implies that gender moderates the effect of these of non-relevant systematic sources of variance on the observed responses. Most likely, however, it is not gender per se that acts as main effect or moderator variable but rather one or more latent variables systematically related to gender. This then could mean that latent first-order personality dimensions on which genders differ significantly ($p<.05$) can explain variance in the observed scores of another latent first-order personality dimension when controlling for that latent first-order personality dimensions. One likely reason for this might be that cultural expectations imposed on males and females as to what constitutes being masculine and feminine, might induce individuals to respond somewhat differently to test stimuli even when controlling for the latent first-order personality dimension being assessed. Males and females might therefore respond systematically differently to items of the *Arrogance* subscale even when controlling for differences in the latent *Arrogance* personality dimension, because of the cultural demand for females to appear more humble than males. This could then manifest itself in uniform bias in items of the *Arrogance* scale. Future studies should explore the impact that such cultural expectations might have on the gender differences between

Achievement Orientation, Empathy, Facilitating, Integrity, Interpersonal Relatedness, Orderliness, Playfulness, Social Intelligence, Traditionalism–Religiosity, Warm-Heartedness, Arrogance, Conflict-Seeking, Deceitfulness, Emotional Balance, Hostility-Egoism and *Intellect.*

It has been stated that inferences are based on dimension scores and not item scores. Measurement bias studies utilising multigroup confirmatory factor analysis are, however, typically conducted on the item level. There is ample literature recommending practical and statistical significance levels to evaluate measurement invariance and equivalence on the item level, such the Cheung and Rensvold (2002) decision rule to evaluate the practical significance of the difference in model fit or the Satorra-Bentler scaled difference test. The former informed the decision on which items displayed uniform bias by evaluating when the difference in fit between the multigroup configural invariance measurement model and the multigroup partial strong invariance measurement model in which specific intercepts were allowed to be freely estimated in both gender groups (based on the rank ordering of the items in terms of absolute differences in the item intercepts) was no longer practically significant. The former also informed the decision on which items displayed error variance bias by evaluating when the difference in fit between the multigroup configural invariance measurement model and the multigroup partial strong partial strict invariance measurement model in which specific error variances were allowed to be freely estimated in both gender groups (based on the rank ordering of the items in terms of absolute differences in the item measurement error variances) was no longer practically significant. A need exists to apply the same methodology to detect bias on the dimension score level. The problem, however is that when a first-order measurement model would be fitted in which each latent first-order (personality) dimension would be represented/operationalised by a single observed dimension scores the resultant measurement model would be under-identified with negative degrees of freedom where the number of unique elements in the observed variance-covariance matric are less than the number of freed parameters in the model (Diamantopoulos & Siguaw, 2000). In the case of the single-group first-order SAPI measurement model in which the latent first-order personality dimensions are operationalised with the subscale dimension scores and the twenty latent first-order personality dimensions are allowed to correlate, 20 λ^{X}_{ij}, 20 $\theta_{\delta ii}$, 20 τ_i and 190 ϕ_{jk} need to be estimated (250 freed parameters in total) whilst there are only 210 unique variance and covariance terms in the observed covariance matrix. In the case of the multigroup configural invariance (first-order) measurement model there would be 500 freed parameters whilst there are only 420 unique variance and covariance terms in the observed covariance matrices[62]. An under-identified single-group or multigroup measurement model cannot be fitted (Diamantopoulos & Siguaw, 2000). However, if the second-order SAPI would be fitted with 20 dimension score indicator variables representing 20 latent first-order personality dimensions, which in turn load onto five second-order latent personality dimensions, then in this single-group second-order measurement model 20 λ^{Y}_{ij}, 20 $\theta_{\varepsilon ii}$, 20 τ_i, and 20 γ_{jp} need to be estimated (80 freed parameters in total) whilst there are 210 unique variance and covariance terms in the observed covariance matrix (130 degrees of freedom). In the case of the multigroup configural invariance (second-order) measurement model there would be 160 freed parameters whilst there are 420 unique variance and covariance terms in the observed covariance matrices (260 degrees of freedom). Multigroup second-order measurement models can now be fitted in which λ_{ij}, τ_i, and $\theta_{\varepsilon ii}$, are constrained to be equal across groups. If a lack of invariance is obtained, essentially the same

[62] The identification problem will not be solved in the multigroup weak, strong strict invariance measurement models either.

procedure as was used in the current study can be used to detect the biased dimension scores that caused the lack of invariance. The tests of the various forms of equivalence also remain the same. The measurement invariance and equivalence taxonomy proposed by Dunbar et al. (2011) will, however, have to be expanded to make provision for possible group differences in the loading of the first-order factors on the second-order factors (i.e. group differences in γ_{jp}).

Biased items in the subscales can be deleted or retained. If they are retained the bias can accumulate to some degree in the dimension score or the bias can effectively cancel itself out in the dimension score. If the bias to some degree accumulates in the dimension score, the question arises whether the bias in the dimension score is practically significant. In considering this question the distinction between construct-referenced inferences derived from dimension scores and criterion-referenced inferences needs to be taken into account. From a construct-referenced dimension score interpretation point of view, practitioners transform the observed dimension scores to norm scores (i.e., percentile ranks, z-scores, stanine, sten or McCall T-scores) in the norm group that describe the relative position of the observed dimension scores in the normative distributions. The critical question is whether the inclusion of the biased items, and their cumulative effect on the dimension score, has an effect on the norm score to which the dimension scores translate. This question is especially pertinent in the case of the less finely-grained norm scales like the sten and stanine scales where ranges to dimension scores transform to the same norm score. The change in the dimension score due to the cumulative effect of the (uniform and/or non-uniform) item bias might therefore not be enough to shift the construct-referenced interpretation of the (biased) dimension score into a different norm score category. This question would be easier to investigate if bias on the dimension score level could be investigated via a second-order SAPI measurement model.

Criterion-referenced norm group interpretation is an alternative option to interpret the observed dimension scores. The question to explore then becomes whether the (selection) decision-making based on these criterion referenced norm scores can be considered fair. SIOP recognises that the term *fairness* can be interpreted in different ways. SIOP distinguishes between the following interpretations of the term: an assessment should ensure equal group outcomes, equal treatment of all groups, comparable opportunities to learn as well as a lack of predictive bias (Society for Industrial and Organizational Psychology Inc., 2003). There are several models that consider the fairness in selection models. The Cleary fairness model in particular argues that a selection model should be regarded as unfair when a common regression equation is used that results in systematic prediction errors, which occurs when the regression lines for multiple groups differ in terms of slope and/or intercept, but that this is not taken into consideration. Practitioners are advised to develop (valid) measures of employee performance (e.g. performance appraisal instruments such as Behavioural Observation Scales) and then evaluate the predictive validity of the SAPI. By taking group membership into consideration on the dimension level, would assist with avoiding systematic group-related prediction errors when bias in the observed dimension scores caused the regression of Y on X_i to shift in terms of intercept and/or slope parameters. Should future studies as previously recommended find real differences between the gender groups in the latent dimensions, practitioners would still be able to apply the Cleary fairness model and thereby prevent unfair discrimination when using the SAPI.

From a criterion-referenced dimension score interpretation point of view, practitioners transform the observed dimension score to a norm score (i.e. expected criterion score or probability of success) in the norm group that describes the relative expected position of the test-taker in the criterion distribution. Bias in the dimension scores bring about a shift in the location group-specific predictor distributions. This shift can potentially cause the regression of the criterion on the predictor to differ in intercept and/or slope across the gender groups. If the difference in the regression of the criterion on the predictor is not taken into account the criterion inferences will suffer from gender-based predictive bias. Such predictive bias will systematically disadvantage the members of the gender-group whose criterion performance is systematically underestimated. Selection decision-making based on such predictively bias criterion inferences will then constitute indirect unfair discrimination against members of the gender group whose criterion performance is systematically underestimated. Selection decision-making based on criterion-referenced inferences need not, however, unavoidably suffer from predictive bias. If the differences in the regression of the criterion on the predictor are taken into account when deriving the criterion inferences, the selection decision-making will be fair despite the measurement bias in the dimension scores. Taking group membership into consideration when deriving criterion inferences from dimension scores could prevent the deletion of biased items

The current study only investigated the SAPI measurement invariance and measurement equivalence for male and female sample groups. The connotative meaning of a construct (like personality) lies in the internal structure of the construct (i.e. in the number and identity of the dimensions that constitute the construct and the manner in which they are related (correlationally and/or structurally) to each other), but also in the manner in which the construct is embedded in a larger nomological network. The current study demonstrated that the multi-group configural invariance measurement model showed close fit ($p>.05$). This constitutes necessary but not sufficient evidence to conclude that the SAPI does not suffer from construct bias. To be in a stronger position to claim a lack of construct bias in the SAPI, the test developers have to prove that the assessment also demonstrates structural invariance and structural equivalence. It is recommended that future research investigates whether the SAPI demonstrates structural invariance and structural equivalence between female and male groups.

Lastly, it is strongly recommended that future research follow suit by applying Dunbar et al.'s (2011) proposed taxonomy when testing for measurement invariance and measurement equivalence, to contribute in creating a taxonomy convention for these analyses.

7.3. Limitations to the study

The following limitations to the current study are acknowledged. Firstly, the demographics of the female group are slightly skewed towards the 20-39 year old female group, and thereby not fully representative of the SA population above 40 years. Secondly, the study used archival data as provided by the test publishers, resulting in a non-random sample. Ideally the study should have used a randomly selected sample to exclude the drawbacks from non-randomised samples. Lastly, Maximum Likelihood estimation was applied in the current study, instead of Robust Maximum Likelihood estimation which is generally recommended as the preferred estimation technique when dealing with data that is not normally distributed. RML demanded

too much computer memory capacity to calculate the inverse of the asymptotic covariance matric and the Satorra-Bentler chi-square statistic that needs to be calculated.

The literature study failed to pick up an important and influential research article by Van Aarde, Meiring and Wiernik (2017) on a South African meta-analytic study on the validity of job performance inferences derived from measures of the Big Five personality traits. This is acknowledged as a limitation[63].

7.5. Conclusions

The SAPI is a personality assessment developed locally in South Africa. The current study evaluated the SAPI in terms of measurement invariance and measurement equivalence between two gender groups, and demonstrated metric – partial scalar - partial conditional probability equivalence. It can therefore be concluded that the SAPI measured the same constructs across the two samples groups, but the item content of the some items were perceived and interpreted differently between the two gender groups. Several recommendations were provided to assist future studies with

The current study contributes to the wide array of research that investigates the SAPI's psychometric properties, by evaluating measurement invariance and measurement equivalence between male and female groups using structural equation modelling. It contributes to measurement invariance and measurement equivalence literature by applying the relatively recent taxonomy as proposed by Dunbar et al. (2011) on the SAPI. Lastly the current study contributes to measurement invariance and measurement equivalence literature by indicating several neglected areas in the analyses that require attention by future research studies.

[63] The external examiner is thanked for bringing this article under the attention of the researcher

APPENDIX A: DESCRIPTIVE ITEM STATISTICS

MALE SAMPLE

	N		Mean	Median	Mode	Std. Deviation	Variance	Skewness	Std. Error of Skewness	Kurtosis	Std. Error of Kurtosis	Minimum	Maximum
	Valid	Missing											
ACH_1	1834	0	3.95856	4.00000	4.000	.898216	.807	-.773	.057	.457	.114	1.000	5.000
ACH_2	1834	0	3.89422	4.00000	4.000	.911669	.831	-.759	.057	.504	.114	1.000	5.000
ACH_3	1834	0	3.73937	4.00000	4.000	.975258	.951	-.703	.057	.204	.114	1.000	5.000
ACH_4	1834	0	4.13195	4.00000	4.000	.683397	.467	-.666	.057	1.259	.114	1.000	5.000
ACH_5	1834	0	3.97110	4.00000	4.000	.818988	.671	-.722	.057	.719	.114	1.000	5.000
ACH_6	1834	0	3.67612	4.00000	4.000	.992502	.985	-.473	.057	-.340	.114	1.000	5.000
ACH_7	1834	0	4.13141	4.00000	4.000	.683901	.468	-.705	.057	1.521	.114	1.000	5.000
ACH_8	1834	0	4.23828	4.00000	4.000	.694799	.483	-.740	.057	.938	.114	1.000	5.000
ACH_9	1834	0	4.00981	4.00000	4.000	.705300	.497	-.630	.057	1.170	.114	1.000	5.000
ACH_10	1834	0	4.12650	4.00000	4.000	.693140	.480	-.803	.057	1.759	.114	1.000	5.000
ACH_11	1834	0	4.13577	4.00000	4.000	.658651	.434	-.553	.057	1.037	.114	1.000	5.000
ARR_1	1834	0	2.32552	2.00000	2.000	.981739	.964	.562	.057	-.139	.114	1.000	5.000
ARR_2	1834	0	2.03544	2.00000	2.000	.937401	.879	.777	.057	.232	.114	1.000	5.000
ARR_3	1834	0	2.02290	2.00000	2.000	.888503	.789	.946	.057	1.011	.114	1.000	5.000
ARR_4	1834	0	2.40513	2.00000	2.000	1.085537	1.178	.520	.057	-.468	.114	1.000	5.000
ARR_5	1834	0	2.03817	2.00000	2.000	.966805	.935	.798	.057	.122	.114	1.000	5.000
ARR_6	1834	0	2.11668	2.00000	2.000	.988485	.977	.844	.057	.291	.114	1.000	5.000
BRO_1	1834	0	3.99564	4.00000	4.000	.902795	.815	-.789	.057	.356	.114	1.000	5.000
BRO_2	1834	0	3.89204	4.00000	4.000	.901179	.812	-.664	.057	.173	.114	1.000	5.000
BRO_3	1834	0	4.21429	4.00000	4.000	.665897	.443	-.761	.057	1.644	.114	1.000	5.000
BRO_4	1834	0	3.99182	4.00000	4.000	.845022	.714	-.745	.057	.406	.114	1.000	5.000
BRO_5	1834	0	3.75900	4.00000	4.000	.833517	.695	-.379	.057	.019	.114	1.000	5.000
BRO_6	1834	0	4.30916	4.00000	4.000	.748049	.560	-1.167	.057	1.900	.114	1.000	5.000

CON_1	1834	0	2.87786	3.00000	3.000	1.050976	1.105	.036	.057	-.685	.114	1.000	5.000
CON_2	1834	0	1.60960	1.00000	1.000	.758466	.575	1.342	.057	2.210	.114	1.000	5.000
CON_3	1834	0	1.84242	2.00000	2.000	.804875	.648	.997	.057	1.282	.114	1.000	5.000
CON_4	1834	0	2.67830	3.00000	2.000	1.031473	1.064	.235	.057	-.676	.114	1.000	5.000
CON_5	1834	0	1.84624	2.00000	1.000	.888994	.790	1.024	.057	.808	.114	1.000	5.000
CON_6	1834	0	1.98364	2.00000	2.000	.870035	.757	.897	.057	.725	.114	1.000	5.000
CON_7	1834	0	2.41658	2.00000	2.000	1.016432	1.033	.286	.057	-.679	.114	1.000	5.000
DEC_1	1834	0	1.79280	2.00000	1.000	.889799	.792	1.203	.057	1.357	.114	1.000	5.000
DEC_2	1834	0	2.88931	3.00000	3.000	.996593	.993	-.128	.057	-.636	.114	1.000	5.000
DEC_3	1834	0	2.78408	3.00000	2.000	1.142469	1.305	.221	.057	-.812	.114	1.000	5.000
DEC_4	1834	0	2.42312	2.00000	2.000	.971964	.945	.456	.057	-.256	.114	1.000	5.000
DEC_5	1834	0	1.94057	2.00000	2.000	.890797	.794	.979	.057	.878	.114	1.000	5.000
DEC_6	1834	0	1.95474	2.00000	1.000	.999521	.999	.987	.057	.412	.114	1.000	5.000
DEC_7	1834	0	2.43839	2.00000	2.000	.915270	.838	.473	.057	-.046	.114	1.000	5.000
EMO_1	1834	0	4.20229	4.00000	4.000	.803684	.646	-1.025	.057	1.374	.114	1.000	5.000
EMO_2	1834	0	3.92966	4.00000	4.000	.806387	.650	-.653	.057	.559	.114	1.000	5.000
EMO_3	1834	0	3.73337	4.00000	4.000	.946068	.895	-.681	.057	.190	.114	1.000	5.000
EMO_4	1834	0	4.05998	4.00000	4.000	.707455	.500	-.780	.057	1.674	.114	1.000	5.000
EMO_5	1834	0	3.75245	4.00000	4.000	.889906	.792	-.499	.057	.034	.114	1.000	5.000
EMO_6	1834	0	3.95911	4.00000	4.000	.761861	.580	-.880	.057	1.719	.114	1.000	5.000
EMO_7	1834	0	3.65540	4.00000	4.000	.849316	.721	-.615	.057	.459	.114	1.000	5.000
EMO_8	1834	0	3.99618	4.00000	4.000	.881703	.777	-.925	.057	1.019	.114	1.000	5.000
EMP_1	1834	0	3.96129	4.00000	4.000	.726796	.528	-.615	.057	.893	.114	1.000	5.000
EMP_2	1834	0	4.03544	4.00000	4.000	.661751	.438	-.660	.057	1.780	.114	1.000	5.000
EMP_3	1834	0	3.90022	4.00000	4.000	.829338	.688	-.703	.057	.589	.114	1.000	5.000
EMP_4	1834	0	3.77863	4.00000	4.000	.821865	.675	-.692	.057	.537	.114	1.000	5.000
EMP_5	1834	0	3.92585	4.00000	4.000	.740061	.548	-.787	.057	1.447	.114	1.000	5.000
EMP_6	1834	0	4.15540	4.00000	4.000	.634828	.403	-.667	.057	1.824	.114	1.000	5.000
EMP_7	1834	0	3.63413	4.00000	4.000	.840045	.706	-.692	.057	.613	.114	1.000	5.000

EPI_1	1834	0	4.35005	4.00000	5.000	.733154	.538	-1.277	.057	2.694	.114	1.000	5.000
EPI_2	1834	0	4.30207	4.00000	4.000	.729943	.533	-1.075	.057	1.728	.114	1.000	5.000
EPI_3	1834	0	4.20284	4.00000	4.000	.619974	.384	-.591	.057	1.738	.114	1.000	5.000
EPI_4	1834	0	4.38986	4.00000	4.000	.586375	.344	-.533	.057	.490	.114	1.000	5.000
EPI_5	1834	0	4.36150	4.00000	4.000	.566053	.320	-.399	.057	.828	.114	1.000	5.000
EPI_6	1834	0	4.45147	5.00000	5.000	.687421	.473	-1.288	.057	2.230	.114	1.000	5.000
FAC_1	1834	0	3.72410	4.00000	4.000	.827248	.684	-.590	.057	.450	.114	1.000	5.000
FAC_2	1834	0	3.71919	4.00000	4.000	.879662	.774	-.582	.057	.181	.114	1.000	5.000
FAC_3	1834	0	3.63740	4.00000	4.000	.808396	.654	-.447	.057	.367	.114	1.000	5.000
FAC_4	1834	0	3.72683	4.00000	4.000	.742351	.551	-.371	.057	.428	.114	1.000	5.000
FAC_5	1834	0	3.85387	4.00000	4.000	.743342	.553	-.684	.057	.991	.114	1.000	5.000
FAC_6	1834	0	3.97219	4.00000	4.000	.702882	.494	-.669	.057	1.255	.114	1.000	5.000
FAC_7	1834	0	3.77590	4.00000	4.000	.774278	.600	-.543	.057	.573	.114	1.000	5.000
FAC_8	1834	0	3.49237	4.00000	4.000	.867172	.752	-.238	.057	-.142	.114	1.000	5.000
FAC_9	1834	0	3.88550	4.00000	4.000	.734154	.539	-.563	.057	.664	.114	1.000	5.000
FAC_10	1834	0	3.96783	4.00000	4.000	.706567	.499	-.586	.057	.840	.114	1.000	5.000
HOS_1	1834	0	2.91821	3.00000	3.000	.966355	.934	-.014	.057	-.517	.114	1.000	5.000
HOS_2	1834	0	1.86641	2.00000	2.000	.870900	.758	.936	.057	.678	.114	1.000	5.000
HOS_3	1834	0	1.72901	2.00000	1.000	.797677	.636	1.106	.057	1.336	.114	1.000	5.000
HOS_4	1834	0	2.18757	2.00000	2.000	.876895	.769	.778	.057	.670	.114	1.000	5.000
HOS_5	1834	0	1.83206	2.00000	2.000	.766979	.588	.956	.057	1.438	.114	1.000	5.000
HOS_6	1834	0	3.44493	4.00000	4.000	.995609	.991	-.595	.057	-.136	.114	1.000	5.000
HOS_7	1834	0	2.19738	2.00000	2.000	1.101640	1.214	.609	.057	-.607	.114	1.000	5.000
HOS_8	1834	0	2.01527	2.00000	2.000	.946344	.896	.770	.057	.087	.114	1.000	5.000
HOS_9	1834	0	2.14340	2.00000	2.000	.878079	.771	.806	.057	.680	.114	1.000	5.000
HOS_10	1834	0	1.92530	2.00000	2.000	.895758	.802	1.050	.057	1.114	.114	1.000	5.000
HOS_11	1834	0	1.95692	2.00000	2.000	.845900	.716	.910	.057	1.098	.114	1.000	5.000
HOS_12	1834	0	1.85878	2.00000	2.000	.813291	.661	1.031	.057	1.525	.114	1.000	5.000
HOS_13	1834	0	2.84569	3.00000	3.000	.967370	.936	.023	.057	-.590	.114	1.000	5.000

HOS_14	1834	0	2.39040	2.00000	2.000	1.004211	1.008	.470	.057	-.470	.114	1.000	5.000
INT_1	1834	0	4.28244	4.00000	4.000	.692763	.480	-.924	.057	1.655	.114	1.000	5.000
INT_2	1834	0	4.08397	4.00000	4.000	.744160	.554	-.725	.057	.966	.114	1.000	5.000
INT_3	1834	0	3.99727	4.00000	4.000	.773605	.598	-.639	.057	.661	.114	1.000	5.000
INT_4	1834	0	4.03980	4.00000	4.000	.716912	.514	-.583	.057	.780	.114	1.000	5.000
INT_5	1834	0	3.45583	4.00000	4.000	1.061893	1.128	-.473	.057	-.349	.114	1.000	5.000
INT_6	1834	0	4.13304	4.00000	4.000	.724272	.525	-.854	.057	1.664	.114	1.000	5.000
INT_7	1834	0	3.78026	4.00000	4.000	.681149	.464	-.461	.057	.956	.114	1.000	5.000
INT_8	1834	0	3.88113	4.00000	4.000	.737909	.545	-.673	.057	.997	.114	1.000	5.000
INT_9	1834	0	3.80862	4.00000	4.000	.743704	.553	-.513	.057	.724	.114	1.000	5.000
INT_10	1834	0	4.19520	4.00000	4.000	.604636	.366	-.563	.057	1.984	.114	1.000	5.000
INT_11	1834	0	3.96728	4.00000	4.000	.751807	.565	-.648	.057	.931	.114	1.000	5.000
INTEG_1	1834	0	4.03053	4.00000	4.000	.766273	.587	-.540	.057	.288	.114	1.000	5.000
INTEG_2	1834	0	4.09815	4.00000	4.000	.696156	.485	-.484	.057	.491	.114	1.000	5.000
INTEG_3	1834	0	4.27699	4.00000	4.000	.670185	.449	-.848	.057	1.715	.114	1.000	5.000
INTEG_4	1834	0	3.90076	4.00000	4.000	.883238	.780	-.690	.057	.263	.114	1.000	5.000
INTEG_5	1834	0	4.14013	4.00000	4.000	.630636	.398	-.576	.057	1.558	.114	1.000	5.000
INTEG_6	1834	0	4.16794	4.00000	4.000	.652443	.426	-.503	.057	.748	.114	1.000	5.000
INTEG_7	1834	0	4.08779	4.00000	4.000	.656856	.431	-.695	.057	1.814	.114	1.000	5.000
INTEG_8	1834	0	4.15649	4.00000	4.000	.748188	.560	-.826	.057	1.180	.114	1.000	5.000
INTEG_9	1834	0	4.17830	4.00000	4.000	.645897	.417	-.686	.057	1.727	.114	1.000	5.000
INTEG_10	1834	0	4.07961	4.00000	4.000	.757713	.574	-.857	.057	1.407	.114	1.000	5.000
INTEG_11	1834	0	4.13359	4.00000	4.000	.673021	.453	-.498	.057	.576	.114	1.000	5.000
INTEG_12	1834	0	4.27590	4.00000	4.000	.602955	.364	-.538	.057	1.308	.114	1.000	5.000
INTEG_13	1834	0	4.04744	4.00000	4.000	.695584	.484	-.483	.057	.669	.114	1.000	5.000
INTER_1	1834	0	3.97655	4.00000	4.000	.843444	.711	-1.059	.057	1.788	.114	1.000	5.000
INTER_2	1834	0	3.87296	4.00000	4.000	.886154	.785	-.818	.057	.740	.114	1.000	5.000
INTER_3	1834	0	4.27699	4.00000	4.000	.616773	.380	-.732	.057	2.224	.114	1.000	5.000
INTER_4	1834	0	4.10414	4.00000	4.000	.658616	.434	-.607	.057	1.450	.114	1.000	5.000

INTER_5	1834	0	3.83751	4.00000	4.000	.771696	.596	-.761	.057	1.122	.114	1.000	5.000
INTER_6	1834	0	3.96020	4.00000	4.000	.673760	.454	-.649	.057	1.556	.114	1.000	5.000
INTER_7	1834	0	3.64395	4.00000	4.000	.802512	.644	-.525	.057	.312	.114	1.000	5.000
INTER_8	1834	0	3.63522	4.00000	4.000	.840519	.706	-.615	.057	.325	.114	1.000	5.000
INTER_9	1834	0	4.01254	4.00000	4.000	.641666	.412	-.731	.057	2.045	.114	1.000	5.000
NEG_1	1834	0	2.24155	2.00000	2.000	1.093051	1.195	.685	.057	-.239	.114	1.000	5.000
NEG_2	1834	0	3.10305	3.00000	4.000	1.166808	1.361	-.096	.057	-.982	.114	1.000	5.000
NEG_3	1834	0	3.33751	3.00000	4.000	1.107451	1.226	-.195	.057	-.811	.114	1.000	5.000
NEG_4	1834	0	3.30589	3.00000	4.000	1.057897	1.119	-.277	.057	-.676	.114	1.000	5.000
NEG_5	1834	0	2.09215	2.00000	2.000	.900211	.810	.775	.057	.575	.114	1.000	5.000
NEG_6	1834	0	2.41985	2.00000	2.000	1.083205	1.173	.467	.057	-.511	.114	1.000	5.000
NEG_7	1834	0	3.55071	4.00000	4.000	1.026592	1.054	-.384	.057	-.499	.114	1.000	5.000
NEG_8	1834	0	2.98582	3.00000	2.000	1.077881	1.162	.091	.057	-.814	.114	1.000	5.000
NEG_9	1834	0	2.93893	3.00000	3.000	.985758	.972	.023	.057	-.588	.114	1.000	5.000
NEG_10	1834	0	2.68157	3.00000	2.000	1.096549	1.202	.284	.057	-.707	.114	1.000	5.000
ORD_1	1834	0	3.95529	4.00000	4.000	.876540	.768	-.633	.057	.117	.114	1.000	5.000
ORD_2	1834	0	4.06870	4.00000	4.000	.711278	.506	-.591	.057	.812	.114	1.000	5.000
ORD_3	1834	0	3.83969	4.00000	4.000	.841786	.709	-.679	.057	.445	.114	1.000	5.000
ORD_4	1834	0	3.92421	4.00000	4.000	.797728	.636	-.502	.057	.270	.114	1.000	5.000
ORD_5	1834	0	3.71156	4.00000	4.000	.898691	.808	-.507	.057	.012	.114	1.000	5.000
ORD_6	1834	0	3.84024	4.00000	4.000	.860160	.740	-.645	.057	.373	.114	1.000	5.000
ORD_7	1834	0	3.82552	4.00000	4.000	.819220	.671	-.515	.057	.293	.114	1.000	5.000
ORD_8	1834	0	4.00709	4.00000	4.000	.668400	.447	-.469	.057	.971	.114	1.000	5.000
ORD_9	1834	0	3.93675	4.00000	4.000	.750582	.563	-.570	.057	.560	.114	1.000	5.000
ORD_10	1834	0	3.88386	4.00000	4.000	.827205	.684	-.661	.057	.580	.114	1.000	5.000
ORD_11	1834	0	3.91385	4.00000	4.000	.721575	.521	-.628	.057	1.010	.114	1.000	5.000
ORD_12	1834	0	3.84188	4.00000	4.000	.780331	.609	-.585	.057	.643	.114	1.000	5.000
ORD_13	1834	0	3.85660	4.00000	4.000	.789762	.624	-.565	.057	.646	.114	1.000	5.000
PLA_1	1834	0	3.91603	4.00000	4.000	.905541	.820	-.642	.057	.117	.114	1.000	5.000

PLA_2	1834	0	3.75191	4.00000	4.000	.967592	.936	-.779	.057	.331	.114	1.000	5.000
PLA_3	1834	0	3.60796	4.00000	4.000	.996755	.994	-.511	.057	-.339	.114	1.000	5.000
PLA_4	1834	0	3.83043	4.00000	4.000	.817252	.668	-.622	.057	.684	.114	.000	5.000
PLA_5	1834	0	3.92857	4.00000	4.000	.734781	.540	-.730	.057	1.323	.114	1.000	5.000
PLA_6	1834	0	3.30807	3.00000	4.000	1.003252	1.007	-.211	.057	-.520	.114	1.000	5.000
SOC_1	1834	0	3.07906	3.00000	3.000	1.083411	1.174	.072	.057	-.735	.114	1.000	5.000
SOC_2	1834	0	3.70229	4.00000	4.000	.920000	.846	-.665	.057	.303	.114	1.000	5.000
SOC_3	1834	0	3.45474	4.00000	4.000	1.058760	1.121	-.350	.057	-.615	.114	1.000	5.000
SOC_4	1834	0	3.10632	3.00000	4.000	1.121205	1.257	.011	.057	-.882	.114	1.000	5.000
SOC_5	1834	0	3.47437	4.00000	4.000	1.053738	1.110	-.413	.057	-.565	.114	1.000	5.000
SOC_6	1834	0	4.04635	4.00000	4.000	.765835	.587	-.815	.057	1.045	.114	1.000	5.000
SOC_7	1834	0	3.65485	4.00000	4.000	.947193	.897	-.563	.057	-.106	.114	1.000	5.000
SOCIN_1	1834	0	3.81516	4.00000	4.000	.755536	.571	-.493	.057	.611	.114	1.000	5.000
SOCIN_2	1834	0	3.89531	4.00000	4.000	.768977	.591	-.698	.057	1.048	.114	1.000	5.000
SOCIN_3	1834	0	3.92421	4.00000	4.000	.733601	.538	-.794	.057	1.570	.114	1.000	5.000
SOCIN_4	1834	0	3.82170	4.00000	4.000	.762135	.581	-.687	.057	.935	.114	1.000	5.000
TRA_1	1834	0	3.32170	3.00000	4.000	1.036749	1.075	-.308	.057	-.425	.114	1.000	5.000
TRA_2	1834	0	3.29444	4.00000	4.000	1.296477	1.681	-.440	.057	-.912	.114	1.000	5.000
TRA_3	1834	0	3.92803	4.00000	4.000	.884028	.782	-.827	.057	.709	.114	1.000	5.000
TRA_4	1834	0	3.31189	4.00000	4.000	1.336793	1.787	-.415	.057	-.998	.114	1.000	5.000
WAR_1	1834	0	4.18757	4.00000	4.000	.718957	.517	-.728	.057	.915	.114	1.000	5.000
WAR_2	1834	0	4.05780	4.00000	4.000	.747381	.559	-.824	.057	1.526	.114	1.000	5.000
WAR_3	1834	0	3.97983	4.00000	4.000	.677055	.458	-.504	.057	.871	.114	1.000	5.000
WAR_4	1834	0	4.24209	4.00000	4.000	.605715	.367	-.512	.057	1.415	.114	1.000	5.000
WAR_5	1834	0	3.86041	4.00000	4.000	.724547	.525	-.592	.057	1.060	.114	1.000	5.000
WAR_6	1834	0	3.81134	4.00000	4.000	.769978	.593	-.431	.057	.149	.114	1.000	5.000
WAR_7	1834	0	3.96565	4.00000	4.000	.678899	.461	-.743	.057	1.701	.114	1.000	5.000
WAR_8	1834	0	3.83860	4.00000	4.000	.720000	.518	-.644	.057	.979	.114	1.000	5.000
WAR_9	1834	0	3.93457	4.00000	4.000	.664815	.442	-.651	.057	1.359	.114	1.000	5.000

WAR_10	1834	0	3.77263	4.00000	4.000	.754758	.570	-.649	.057	.775	.114	1.000	5.000
WAR_11	1834	0	3.90513	4.00000	4.000	.648761	.421	-.506	.057	1.080	.114	1.000	5.000

FEMALE SAMPLE

	N		Mean	Median	Mode	Std. Deviation	Variance	Skewness	Std. Error of Skewness	Kurtosis	Std. Error of Kurtosis	Minimum	Maximum
	Valid	Missing											
ACH_1	2420	0	4.13843	4.00000	4.000	.835505	.698	-.937	.050	.969	.099	1.000	5.000
ACH_2	2420	0	3.97562	4.00000	4.000	.890342	.793	-.786	.050	.458	.099	1.000	5.000
ACH_3	2420	0	3.85909	4.00000	4.000	.944721	.892	-.867	.050	.657	.099	1.000	5.000
ACH_4	2420	0	4.18678	4.00000	4.000	.651762	.425	-.648	.050	1.474	.099	1.000	5.000
ACH_5	2420	0	4.05661	4.00000	4.000	.761095	.579	-.641	.050	.510	.099	1.000	5.000
ACH_6	2420	0	3.40207	3.00000	4.000	1.051739	1.106	-.245	.050	-.657	.099	1.000	5.000
ACH_7	2420	0	4.22603	4.00000	4.000	.646599	.418	-.512	.050	.565	.099	1.000	5.000
ACH_8	2420	0	4.40496	4.00000	4.000	.610349	.373	-.687	.050	.537	.099	1.000	5.000
ACH_9	2420	0	4.05124	4.00000	4.000	.732984	.537	-.622	.050	.704	.099	1.000	5.000
ACH_10	2420	0	4.21446	4.00000	4.000	.632486	.400	-.532	.050	1.021	.099	1.000	5.000
ACH_11	2420	0	4.12355	4.00000	4.000	.653814	.427	-.425	.050	.483	.099	1.000	5.000
ARR_1	2420	0	2.07893	2.00000	2.000	.908807	.826	.849	.050	.678	.099	1.000	5.000
ARR_2	2420	0	1.76983	2.00000	1.000	.837524	.701	1.130	.050	1.339	.099	1.000	5.000
ARR_3	2420	0	1.82107	2.00000	2.000	.847028	.717	1.227	.050	1.881	.099	1.000	5.000
ARR_4	2420	0	2.23471	2.00000	2.000	1.063251	1.131	.577	.050	-.541	.099	1.000	5.000
ARR_5	2420	0	1.74050	2.00000	1.000	.827651	.685	1.105	.050	1.055	.099	1.000	5.000
ARR_6	2420	0	1.89587	2.00000	2.000	.897325	.805	1.085	.050	1.167	.099	1.000	5.000
BRO_1	2420	0	3.92769	4.00000	4.000	.921998	.850	-.689	.050	.042	.099	1.000	5.000
BRO_2	2420	0	3.89132	4.00000	4.000	.920668	.848	-.696	.050	.142	.099	1.000	5.000
BRO_3	2420	0	4.19008	4.00000	4.000	.667116	.445	-.606	.050	.934	.099	1.000	5.000

BRO_4	2420	0	3.96198	4.00000	4.000	.822509	.677	-.683	.050	.374	.099	1.000	5.000
BRO_5	2420	0	3.61488	4.00000	4.000	.846915	.717	-.291	.050	-.205	.099	1.000	5.000
BRO_6	2420	0	4.42686	5.00000	5.000	.678633	.461	-1.175	.050	1.937	.099	1.000	5.000
CON_1	2420	0	2.55124	3.00000	2.000	.994122	.988	.219	.050	-.540	.099	1.000	5.000
CON_2	2420	0	1.59380	1.00000	1.000	.687004	.472	.889	.050	.212	.099	1.000	4.000
CON_3	2420	0	1.81198	2.00000	2.000	.791504	.626	.968	.050	1.115	.099	1.000	5.000
CON_4	2420	0	2.37562	2.00000	2.000	.968619	.938	.478	.050	-.252	.099	1.000	5.000
CON_5	2420	0	1.85083	2.00000	2.000	.856159	.733	.876	.050	.499	.099	1.000	5.000
CON_6	2420	0	1.83306	2.00000	2.000	.787604	.620	.908	.050	1.006	.099	1.000	5.000
CON_7	2420	0	2.50083	2.00000	2.000	.995857	.992	.187	.050	-.606	.099	1.000	5.000
DEC_1	2420	0	1.67107	2.00000	1.000	.805668	.649	1.319	.050	1.980	.099	1.000	5.000
DEC_2	2420	0	2.76074	3.00000	3.000	1.019948	1.040	.000	.050	-.657	.099	1.000	5.000
DEC_3	2420	0	2.57810	2.00000	2.000	1.134276	1.287	.346	.050	-.724	.099	1.000	5.000
DEC_4	2420	0	2.24421	2.00000	2.000	.974387	.949	.588	.050	-.178	.099	1.000	5.000
DEC_5	2420	0	1.75083	2.00000	2.000	.811439	.658	1.288	.050	2.229	.099	1.000	5.000
DEC_6	2420	0	1.74669	2.00000	1.000	.868284	.754	1.219	.050	1.409	.099	1.000	5.000
DEC_7	2420	0	2.37769	2.00000	2.000	.914898	.837	.335	.050	-.335	.099	1.000	5.000
EMO_1	2420	0	4.38760	5.00000	5.000	.734894	.540	-1.159	.050	1.449	.099	1.000	5.000
EMO_2	2420	0	3.70041	4.00000	4.000	.839250	.704	-.474	.050	.128	.099	1.000	5.000
EMO_3	2420	0	3.49339	4.00000	4.000	.974858	.950	-.409	.050	-.331	.099	1.000	5.000
EMO_4	2420	0	3.98802	4.00000	4.000	.725902	.527	-.728	.050	1.447	.099	1.000	5.000
EMO_5	2420	0	3.57149	4.00000	4.000	.898907	.808	-.483	.050	.177	.099	1.000	5.000
EMO_6	2420	0	3.97645	4.00000	4.000	.783157	.613	-.698	.050	.883	.099	1.000	5.000
EMO_7	2420	0	3.64256	4.00000	4.000	.881066	.776	-.443	.050	-.085	.099	1.000	5.000
EMO_8	2420	0	3.98430	4.00000	4.000	.848948	.721	-.753	.050	.599	.099	1.000	5.000
EMP_1	2420	0	4.16116	4.00000	4.000	.652894	.426	-.578	.050	1.143	.099	1.000	5.000
EMP_2	2420	0	4.16116	4.00000	4.000	.611029	.373	-.440	.050	1.230	.099	1.000	5.000
EMP_3	2420	0	4.15000	4.00000	4.000	.763077	.582	-.913	.050	1.309	.099	1.000	5.000
EMP_4	2420	0	4.01901	4.00000	4.000	.765936	.587	-.828	.050	1.241	.099	1.000	5.000

EMP_5	2420	0	4.13430	4.00000	4.000	.663011	.440	-.675	.050	1.497	.099	1.000	5.000
EMP_6	2420	0	4.32025	4.00000	4.000	.612940	.376	-.589	.050	.931	.099	1.000	5.000
EMP_7	2420	0	3.94380	4.00000	4.000	.739361	.547	-.881	.050	1.733	.099	1.000	5.000
EPI_1	2420	0	4.40744	5.00000	5.000	.695544	.484	-1.286	.050	2.678	.099	1.000	5.000
EPI_2	2420	0	4.26240	4.00000	4.000	.726962	.528	-.912	.050	1.155	.099	1.000	5.000
EPI_3	2420	0	4.15661	4.00000	4.000	.625900	.392	-.472	.050	.974	.099	1.000	5.000
EPI_4	2420	0	4.37273	4.00000	4.000	.590611	.349	-.504	.050	.307	.099	2.000	5.000
EPI_5	2420	0	4.29380	4.00000	4.000	.564232	.318	-.316	.050	1.033	.099	1.000	5.000
EPI_6	2420	0	4.54959	5.00000	5.000	.585426	.343	-1.127	.050	1.676	.099	1.000	5.000
FAC_1	2420	0	3.75289	4.00000	4.000	.785416	.617	-.401	.050	.116	.099	1.000	5.000
FAC_2	2420	0	3.71570	4.00000	4.000	.856945	.734	-.497	.050	.027	.099	1.000	5.000
FAC_3	2420	0	3.76157	4.00000	4.000	.752404	.566	-.317	.050	.116	.099	1.000	5.000
FAC_4	2420	0	3.76198	4.00000	4.000	.728815	.531	-.184	.050	-.134	.099	1.000	5.000
FAC_5	2420	0	3.74711	4.00000	4.000	.775618	.602	-.582	.050	.589	.099	1.000	5.000
FAC_6	2420	0	4.05744	4.00000	4.000	.654117	.428	-.511	.050	1.148	.099	1.000	5.000
FAC_7	2420	0	3.84380	4.00000	4.000	.738410	.545	-.484	.050	.511	.099	1.000	5.000
FAC_8	2420	0	3.56777	4.00000	4.000	.840256	.706	-.237	.050	-.166	.099	1.000	5.000
FAC_9	2420	0	3.96529	4.00000	4.000	.718585	.516	-.517	.050	.600	.099	1.000	5.000
FAC_10	2420	0	4.01240	4.00000	4.000	.703628	.495	-.566	.050	.744	.099	1.000	5.000
HOS_1	2420	0	2.80620	3.00000	3.000	.946285	.895	.031	.050	-.428	.099	1.000	5.000
HOS_2	2420	0	1.73967	2.00000	1.000	.827392	.685	1.142	.050	1.319	.099	1.000	5.000
HOS_3	2420	0	1.65579	2.00000	1.000	.760830	.579	1.069	.050	.979	.099	1.000	5.000
HOS_4	2420	0	2.00702	2.00000	2.000	.840500	.706	.844	.050	.868	.099	1.000	5.000
HOS_5	2420	0	1.67273	2.00000	2.000	.703385	.495	1.098	.050	2.151	.099	1.000	5.000
HOS_6	2420	0	3.35579	4.00000	4.000	.988706	.978	-.461	.050	-.343	.099	1.000	5.000
HOS_7	2420	0	2.04380	2.00000	1.000	1.025382	1.051	.757	.050	-.268	.099	1.000	5.000
HOS_8	2420	0	1.96570	2.00000	2.000	.909314	.827	.702	.050	-.048	.099	1.000	5.000
HOS_9	2420	0	1.94339	2.00000	2.000	.815630	.665	.992	.050	1.544	.099	1.000	5.000
HOS_10	2420	0	1.65372	2.00000	1.000	.725664	.527	1.128	.050	1.754	.099	1.000	5.000

HOS_11	2420	0	1.78306	2.00000	2.000	.768086	.590	1.043	.050	1.683	.099	1.000	5.000
HOS_12	2420	0	1.69917	2.00000	2.000	.740099	.548	1.188	.050	2.412	.099	1.000	5.000
HOS_13	2420	0	2.55661	2.00000	2.000	.953496	.909	.300	.050	-.399	.099	1.000	5.000
HOS_14	2420	0	2.10702	2.00000	2.000	.926680	.859	.691	.050	.045	.099	1.000	5.000
INT_1	2420	0	4.28017	4.00000	4.000	.684092	.468	-.895	.050	1.724	.099	1.000	5.000
INT_2	2420	0	3.83430	4.00000	4.000	.756549	.572	-.363	.050	.059	.099	1.000	5.000
INT_3	2420	0	3.85289	4.00000	4.000	.802958	.645	-.500	.050	.175	.099	1.000	5.000
INT_4	2420	0	3.95331	4.00000	4.000	.697756	.487	-.463	.050	.550	.099	1.000	5.000
INT_5	2420	0	3.48388	4.00000	4.000	1.038405	1.078	-.449	.050	-.355	.099	1.000	5.000
INT_6	2420	0	4.25207	4.00000	4.000	.663287	.440	-.816	.050	1.730	.099	1.000	5.000
INT_7	2420	0	3.73719	4.00000	4.000	.659103	.434	-.300	.050	.424	.099	1.000	5.000
INT_8	2420	0	3.87603	4.00000	4.000	.740605	.548	-.605	.050	.747	.099	1.000	5.000
INT_9	2420	0	3.66529	4.00000	4.000	.766950	.588	-.263	.050	-.086	.099	1.000	5.000
INT_10	2420	0	4.20248	4.00000	4.000	.616966	.381	-.601	.050	1.830	.099	1.000	5.000
INT_11	2420	0	3.92107	4.00000	4.000	.750892	.564	-.486	.050	.290	.099	1.000	5.000
INTEG_1	2420	0	4.03512	4.00000	4.000	.738146	.545	-.463	.050	.121	.099	1.000	5.000
INTEG_2	2420	0	4.14174	4.00000	4.000	.702086	.493	-.570	.050	.576	.099	1.000	5.000
INTEG_3	2420	0	4.40537	4.00000	5.000	.638871	.408	-.995	.050	1.968	.099	1.000	5.000
INTEG_4	2420	0	3.95868	4.00000	4.000	.786373	.618	-.570	.050	.280	.099	1.000	5.000
INTEG_5	2420	0	4.19669	4.00000	4.000	.613468	.376	-.336	.050	.369	.099	2.000	5.000
INTEG_6	2420	0	4.23554	4.00000	4.000	.597550	.357	-.314	.050	.519	.099	1.000	5.000
INTEG_7	2420	0	4.14752	4.00000	4.000	.648187	.420	-.681	.050	1.729	.099	1.000	5.000
INTEG_8	2420	0	4.21033	4.00000	4.000	.723989	.524	-.904	.050	1.708	.099	1.000	5.000
INTEG_9	2420	0	4.29215	4.00000	4.000	.628801	.395	-.688	.050	1.510	.099	1.000	5.000
INTEG_10	2420	0	4.05083	4.00000	4.000	.715605	.512	-.657	.050	1.017	.099	1.000	5.000
INTEG_11	2420	0	4.27893	4.00000	4.000	.622283	.387	-.421	.050	.144	.099	1.000	5.000
INTEG_12	2420	0	4.26281	4.00000	4.000	.585357	.343	-.300	.050	.476	.099	1.000	5.000
INTEG_13	2420	0	4.07355	4.00000	4.000	.668461	.447	-.393	.050	.538	.099	1.000	5.000
INTER_1	2420	0	4.05537	4.00000	4.000	.868011	.753	-1.196	.050	2.074	.099	1.000	5.000

INTER_2	2420	0	3.89504	4.00000	4.000	.823221	.678	-.587	.050	.376	.099	1.000	5.000
INTER_3	2420	0	4.28223	4.00000	4.000	.562812	.317	-.409	.050	1.888	.099	1.000	5.000
INTER_4	2420	0	4.18388	4.00000	4.000	.602863	.363	-.331	.050	.675	.099	1.000	5.000
INTER_5	2420	0	3.87273	4.00000	4.000	.781345	.611	-.778	.050	1.123	.099	1.000	5.000
INTER_6	2420	0	4.05950	4.00000	4.000	.651716	.425	-.661	.050	1.801	.099	1.000	5.000
INTER_7	2420	0	3.67893	4.00000	4.000	.785231	.617	-.521	.050	.316	.099	1.000	5.000
INTER_8	2420	0	3.68140	4.00000	4.000	.801859	.643	-.422	.050	.090	.099	1.000	5.000
INTER_9	2420	0	4.02769	4.00000	4.000	.612379	.375	-.620	.050	2.109	.099	1.000	5.000
NEG_1	2420	0	3.19215	3.00000	2.000	1.206920	1.457	-.030	.050	-1.024	.099	1.000	5.000
NEG_2	2420	0	3.29711	3.00000	4.000	1.159354	1.344	-.235	.050	-.853	.099	1.000	5.000
NEG_3	2420	0	3.51033	4.00000	4.000	1.091275	1.191	-.317	.050	-.745	.099	1.000	5.000
NEG_4	2420	0	3.38678	4.00000	4.000	1.092266	1.193	-.342	.050	-.659	.099	1.000	5.000
NEG_5	2420	0	2.16818	2.00000	2.000	.880300	.775	.631	.050	.325	.099	1.000	5.000
NEG_6	2420	0	2.39793	2.00000	2.000	1.071598	1.148	.420	.050	-.573	.099	1.000	5.000
NEG_7	2420	0	3.93967	4.00000	4.000	.970246	.941	-.817	.050	.224	.099	1.000	5.000
NEG_8	2420	0	3.27851	3.00000	4.000	1.086632	1.181	-.109	.050	-.860	.099	1.000	5.000
NEG_9	2420	0	2.82190	3.00000	3.000	.964487	.930	.118	.050	-.445	.099	1.000	5.000
NEG_10	2420	0	2.67190	3.00000	2.000	1.069222	1.143	.342	.050	-.604	.099	1.000	5.000
ORD_1	2420	0	4.09669	4.00000	4.000	.841605	.708	-.771	.050	.341	.099	1.000	5.000
ORD_2	2420	0	4.16653	4.00000	4.000	.692453	.479	-.645	.050	.927	.099	1.000	5.000
ORD_3	2420	0	3.83802	4.00000	4.000	.847754	.719	-.582	.050	.163	.099	1.000	5.000
ORD_4	2420	0	4.05702	4.00000	4.000	.755887	.571	-.606	.050	.448	.099	1.000	5.000
ORD_5	2420	0	3.94091	4.00000	4.000	.877006	.769	-.684	.050	.271	.099	1.000	5.000
ORD_6	2420	0	4.03388	4.00000	4.000	.835652	.698	-.834	.050	.773	.099	1.000	5.000
ORD_7	2420	0	3.90248	4.00000	4.000	.846896	.717	-.610	.050	.168	.099	1.000	5.000
ORD_8	2420	0	4.05909	4.00000	4.000	.665251	.443	-.421	.050	.549	.099	1.000	5.000
ORD_9	2420	0	4.04628	4.00000	4.000	.713890	.510	-.620	.050	.850	.099	1.000	5.000
ORD_10	2420	0	4.06860	4.00000	4.000	.793372	.629	-.804	.050	.880	.099	1.000	5.000
ORD_11	2420	0	4.04835	4.00000	4.000	.672296	.452	-.498	.050	.847	.099	1.000	5.000

ORD_12	2420	0	3.97686	4.00000	4.000	.767978	.590	-.662	.050	.728	.099	1.000	5.000
ORD_13	2420	0	3.98388	4.00000	4.000	.765192	.586	-.643	.050	.814	.099	1.000	5.000
PLA_1	2420	0	4.11488	4.00000	4.000	.850568	.723	-.794	.050	.266	.099	1.000	5.000
PLA_2	2420	0	3.63719	4.00000	4.000	.980473	.961	-.612	.050	-.118	.099	1.000	5.000
PLA_3	2420	0	3.49504	4.00000	4.000	1.003701	1.007	-.409	.050	-.443	.099	1.000	5.000
PLA_4	2420	0	3.81529	4.00000	4.000	.814410	.663	-.596	.050	.422	.099	1.000	5.000
PLA_5	2420	0	4.05950	4.00000	4.000	.678444	.460	-.677	.050	1.448	.099	1.000	5.000
PLA_6	2420	0	3.22893	3.00000	4.000	1.005396	1.011	-.106	.050	-.674	.099	1.000	5.000
SOC_1	2420	0	3.36570	3.00000	3.000	1.068809	1.142	-.127	.050	-.757	.099	1.000	5.000
SOC_2	2420	0	3.83347	4.00000	4.000	.850525	.723	-.624	.050	.327	.099	1.000	5.000
SOC_3	2420	0	3.66942	4.00000	4.000	.970320	.942	-.485	.050	-.305	.099	1.000	5.000
SOC_4	2420	0	3.12190	3.00000	4.000	1.118272	1.251	.011	.050	-.946	.099	1.000	5.000
SOC_5	2420	0	3.57975	4.00000	4.000	1.037058	1.075	-.534	.050	-.348	.099	1.000	5.000
SOC_6	2420	0	4.11777	4.00000	4.000	.715804	.512	-.773	.050	1.322	.099	1.000	5.000
SOC_7	2420	0	3.80124	4.00000	4.000	.890768	.793	-.696	.050	.289	.099	1.000	5.000
SOCIN_1	2420	0	3.93058	4.00000	4.000	.699122	.489	-.485	.050	.730	.099	1.000	5.000
SOCIN_2	2420	0	4.08760	4.00000	4.000	.640841	.411	-.542	.050	1.289	.099	1.000	5.000
SOCIN_3	2420	0	4.05207	4.00000	4.000	.659287	.435	-.809	.050	2.299	.099	1.000	5.000
SOCIN_4	2420	0	3.94711	4.00000	4.000	.691659	.478	-.785	.050	1.870	.099	1.000	5.000
TRA_1	2420	0	3.39132	3.00000	3.000	.961950	.925	-.322	.050	-.155	.099	1.000	5.000
TRA_2	2420	0	3.85661	4.00000	4.000	1.147862	1.318	-1.044	.050	.389	.099	1.000	5.000
TRA_3	2420	0	4.04711	4.00000	4.000	.772253	.596	-.787	.050	1.133	.099	1.000	5.000
TRA_4	2420	0	3.73636	4.00000	4.000	1.205918	1.454	-.858	.050	-.137	.099	1.000	5.000
WAR_1	2420	0	4.30248	4.00000	4.000	.661523	.438	-.688	.050	.687	.099	1.000	5.000
WAR_2	2420	0	4.27025	4.00000	4.000	.652929	.426	-.582	.050	.505	.099	1.000	5.000
WAR_3	2420	0	4.12603	4.00000	4.000	.624218	.390	-.514	.050	1.645	.099	1.000	5.000
WAR_4	2420	0	4.32562	4.00000	4.000	.579165	.335	-.418	.050	.930	.099	1.000	5.000
WAR_5	2420	0	3.93554	4.00000	4.000	.685871	.470	-.440	.050	.635	.099	1.000	5.000
WAR_6	2420	0	3.96488	4.00000	4.000	.721722	.521	-.621	.050	1.063	.099	1.000	5.000

WAR_7	2420	0	4.07686	4.00000	4.000	.630526	.398	-.467	.050	1.047	.099	1.000	5.000
WAR_8	2420	0	3.84504	4.00000	4.000	.721400	.520	-.525	.050	.630	.099	1.000	5.000
WAR_9	2420	0	4.11364	4.00000	4.000	.582134	.339	-.331	.050	1.438	.099	1.000	5.000
WAR_10	2420	0	3.98512	4.00000	4.000	.685727	.470	-.605	.050	1.211	.099	1.000	5.000
WAR_11	2420	0	4.03264	4.00000	4.000	.616844	.380	-.506	.050	1.652	.099	1.000	5.000

APPENDIX B: GOODNESS OF FIT STATISTICS FOR THE SAPI SINGLE GROUP MEASUREMENT MODEL: FEMALE

Goodness of Fit Statistics

Degrees of Freedom = 14005
Minimum Fit Function Chi-Square = 65565.28 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 90548.77 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 76543.77
90 Percent Confidence Interval for NCP = (75592.85 ; 77498.16)
Minimum Fit Function Value = 27.10
Population Discrepancy Function Value (F0) = 31.64
90 Percent Confidence Interval for F0 = (31.25 ; 32.04)
Root Mean Square Error of Approximation (RMSEA) = 0.048
90 Percent Confidence Interval for RMSEA = (0.047 ; 0.048)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.00
Expected Cross-Validation Index (ECVI) = 37.87
90 Percent Confidence Interval for ECVI = (37.48 ; 38.27)
ECVI for Saturated Model = 12.02
ECVI for Independence Model = 538.44
Chi-Square for Independence Model with 14365 Degrees of Freedom = 1302151.77
Independence AIC = 1302491.77
Model AIC = 91608.77
Saturated AIC = 29070.00
Independence CAIC = 1303646.33
Model CAIC = 95208.28
Saturated CAIC = 127784.78
Normed Fit Index (NFI) = 0.95
Non-Normed Fit Index (NNFI) = 0.96
Parsimony Normed Fit Index (PNFI) = 0.93
Comparative Fit Index (CFI) = 0.96
Incremental Fit Index (IFI) = 0.96
Relative Fit Index (RFI) = 0.95
Critical N (CN) = 532.18
Root Mean Square Residual (RMR) = 0.040
Standardized RMR = 0.059
Goodness of Fit Index (GFI) = 0.69
Adjusted Goodness of Fit Index (AGFI) = 0.68
Parsimony Goodness of Fit Index (PGFI) = 0.67

APPENDIX C: GOODNESS OF FIT STATISTICS FOR THE SAPI SINGLE GROUP MEASUREMENT MODEL: MALE

Degrees of Freedom = 14005
Minimum Fit Function Chi-Square = 55228.101 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 76246.177 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 62241.177
90 Percent Confidence Interval for NCP = (61376.166 ; 63110.409)
Minimum Fit Function Value = 30.130
Population Discrepancy Function Value (F0) = 33.956
90 Percent Confidence Interval for F0 = (33.484 ; 34.430)
Root Mean Square Error of Approximation (RMSEA) = 0.0492
90 Percent Confidence Interval for RMSEA = (0.0489 ; 0.0496)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000
Expected Cross-Validation Index (ECVI) = 42.175
90 Percent Confidence Interval for ECVI = (41.703 ; 42.649)
ECVI for Saturated Model = 15.859
ECVI for Independence Model = 616.460
Chi-Square for Independence Model with 14365 Degrees of Freedom = 1129631.890
Independence AIC = 1129971.890
Model AIC = 77306.177
Saturated AIC = 29070.000
Independence CAIC = 1131079.313
Model CAIC = 80758.732
Saturated CAIC = 123754.691
Normed Fit Index (NFI) = 0.951
Non-Normed Fit Index (NNFI) = 0.962
Parsimony Normed Fit Index (PNFI) = 0.927
Comparative Fit Index (CFI) = 0.963
Incremental Fit Index (IFI) = 0.963
Relative Fit Index (RFI) = 0.950
Critical N (CN) = 478.840
Root Mean Square Residual (RMR) = 0.0454
Standardized RMR = 0.0618
Goodness of Fit Index (GFI) = 0.671
Adjusted Goodness of Fit Index (AGFI) = 0.659
Parsimony Goodness of Fit Index (PGFI) = 0.647

APPENDIX D: GOODNESS OF FIT STATISTICS FOR THE SAPI CONFIGURAL INVARIANCE MEASUREMENT MODEL

Contribution to Chi-Square = 65565.281
Percentage Contribution to Chi-Square = 54.279
Root Mean Square Residual (RMR) = 0.0401
Standardized RMR = 0.0595
Goodness of Fit Index (GFI) = 0.694
Degrees of Freedom = 28010
Minimum Fit Function Chi-Square = 120793.382 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 166794.946 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 138784.946
90 Percent Confidence Interval for NCP = (0.0 ; 0.0)
Minimum Fit Function Value = 28.409
Population Discrepancy Function Value (F0) = 32.640
90 Percent Confidence Interval for F0 = (0.0 ; 0.0)
Root Mean Square Error of Approximation (RMSEA) = 0.0483
90 Percent Confidence Interval for RMSEA = (0.0 ; 0.0)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000
Expected Cross-Validation Index (ECVI) = 39.886
90 Percent Confidence Interval for ECVI = (7.166 ; 7.166)
ECVI for Saturated Model = 6.837
ECVI for Independence Model = 571.995
Chi-Square for Independence Model with 28730 Degrees of Freedom = 2431783.659
Independence AIC = 2432463.659
Model AIC = 110774.946
Saturated AIC = 58140.000
Independence CAIC = 2434964.568
Model BIC = -67232.658
Model CAIC = -95242.658
Saturated CAIC = 271967.728
Normed Fit Index (NFI) = 0.931
Non-Normed Fit Index (NNFI) = 0.941
Parsimony Normed Fit Index (PNFI) = 0.908
Comparative Fit Index (CFI) = 0.942
Incremental Fit Index (IFI) = 0.942
Relative Fit Index (RFI) = 0.930
Critical N (CN) = 729.153
Contribution to Chi-Square = 55228.101
Percentage Contribution to Chi-Square = 45.721
Root Mean Square Residual (RMR) = 0.0454
Standardized RMR = 0.0618
Goodness of Fit Index (GFI) = 0.671

APPENDIX E: GOODNESS OF FIT STATISTICS FOR THE SAPI WEAK INVARIANCE MEASUREMENT MODEL

Contribution to Chi-Square = 65716.991
Percentage Contribution to Chi-Square = 54.256
Root Mean Square Residual (RMR) = 0.0401
Standardized RMR = 0.0595
Goodness of Fit Index (GFI) = 0.694
Degrees of Freedom = 28160
Minimum Fit Function Chi-Square = 121124.733 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 167549.691 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 139389.691
90 Percent Confidence Interval for NCP = (0.0 ; 0.0)
Minimum Fit Function Value = 28.487
Population Discrepancy Function Value (F0) = 32.782
90 Percent Confidence Interval for F0 = (0.0 ; 0.0)
Root Mean Square Error of Approximation (RMSEA) = 0.0483
90 Percent Confidence Interval for RMSEA = (0.0 ; 0.0)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000
Expected Cross-Validation Index (ECVI) = 39.993
90 Percent Confidence Interval for ECVI = (7.131 ; 7.131)
ECVI for Saturated Model = 6.837
ECVI for Independence Model = 571.995
Chi-Square for Independence Model with 28730 Degrees of Freedom = 2431783.659
Independence AIC = 2432463.659
Model AIC = 111229.691
Saturated AIC = 58140.000
Independence CAIC = 2434964.568
Model BIC = -67731.185
Model CAIC = -95891.185
Saturated CAIC = 271967.728
Normed Fit Index (NFI) = 0.931
Non-Normed Fit Index (NNFI) = 0.941
Parsimony Normed Fit Index (PNFI) = 0.913
Comparative Fit Index (CFI) = 0.942
Incremental Fit Index (IFI) = 0.942
Relative Fit Index (RFI) = 0.930
Critical N (CN) = 729.717
Contribution to Chi-Square = 55407.742
Percentage Contribution to Chi-Square = 45.744
Root Mean Square Residual (RMR) = 0.0460
Standardized RMR = 0.0627
Goodness of Fit Index (GFI) = 0.670

APPENDIX F: GOODNESS OF FIT STATISTICS FOR THE SAPI STRONG INVARIANCE MEASUREMENT MODEL

Contribution to Chi-Square = 67286.796
Percentage Contribution to Chi-Square = 53.853
Root Mean Square Residual (RMR) = 0.0414
Standardized RMR = 0.0607
Goodness of Fit Index (GFI) = 0.686
Degrees of Freedom = 28330
Minimum Fit Function Chi-Square = 124944.721 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 175326.865 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 146996.865
90 Percent Confidence Interval for NCP = (0.0 ; 0.0)
Minimum Fit Function Value = 29.385
Population Discrepancy Function Value (F0) = 34.571
90 Percent Confidence Interval for F0 = (0.0 ; 0.0)
Root Mean Square Error of Approximation (RMSEA) = 0.0494
90 Percent Confidence Interval for RMSEA = (0.0 ; 0.0)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000
Expected Cross-Validation Index (ECVI) = 41.742
90 Percent Confidence Interval for ECVI = (7.091 ; 7.091)
ECVI for Saturated Model = 6.837
ECVI for Independence Model = 571.995
Chi-Square for Independence Model with 28730 Degrees of Freedom = 2431783.659
Independence AIC = 2432463.659
Model AIC = 118666.865
Saturated AIC = 58140.000
Independence CAIC = 2434964.568
Model BIC = -61374.386
Model CAIC = -89704.386
Saturated CAIC = 271967.728
Normed Fit Index (NFI) = 0.928
Non-Normed Fit Index (NNFI) = 0.938
Parsimony Normed Fit Index (PNFI) = 0.915
Comparative Fit Index (CFI) = 0.939
Incremental Fit Index (IFI) = 0.939
Relative Fit Index (RFI) = 0.927
Critical N (CN) = 701.556
Contribution to Chi-Square = 57657.925
Percentage Contribution to Chi-Square = 46.147
Root Mean Square Residual (RMR) = 0.0490
Standardized RMR = 0.0650
Goodness of Fit Index (GFI) = 0.654
Standardized RMR = 0.0650
Goodness of Fit Index (GFI) = 0.654

APPENDIX G: DIFFERENCE IN TAU BETWEEN MALE AND FEMALE SAMPLE GROUPS

H02				Ha TAU		
	ITEM NUMBER	**FEMALE**	**MALE**	**D**	**|D|**	**RANK-ORDERED D**
H021	NEG_1	3.192	2.242	0.95	0.95	1
H022	TRA_2	3.857	3.294	0.563	0.563	2
H023	TRA_4	3.736	3.312	0.424	0.424	3
H024	NEG_7	3.94	3.551	0.389	0.389	4
H025	CON_1	2.551	2.878	-0.327	0.327	5
H026	EMP_7	3.944	3.634	0.31	0.31	6
H027	CON_4	2.376	2.678	-0.302	0.302	7
H028	ARR_5	1.74	2.038	-0.298	0.298	8
H029	NEG_8	3.279	2.986	0.293	0.293	9
H0210	HOS_13	2.557	2.846	-0.289	0.289	10
H0211	SOC_1	3.366	3.079	0.287	0.287	11
H0212	HOS_14	2.107	2.39	-0.283	0.283	12
H0213	ACH_6	3.402	3.676	-0.274	0.274	13
H0214	HOS_10	1.654	1.925	-0.271	0.271	14
H0215	ARR_2	1.77	2.035	-0.265	0.265	15
H0216	EMP_3	4.15	3.9	0.25	0.25	16
H0217	INT_2	3.834	4.084	-0.25	0.25	17
H0218	ARR_1	2.079	2.326	-0.247	0.247	18
H0219	EMP_4	4.019	3.779	0.24	0.24	19
H0220	EMO_3	3.493	3.733	-0.24	0.24	19
H0221	EMO_2	3.7	3.93	-0.23	0.23	21
H0222	ORD_5	3.941	3.712	0.229	0.229	22
H0223	ARR_6	1.896	2.117	-0.221	0.221	23
H0224	SOC_3	3.669	3.455	0.214	0.214	24
H0225	WAR_10	3.985	3.773	0.212	0.212	25
H0226	WAR_2	4.27	4.058	0.212	0.212	25
H0227	EMP_5	4.134	3.926	0.208	0.208	27
H0228	DEC_6	1.747	1.955	-0.208	0.208	28
H0229	DEC_3	2.578	2.784	-0.206	0.206	29
H0230	ARR_3	1.821	2.023	-0.202	0.202	30
H0231	EMP_1	4.161	3.961	0.2	0.2	31
H0232	HOS_9	1.943	2.143	-0.2	0.2	31
H0233	PLA_1	4.115	3.916	0.199	0.199	33
H0234	ORD_6	4.034	3.84	0.194	0.194	34
H0235	NEG_2	3.297	3.103	0.194	0.194	34
H0236	SOCIN_2	4.088	3.895	0.193	0.193	36
H0237	DEC_5	1.751	1.941	-0.19	0.19	37
H0238	EMO_1	4.388	4.202	0.186	0.186	38
H0239	ORD_10	4.069	3.884	0.185	0.185	39
H0240	HOS_4	2.007	2.188	-0.181	0.181	40
H0241	EMO_5	3.571	3.752	-0.181	0.181	41
H0242	WAR_9	4.114	3.935	0.179	0.179	42
H0243	DEC_4	2.244	2.423	-0.179	0.179	42
H0244	ACH_1	4.138	3.959	0.179	0.179	42
H0245	HOS_11	1.783	1.957	-0.174	0.174	45
H0246	NEG_3	3.51	3.338	0.172	0.172	46

H0247	ARR_4	2.235	2.405	-0.17	0.17	47
H0248	ACH_8	4.405	4.238	0.167	0.167	48
H0249	EMP_6	4.32	4.155	0.165	0.165	49
H0250	HOS_12	1.699	1.859	-0.16	0.16	50
H0251	HOS_5	1.673	1.832	-0.159	0.159	51
H0252	WAR_6	3.965	3.811	0.154	0.154	52
H0253	HOS_7	2.044	2.197	-0.153	0.153	53
H0254	CON_6	1.833	1.984	-0.151	0.151	54
H0255	WAR_3	4.126	3.98	0.146	0.146	55
H0256	SOC_7	3.801	3.655	0.146	0.146	55
H0257	INTEG_11	4.279	4.134	0.145	0.145	57
H0258	INT_9	3.665	3.809	-0.144	0.144	58
H0259	INT_3	3.853	3.997	-0.144	0.144	59
H0260	BRO_5	3.615	3.759	-0.144	0.144	59
H0261	ORD_1	4.097	3.955	0.142	0.142	61
H0262	ORD_12	3.977	3.842	0.135	0.135	62
H0263	ORD_11	4.048	3.914	0.134	0.134	63
H0264	ORD_4	4.057	3.924	0.133	0.133	64
H0265	SOC_2	3.833	3.702	0.131	0.131	65
H0266	PLA_5	4.06	3.929	0.131	0.131	66
H0267	WAR_11	4.033	3.905	0.128	0.128	67
H0268	INTEG_3	4.405	4.277	0.128	0.128	68
H0269	SOCIN_3	4.052	3.924	0.128	0.128	69
H0270	DEC_2	2.761	2.889	-0.128	0.128	69
H0271	ORD_13	3.984	3.857	0.127	0.127	71
H0272	HOS_2	1.74	1.866	-0.126	0.126	72
H0273	EMP_2	4.161	4.035	0.126	0.126	73
H0274	SOCIN_4	3.947	3.822	0.125	0.125	74
H0275	FAC_3	3.762	3.637	0.125	0.125	74
H0276	DEC_1	1.671	1.793	-0.122	0.122	76
H0277	ACH_3	3.859	3.739	0.12	0.12	77
H0278	TRA_3	4.047	3.928	0.119	0.119	78
H0279	INT_6	4.252	4.133	0.119	0.119	78
H0280	BRO_6	4.427	4.309	0.118	0.118	80
H0281	NEG_9	2.822	2.939	-0.117	0.117	81
H0282	SOCIN_1	3.931	3.815	0.116	0.116	82
H0283	PLA_2	3.637	3.752	-0.115	0.115	83
H0284	INTEG_9	4.292	4.178	0.114	0.114	84
H0285	WAR_1	4.302	4.188	0.114	0.114	85

APPENDIX H: GOODNESS OF FIT STATISTICS FOR THE SAPI PARTIAL STRONG INVARIANCE MEASUREMENT MODEL

Contribution to Chi-Square = 65943.322
Percentage Contribution to Chi-Square = 54.204
Root Mean Square Residual (RMR) = 0.0401
Standardized RMR = 0.0595
Goodness of Fit Index (GFI) = 0.693
Degrees of Freedom = 28245
Minimum Fit Function Chi-Square = 121657.102 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 168032.538 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 139787.538
90 Percent Confidence Interval for NCP = (0.0 ; 0.0)
Minimum Fit Function Value = 28.612
Population Discrepancy Function Value (F0) = 32.876
90 Percent Confidence Interval for F0 = (0.0 ; 0.0)
Root Mean Square Error of Approximation (RMSEA) = 0.0482
90 Percent Confidence Interval for RMSEA = (0.0 ; 0.0)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000
Expected Cross-Validation Index (ECVI) = 40.066
90 Percent Confidence Interval for ECVI = (7.111 ; 7.111)
ECVI for Saturated Model = 6.837
ECVI for Independence Model = 571.995
Chi-Square for Independence Model with 28730 Degrees of Freedom = 2431783.659
Independence AIC = 2432463.659
Model AIC = 111542.538
Saturated AIC = 58140.000
Independence CAIC = 2434964.568
Model BIC = -67958.525
Model CAIC = -96203.525
Saturated CAIC = 271967.728
Normed Fit Index (NFI) = 0.931
Non-Normed Fit Index (NNFI) = 0.941
Parsimony Normed Fit Index (PNFI) = 0.915
Comparative Fit Index (CFI) = 0.942
Incremental Fit Index (IFI) = 0.942
Relative Fit Index (RFI) = 0.930
Critical N (CN) = 729.795
Contribution to Chi-Square = 55713.780
Percentage Contribution to Chi-Square = 45.796
Root Mean Square Residual (RMR) = 0.0461
Standardized RMR = 0.0628
Goodness of Fit Index (GFI) = 0.669

APPENDIX I: GOODNESS OF FIT STATISTICS FOR THE SAPI STRICT INVARIANCE MEASUREMENT MODEL

Contribution to Chi-Square = 66446.494
Percentage Contribution to Chi-Square = 54.139
Root Mean Square Residual (RMR) = 0.0402
Standardized RMR = 0.0589
Goodness of Fit Index (GFI) = 0.691
Degrees of Freedom = 28415
Minimum Fit Function Chi-Square = 122733.147 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 169832.984 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 141417.984
90 Percent Confidence Interval for NCP = (0.0 ; 0.0)
Minimum Fit Function Value = 28.865
Population Discrepancy Function Value (F0) = 33.259
90 Percent Confidence Interval for F0 = (0.0 ; 0.0)
Root Mean Square Error of Approximation (RMSEA) = 0.0484
90 Percent Confidence Interval for RMSEA = (0.0 ; 0.0)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000
Expected Cross-Validation Index (ECVI) = 40.410
90 Percent Confidence Interval for ECVI = (7.071 ; 7.071)
ECVI for Saturated Model = 6.837
ECVI for Independence Model = 571.995
Chi-Square for Independence Model with 28730 Degrees of Freedom = 2431783.659
Independence AIC = 2432463.659
Model AIC = 113002.984
Saturated AIC = 58140.000
Independence CAIC = 2434964.568
Model BIC = -67578.454
Model CAIC = -95993.454
Saturated CAIC = 271967.728
Normed Fit Index (NFI) = 0.930
Non-Normed Fit Index (NNFI) = 0.940
Parsimony Normed Fit Index (PNFI) = 0.920
Comparative Fit Index (CFI) = 0.941
Incremental Fit Index (IFI) = 0.941
Relative Fit Index (RFI) = 0.929
Critical N (CN) = 726.367
Contribution to Chi-Square = 56286.653
Percentage Contribution to Chi-Square = 45.861
Root Mean Square Residual (RMR) = 0.0462
Standardized RMR = 0.0639
Goodness of Fit Index (GFI) = 0.668

APPENDIX J: DIFFERENCE IN THETA-DELTA BETWEEN MALE AND FEMALE SAMPLE GROUPS

H03_Add to previous H02	Item	Female	Male	D	|D|	RANK-ORDERED D
H031	EPI_6	0.649	0.879	(0.23)	0.23	1
H032	HOS_10	0.441	0.664	(0.22)	0.223	2
H033	TRA_3	0.679	0.9	(0.22)	0.221	3
H034	DEC_6	0.666	0.865	(0.20)	0.199	4
H035	SOCIN_2	0.521	0.708	(0.19)	0.187	5
H036	INTEG_11	0.594	0.774	(0.18)	0.18	6
H037	BRO_6	0.656	0.835	(0.18)	0.179	7
H038	ACH_6	0.867	0.694	0.17	0.173	8
H039	WAR_2	0.63	0.79	(0.16)	0.16	9
H0310	CON_2	0.54	0.695	(0.16)	0.155	10
H0311	ACH_8	0.533	0.687	(0.15)	0.154	11
H0312	EMP_7	0.582	0.729	(0.15)	0.147	12
H0313	ARR_6	0.542	0.674	(0.13)	0.132	13
H0314	HOS_12	0.637	0.767	(0.13)	0.13	14
H0315	EMP_3	0.688	0.817	(0.13)	0.129	15
H0316	INT_10	0.814	0.69	0.12	0.124	16
H0317	INT_11	0.775	0.654	0.12	0.121	17
H0318	INTEG_9	0.672	0.792	(0.12)	0.12	18
H0319	ACH_10	0.496	0.612	(0.12)	0.116	19
H0320	BRO_4	0.621	0.737	(0.12)	0.116	19
H0321	EMP_6	0.7	0.815	(0.12)	0.115	21
H0322	HOS_11	0.602	0.716	(0.11)	0.114	22
H0323	NEG_3	0.555	0.666	(0.11)	0.111	23
H0324	INTER_2	0.777	0.884	(0.11)	0.107	24
H0325	WAR_9	0.435	0.541	(0.11)	0.106	25
H0326	INTEG_3	0.684	0.79	(0.11)	0.106	26
H0327	INTEG_4	0.72	0.824	(0.10)	0.104	27
H0328	INTER_1	0.911	0.807	0.10	0.104	27
H0329	NEG_7	0.812	0.916	(0.10)	0.104	27
H0330	HOS_7	0.755	0.858	(0.10)	0.103	30
H0331	HOS_5	0.587	0.688	(0.10)	0.101	31
H0332	WAR_10	0.589	0.69	(0.10)	0.101	31
H0333	WAR_1	0.645	0.744	(0.10)	0.099	33
H0334	DEC_1	0.592	0.686	(0.09)	0.094	34
H0335	EMP_2	0.576	0.669	(0.09)	0.093	35
H0336	SOCIN_1	0.586	0.679	(0.09)	0.093	35
H0337	INTER_3	0.704	0.794	(0.09)	0.09	37
H0338	INT_6	0.716	0.805	(0.09)	0.089	38
H0339	EMO_6	0.626	0.537	0.09	0.089	39
H0340	SOCIN_3	0.387	0.476	(0.09)	0.089	39
H0341	ARR_2	0.569	0.652	(0.08)	0.083	41
H0342	EMO_4	0.655	0.572	0.08	0.083	41
H0343	EMO_2	0.768	0.685	0.08	0.083	43
H0344	WAR_3	0.631	0.714	(0.08)	0.083	43
H0345	EMO_7	0.768	0.687	0.08	0.081	45
H0346	EPI_5	0.58	0.499	0.08	0.081	45
H0347	INTEG_2	0.719	0.639	0.08	0.08	47

H0348	SOC_6	0.548	0.627	(0.08)	0.079	48
H0349	ORD_3	0.827	0.75	0.08	0.077	49
H0350	ACH_2	0.67	0.595	0.08	0.075	50
H0351	CON_4	0.811	0.886	(0.08)	0.075	51
H0352	ARR_5	0.523	0.597	(0.07)	0.074	52
H0353	INT_9	0.605	0.531	0.07	0.074	52
H0354	NEG_1	0.832	0.759	0.07	0.073	54
H0355	FAC_5	0.682	0.61	0.07	0.072	55
H0356	ORD_13	0.737	0.809	(0.07)	0.072	55
H0357	HOS_9	0.718	0.787	(0.07)	0.069	57
H0358	FAC_3	0.62	0.689	(0.07)	0.069	58
H0359	WAR_7	0.634	0.703	(0.07)	0.069	58
H0360	INTEG_8	0.754	0.821	(0.07)	0.067	60
H0361	SOCIN_4	0.357	0.423	(0.07)	0.066	61
H0362	EMP_5	0.535	0.601	(0.07)	0.066	62
H0363	INTER_5	0.754	0.691	0.06	0.063	63
H0364	NEG_6	0.587	0.647	(0.06)	0.06	64
H0365	INT_3	0.69	0.63	0.06	0.06	65
H0366	SOC_7	0.258	0.317	(0.06)	0.059	66
H0367	INT_2	0.818	0.759	0.06	0.059	67
H0368	ORD_7	0.714	0.655	0.06	0.059	67
H0369	WAR_11	0.576	0.634	(0.06)	0.058	69
H0370	DEC_2	0.907	0.85	0.06	0.057	70
H0371	TRA_1	0.811	0.866	(0.05)	0.055	71
H0372	PLA_3	0.464	0.41	0.05	0.054	72
H0373	CON_1	0.639	0.692	(0.05)	0.053	73
H0374	PLA_1	0.628	0.68	(0.05)	0.052	74
H0375	ACH_4	0.637	0.689	(0.05)	0.052	75
H0376	DEC_3	0.659	0.71	(0.05)	0.051	76
H0377	EPI_1	0.708	0.758	(0.05)	0.05	77
H0378	EPI_3	0.686	0.636	0.05	0.05	77
H0379	HOS_14	0.593	0.642	(0.05)	0.049	79
H0380	CON_6	0.603	0.651	(0.05)	0.048	80
H0381	INTEG_5	0.708	0.756	(0.05)	0.048	80
H0382	FAC_7	0.378	0.426	(0.05)	0.048	82
H0383	PLA_5	0.555	0.603	(0.05)	0.048	83
H0384	DEC_7	0.676	0.629	0.05	0.047	84
H0385	INT_1	0.768	0.721	0.05	0.047	84
H0386	HOS_3	0.678	0.724	(0.05)	0.046	86
H0387	SOC_3	0.531	0.577	(0.05)	0.046	86
H0388	TRA_2	0.282	0.327	(0.05)	0.045	88
H0389	DEC_5	0.572	0.615	(0.04)	0.043	89
H0390	INTEG_13	0.636	0.593	0.04	0.043	89
H0391	INTEG_10	0.748	0.79	(0.04)	0.042	91
H0392	ORD_9	0.548	0.59	(0.04)	0.042	92
H0393	WAR_6	0.695	0.736	(0.04)	0.041	93
H0394	EPI_4	0.47	0.43	0.04	0.04	94
H0395	BRO_5	0.593	0.553	0.04	0.04	95
H0396	INT_8	0.682	0.643	0.04	0.039	96
H0397	ORD_11	0.48	0.519	(0.04)	0.039	96
H0398	EMP_4	0.536	0.575	(0.04)	0.039	98
H0399	ACH_7	0.469	0.507	(0.04)	0.038	99

H03100	PLA_2	0.572	0.534	0.04	0.038	100
H03101	HOS_2	0.735	0.772	(0.04)	0.037	101
H03102	NEG_5	0.645	0.682	(0.04)	0.037	102

APPENDIX K: GOODNESS OF FIT STATISTICS FOR THE SAPI PARTIAL STRICT INVARIANCE MEASUREMENT MODEL

Contribution to Chi-Square = 65964.752
Percentage Contribution to Chi-Square = 54.201
Root Mean Square Residual (RMR) = 0.0401
Standardized RMR = 0.0595
Goodness of Fit Index (GFI) = 0.694
Degrees of Freedom = 28313
Minimum Fit Function Chi-Square = 121703.915 (P = 0.0)
Normal Theory Weighted Least Squares Chi-Square = 168119.255 (P = 0.0)
Estimated Non-centrality Parameter (NCP) = 139806.255
90 Percent Confidence Interval for NCP = (0.0 ; 0.0)
Minimum Fit Function Value = 28.623
Population Discrepancy Function Value (F0) = 32.880
90 Percent Confidence Interval for F0 = (0.0 ; 0.0)
Root Mean Square Error of Approximation (RMSEA) = 0.0482
90 Percent Confidence Interval for RMSEA = (0.0 ; 0.0)
P-Value for Test of Close Fit (RMSEA < 0.05) = 1.000
Expected Cross-Validation Index (ECVI) = 40.055
90 Percent Confidence Interval for ECVI = (7.095 ; 7.095)
ECVI for Saturated Model = 6.837
ECVI for Independence Model = 571.995
Chi-Square for Independence Model with 28730 Degrees of Freedom = 2431783.659
Independence AIC = 2432463.659
Model AIC = 111493.255
Saturated AIC = 58140.000
Independence CAIC = 2434964.568
Model BIC = -68439.958
Model CAIC = -96752.958
Saturated CAIC = 271967.728
Normed Fit Index (NFI) = 0.931
Non-Normed Fit Index (NNFI) = 0.941
Parsimony Normed Fit Index (PNFI) = 0.917
Comparative Fit Index (CFI) = 0.942
Incremental Fit Index (IFI) = 0.942
Relative Fit Index (RFI) = 0.930
Critical N (CN) = 731.156
Contribution to Chi-Square = 55739.162
Percentage Contribution to Chi-Square = 45.799
Root Mean Square Residual (RMR) = 0.0461
Standardized RMR = 0.0628
Goodness of Fit Index (GFI) = 0.669

APPENDIX L: IDENTIFYING THE LATENT FIRST-ORDER PERSONALITY DIMENSIONS IMPACTED MOST BY BIASED ITEMS

Dimension	Code	Positive or negative trait	% Items biased			Group favoured with uniform bias:[64]	Group favoured with error variance bias: Positive traits: group favoured whose error variance is smaller Negative traits: group favoured whose error variance is larger
			Tau	Theta-delta	Tau and/or Theta-delta		
Achievement Orientation	ACH	Positive	36%	55%	73%	Female group advantaged	Female group advantaged
Arrogance	ARR	*Negative*	100%	50%	100%	Male group advantaged	Male group advantaged
Broad-Mindedness	BRO	Positive	33%	50%	50%	Male group advantaged	Female group advantaged
Conflict-Seeking	CON	*Negative*	43%	57%	57%	Male group advantaged	Male group advantaged
Deceitfulness	DEC	*Negative*	86%	86%	100%	Male group advantaged	Male group advantaged
Emotional Balance	EMO	Positive	50%	50%	88%	Male group advantaged	Male group advantaged
Empathy	EMP	Positive	100%	86%	100%	Female group advantaged	Female group advantaged
Epistemic Curiosity	EPI	Positive	0%	83%	83%	Male group advantaged	Female group advantaged
Facilitating	FAC	Positive	10%	30%	30%	Female group advantaged	Female group advantaged
Hostility–Egoism	HOS	*Negative*	71%	64%	79%	Male group advantaged	Male group advantaged
Integrity	INTEG	Positive	36%	73%	69%	Female group advantaged	Female group advantaged
Intellect	INT	Positive	23%	69%	73%	Male group advantaged	Male group advantaged
Interpersonal Relatedness	INTER	Positive	0%	44%	44%	Female group advantaged	Female group advantaged
Negative Emotionality	NEG	*Negative*	60%	50%	80%	Female group advantaged	Male group advantaged
Orderliness	ORD	Positive	62%	38%	85%	Female group advantaged	Female group advantaged
Playfulness	PLA	Positive	50%	67%	67%	Female group advantaged	Female group advantaged
Sociability	SOC	Positive	57%	43%	71%	Female group advantaged	Male group advantaged
Social Intelligence	SOCIN	Positive	100%	100%	100%	Female group advantaged	Female group advantaged
Traditionalism–Religiosity	TRA	Positive	75%	75%	100%	Female group advantaged	Female group advantaged
Warm-Heartedness	WAR	Positive	64%	73%	73%	Female group advantaged	Female group advantaged

[64] The group who is advantaged with uniform bias is the group who scored higher on average for the first-order dimension.

REFERENCE LIST

Abrahams, F., & Mauer, K. (1999). Qualitative and statistical impacts of home language on responses to the items of the Sixteen Personality Factor Questionnaire (16PF) in South Africa. *South African Journal of Psychology, 29*(2), 76-86.

Albertyn, L. (2003). Psychodynamic perspectives. In Z. Bergh, & A. Theron, *Psychology in the work context* (2nd ed., pp. 303-318). South Africa: Oxford University Press.

Allport, G. (1963). *Pattern and growth in personality.* London: Holt, Rinehart & Winston.

Allport, G., & Odbert, H. (1936). Traitnames. A psycho-lexical study. *Psychological Monographs*, i.

Anderson, P., & Lewis, C. (1998). *PAPI Technical Manual.* London: PA Consulting/Cubiks.

Ayman, R., & Korabik, K. (2010). Leadership: Why gender and culture matter. *American Psychologist, 65*(3), 157-170.

Babbie, E. (2010). *The practice of social research* (12th ed.). Wadsworth: Cengage Learning.

Babbie, E., & Mouton, J. (2001). *The practice of social research.* Cape Town: Oxford University Press.

Bandalos, D., Gerke, J., & Finney, S. (2001). A model of statistics performance based on achievement goal theory. *Journal of Educational Psychology, 95*(3), 604-616.

Bandolos, D. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*(1), 78-102.

Barrick, M., & Mount, M. (2005). Yes, personality matters: Moving on to more imporant matters. *Human performance, 18*(4), 259-372.

Becker, P. (1999). Beyond the Big Five. *Personality and individual differences, 26*, 511-530.

Bedell, B., Van Eeden, R., & Van Staden, F. (1999). Culture as moderator variable in psychological test performance: Issues and trends in South Africa. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 25*(3), 1-7.

Beery, J., Poortinga, Y., Segall, M., & Dasen, P. (2002). *Cross-cultural psychology. Research and applications* (2nd ed.). Cambridge, UK: Cambridge University Press.

Bergh, Z. (2014). The nature of personality and fundamental assumptions in personality study. In Z. Bergh, & D. Geldenhuys, *Psychology in the Work Context* (5th ed., pp. 290-315). Oxford University Press.

Bester, A. (2008). *The establishment of implicit perspectives of personality among Afrikaans speaking people in South Africa.* Unpublished masters' thesis, North-West University, Potschefstroom, South Africa.

Binning, J., & Barrett, G. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*(3), 478-494.

Borman, W., & Mototwidlo, S. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human performance, 10*(2), 99-109.

Bornstein, M., Putnick, D., Bradley, R., Deater-Deckard, K., & Lansford, J. (2016). I. GENDER IN LOW-AND MIDDLE-INCOME COUNTRIES: INTRODUCTION. *Monographs of the Society for Research in Child Development, 81*(1), 7-23.

Borsmboom, D., Mellenbergh, G., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203-219.

Brits, N. (2011). *An explorative investigation into the construct validity of a development assessment centre.* Master's thesis, University of Stellenbosch.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen, & J. Long, *Testing structural equation models.* Newbury Park: Sage Publications.

Bruwer, M. (2016). *Assessing the nomological network of the South African Personality Inventory among industrial psychologists.* Unpublished masters' thesis, Nort-West University, Potchefstroom, South Africa.

Byrne, B., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of cross-cultural psychology, 34*(2), 155-175.

Byrne, B., Shavelson, R., & Muthén, B. (1989). Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement In variance. *Psychological Bulletin, 105*(3), 456-466.

Cattell, R. (1965). *The scientific analysis of personality.* Baltimore: Penguin.

Chen, F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005-1018.

Cheung, F., Cheung, S., Wada, S., & ZhangJ. (2003). Indigenous Measures of Personality Assessment in Asian Countries: A Review. *Psychological Assessment, 15*(3), 280-289.

Cheung, F., Leung, K., Zhang, J., Sun, H., Gan, Y., & Xie, D. (2001). Indigenous Chinese personality constructs: Is the Five Factor Model complete? *Journal of cross-cultural psychology*, 407-433.

Cheung, F., Van de Vijver, F., & Leong, F. (2011, January 24). Toward a new approach to the study of personality in culture. *American Psychologist*, 1-11.

Cheung, G., & Rensvold, R. (1998). Cross-cultural comparisons using non-invariant measurement items. *Applied Behavioral Science Review, 6*(1), 93-110.

Cheung, G., & Rensvold, R. (1999). Testing Factorial Invariance across Groups: A Reconceptualization and Proposed New Method. *Journal of Management, 25*(1), 1-27.

Cheung, G., & Rensvold, R. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255.

Chrystal, E. (2012). *The construction of an indigenous emotional stability scale.* Unpublished masters' thesis, University of Johannesburg, Johannesburg, South Africa.

Cloninger, S. (2009). Conceptual issues in personality theory. In P. Corr, & G. Matthews, *The Cambridge handbook of personality psychology.* New York: Cambridge University Press.

Cohen, F. (2013). *Validation of the Emotional Stability Scale of the South African Personality Inventory.* Unpublished masters' thesis, University of Johannesburg, Johannesburg, South Africa.

Colbert, A., Barrick, M., & Bradley, B. (2014). Personality and leadership composition in top management teams: Implications for organizational effectiveness. *Personnel Psychology, 67*, 351-387.

Costa Jnr., P., Terracciano, A., & McCrae, R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journai of Personality and Social Psychology, 81*(2), 322-331.

Crush, J., & Williams, V. (2001). *Gender concerns in South African Migration Policy. Migration Policy Brief No. 4.* Southern African Migration Project.

Davis, S. (2014). *The measurement invariance and measurement equivalence of the Sources of Work Stress Inventory (SWSI) across gender groups in South Africa.* Stellenbosch: Unpublished masters' thesis, University of Stellenbosch.

De Beer, M. (2004). Use of differential item functioning (DIF) analysis for bias analysis in test construction. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 30*(4), 52-58.

De Raad, B., Barelds, D., Levert, E., Pstendorf, F., Mlačić, B., Blas, L., et al. (2010). Only Three Factors of Personality Description Are Fully ReplicableAcross Languages: A Comparison of 14 Trait Taxonomies. *Journal of personality and social psychology, 98*(1), 160-173.

Deacon, G. (2016). *Exploring the construct validity of the Social Desirability Scale of the South African Personality Inventory .* Unpublished masters' thesis, University of Stellenbosch, Stellenbosch, South Africa.

Department of Higher Education and Training. (2013). *White paper on post-school education and training: Building an expanded, effective and integrated post-school system.* Pretoria: DHET.

Diamantopoulos, A., & Siguaw, J. (2000). *Introducing LISERAL.* London: Sage Publications Ltd.

Donnelly, C. (2009). *A multi-group structural equation modelling Investigation of the measurement invariance of the Campbell Interest and Skill Survey (ciss) across gender Groups in South Africa.* Stellenbosch: Unpublished masters' thesis, University of Stellenbosch.

Douglas, S., & Martinko, M. (2001). Exploring the role of individual differences in the prediction of workplace. *Journal of Applied Psychology, 86*(4), 547-559.

Douglas, S., & Martinko, M. (2001). Exploring the role of individual differences in the prediction of workplace aggression. *Journal of Applied Psychology*(86), 547-559.

Du Plessiss, P. (1987). Die moderne ondernemer en die doelwit van die onderneming. In P. Du Plessiss, *Toegepaste Bedrysfekonomie: 'n Inleidende Oorsig.* Pretoria: HAUM Opvoedkundige Uitgewery.

Du Toit, M., Du Toit, S., & Hawkins, D. (2001). *Interactive LISREL: User's guide.* Lincolnwood IL: Scientific Software International.

Dunbar, H., Theron, C., & Spangenberg, H. (2011). A cross-validation study of the Performance Index. *Bestuursdinamika/Management Dynamics, 20*(3), 2-24.

Dunbar-Isaacson, H. (2006). *An investigation into the measurement invariance of the performance index.* Stellenbosch, South Africa: Unpublished masters' thesis, University of Stellenbosch.

Eagly, A., Johannesen-Schmidt, M., & Van Engen, M. (2003). Transformational, Transactional, and Laissez-Faire Leadership Styles: A Meta-Analysis Comparing Women and Men. *Psychological Bulletin, 129*(4), 569-597.

Fakir, S., & Laher, S. (2015). Perceptions of the utility of personality assessment for personnel selection in the South African context: An exploratory study. *Journal of Psychology in Africa, 25*(5), 482-485.

Farrell, A. (2010). Insufficient discriminant validity: A comment on Bove, Pervan, Beatty, and Shiu (2009). *Journal of Business Research, 63*(3), 324-327.

Feingold, A. (1994). Gender Differences in Personality: A Meta-Analysis. *Psychological Bulletin, 116*(3), 429-456.

Fetvadjiev, V., Meiring, D., Van de Vijver, F., Nel, J., & Hill, C. (2015). The South African Personality Inventory (SAPI): A Culture-Informed Instrument for the Country's Main Ethnocultural Groups. *Psychological Assessment, 27*(3), 827-837.

Foxcroft, C. (2004). Planning a psychological test in the multicultural Sout African context. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 30*(4), 8-15.

Foxcroft, C., & Roodt, G. (2010). *Introduction to psychological assessment in the South African context* (3rd ed.). Cape Town: Oxford University Press Southern Africa (Pty) Ltd.

French, L. (2011). *Construct validation of a preliminary Relationship Harmony Scale within the South African Personality Inventory.* Unpublished masters' thesis, University of Johannesburg, Johannesburg, South Africa.

Furnham, A. (2008). *Personality and intelligence at work: exploring and explaining individual differences at work.* London: Routledge.

Geddes, T. (2012). *Validating an indigenous Extraversion Personality Scale: A cross-cultural study.* Unpublished masters' thesis, University of Johannesburg, Johannesburg, South Africa.

Gonzalez-Mulé, E., DeGeest, D., Kiersch, C., & Mount, M. (2013). Gender differences in personality predictors of counterproductive behavior. *Journal of Managerial Psychology, 28*(4), 333-353.

Graham, J. (2006, December). Congeneric and (essentially) tau-equivalent estamates of score reliability. *Educational and Psychological Measurement, 66*(6), 930-944.

Graham, J., Olchowksi, A., & Gilreath, T. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Preventative Science, 8*(3), 208-213.

Grijalva, E., Newman, D., Tay, L., Donnellan, M., Harms, P., Robins, R., et al. (2015, March). Gender differences in Narcissism: A meta-analytic review. *Psychological Bulletin*, 261-310.

Guenole, N. (2014). Maladaptive Personality at Work: Exploring the Darkness. *Industrial and Organizational Psychology, 7*, 85-97.

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in psychology, 5*, 980.

Guion, R., & Gottier, R. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18*, 135-164.

Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson Education Inc.

Heaven, P., & Pretorius, A. (1998). Personality structure among Black and White South Africans. *The Journal of Social Psychology, 138*(5), 664-666.

Hill, C., Nel, J., Van de Vijver, F., Meiring, D., Valchev, V., Adams, B., et al. (2013). Developing and testing items for the South African Personality Inventory (SAPI). *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 39*(1), 13.

Hogan, R. (2005). In defense of personality measurement: New wine for old whiners. *Human Performance, 18*(4), 331-341.

Holtzkamp, J. (2013). *Measurement Invariance of the second edition of the Fifteen Factor Personality Questionnaire (15FQ+) over different ethnic groups in South Africa.* Unpublished masters' thesis, University of Stellenbosch, Stellenbosch, South Africa.

Horak, S. (2012). *The cross-cultural validation of the Conscientiousness Scale of the South African Personality Inventory.* Unpublished masters' thesis, University of Johannesburg, Johannesburg, South Africa.

Hough, L. (2003). Emerging trends and needs in personality research and practive. In M. Barrick, & A. Ryan, *Personality at work: reconsidering the role of personality in organizations* (pp. 289-315). San Francisco: John Wiley & Sons.

Howard, A., & Thomas, J. (2010). Executive And Manager Assessment. In J. Scott, & D. Reynolds, *Handbook of Workplace Assessment: Evidence-Based Practices for Selecting and Developing Organizational Talent* (pp. 395-436). San Francisco: Jossey-Bass A Wiley Imprint.

Huang, J., Ryan, A., Zabel, K., & Palmer, A. (2014). Personality and Adaptive Performance at Work: A Meta-Analytic Investigation. *Journal of Applied Psychology, 99*(1), 162-179.

International Test Commission. (2001). ITC guidelines on test use. *International Journal of Testing, 1*(2), 93-114.

International Test Commission. (2012). *ITC guidelines for quality control in scoring, test analysis, and reporting of test scores.* International Test Commission.

Jaušovec, N., & Jaušovec, K. (2007). Personality, gender and brain osciilations. *International Journal of Psychophysiology*, 215-224.

Jöreskog, K., & Sörbom, D. (1993). *LISREL 8: Structural Equation Modelling with the SIMPLIS Command Language.* United States of America: Scientific Software International, Inc.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

Judge, T., Bono, J., Ilies, R., & Gerhardt, M. (2002). Personality and Leadership: A Qualitative and Quantitative Review. *Journal of Applied Psychology, 87*(4), 765-780.

Judge, T., Higgins, C., Thoresen, C., & Barrick, M. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology, 52*(3), 621-652.

Jung, E., & Yoon, M. (2016). Comparisons of Three Empirical Methods for Partial Factorial Invariance: Forward, Backward, and Factor-Ratio Tests. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 567-584.

Kelloway, E. (1998). *Using LISREL for structural equation modelling: A researcher's guide.* United States of America: Thousand Oaks Calif. : Sage.

Kemp, T. (2013). *An exploration of social desirability within the white Afrikaans-speaking group.* Unpublished masters' thesis, North-West University, Potschefstroom, South Africa.

Kerlinger, F. (1973). *Foundations of behavioural research* (2nd ed.). New York: Holt, Rinehart and Winston, Inc.

Kerlinger, F., & Lee, H. (2000). *Foundations of behavioural research* (4th ed.). Fort Worth, Tex: Harcourt College Publishers.

Labuschagne, A. (2010). *South African Personality Inventory: The development of an investigation into the psychometric properties of the intellect cluster.* Unpublished masters' thesis, North-West University, Potchefstroom, South Africa.

Laher, S. (2008). EStructural equivalence and the NEO-PI-R: Implications for the applicability of the Five-Factor Model of personality in an African context. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 34*(1), 76-80.

Leong, F., Leung, K., & Cheung, F. (2010). Integrating Cross-Cultural Psychology Research Methods Into Ethnic Minority Psychology. *Cultural Diversity and Ethnic Minority Psychology, 16*(4), 590-597.

Little, T. (1997). Mean and Covariance Structures (MACS) Analyses of Cross-Cultural Data: Practical and Theoretical Issues. *Multivariate Behavioral Research, 32*(1), 53-76.

Little, T., Cunningham, W., Shahar, G., & Widaman, K. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modelling, 9*, 151-173.

Lubbe, M. (2012). *Test mode equivalence in a South African personality context: Paper-and-pencil vs computerised testing.* Unpublished masters' thesis, North-West University, Potchefstroom, South Africa.

MacCallum, R. (1995). Model specification: procedures, strategies and related issues. In R. Hoyle, *Structural equation modelling; concepts, issues and applications.* Thousend Oaks, California: Sage Publications.

MacCallum, R., Browne, M., & Sugawara, H. (1996). Power analysis and determination of sample size for Covariance structure modeling. *Psychological Methods, 1*(2), 130-149.

MacCallum, R., Browne, M., & Sugawara, H. (1996). Power analysis and determination of sample size for covariance structure modeling . *Psychological Methods, 1*(2), 30-149.

Marsella, A., Dubanoski, J., Hamada, W., & Morse, H. (2000). The measurement of personality across cultures: Historical, conceptual, and methodological issues and considerations. *American Behavioural Scientist, 44*(1), 41-62.

Marsh, H., Hau, K., Balla, J., & Grayston, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181-220.

Mavondo, F., Gabbott, M., & Tsarenko, Y. (2003). Measurement invariance of marketing instruments: An implication across countries. *Journal of Marketing Management, 19*(5-6), 523-540.

McCrae, R., & Costa, P. (2010). The Five-Factor theory of personality. In O. John, R. Robins, & L. Pervin, *Handbook of personality* (3rd ed., pp. 159-181). New York: The Guilford Press.

McWilliams, A., Parhankangas, A., Coupet, J., Welch, E., & Barnum, D. (2016). Strategic decision making for the triple bottom line. *Business Strategy and the Environment, 25*, 193-204.

Mead, A., Johnson, E., & Braddy, P. (2008). Power and sensitivity of alternative fit indices in tests of measurment invariance. *Journal of Applied Psychology, 93*(3), 568-592.

Meade, A., & Kroustalis, C. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods, 9*(3), 369-403.

Meiring, D., Van de Vijver, A., Rothmann, S., & Barrick, M. (2005). Construct, item, and method bias of cognitive and personality tests in South Africa. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 31*(1), 1-8.

Meiring, D., Van de Vijver, F., & Rothmann, S. (2006). Bias in an adapted version of the 15FQ+ in South Arica. *South African Journal of Psychology, 36*(2), 340-356.

Mels, G. (2003). *A workshop on structural equation modelling with LISREL 8.54 for Windows.* Chicago, IL: Scientific Software International.

Mels, G. (2010). *Structural Equation Modelling with LISREL 9 for Windows.* Chicago: ScientificSoftware International.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial. *Psychometrika, 58*(4), 525-543.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Mîndrilă, D. (2010). Maximum Likelihood (ML) and Diagonally Weighted Least Squares (DWLS) Estimation Procedures: A Comparison of Estimation Bias with Ordinal and Multivariate Non-Normal Data. *International Journal of Digital Society, 1*(1), 60-66.

Mischel, W. (1976). *Introduction to personality.* USA: Holt, Rinehart and Winston, Inc.

Morgeson, F., Campion, M., Dipboye, M., Hollenbeck, J., Murphy, K., & Schmitt, N. (2007). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Paychology, 60*, 1029-1049.

Morgeson, F., Campion, M., Dipboye, M., Hollenbeck, J., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683-729.

Morris, M., Leung, K., Ames, D., & Lickel, B. (1999). Views from inside and outside: Intergrating emic and etic insights about culture and justice judgement. *The Academy of Management Review, 24*(4), 781-796.

Morton, N. (2011). *Validation of a preliminary Soft-heartedness Scale within the South African Personality Inventory.* Unpublished masters' thesis, Universty of Johannesburg, Johannesburg, South Africa.

Mouton, S. (2017). *An investigation into the first and second -order factor structure of the South African Personality Inventory (SAPI).* Unpublished masters' thesis, University of Stellenbosch, Stellenbosch, South Africa.

Moyo, S. (2009). *A Preliminary Factor Analytic Investigation Into the First-Orde Factor Structure of the Fifteen Factor Questionnaire Plus On a Sample of Black South African Managers.* Stellenbosch University.

Moyo, S., & Theron, C. (2011). A preliminary factor analytic investigation into the first-order factor structure of the Fifteen Factor Plus (15FQ+) on a sample of Black South African managers. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde , 37*(1), 1-22.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*(2), 171-189.

Muthen, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British journal of mathematical and statistical psychology, 45*(1), 19-30.

Nel, J., Valchev, V., Rothmans, S., Van de Vijver, F., Meiring, D., & De Bruin, G. (2012). Exploring the personality structure in the 11 languages of South Africa. *Journal of personality, 80*, 915-948.

Noe, R., Hollenbeck, J., Gerhart, B., & Wright, P. (2010). *Human Resource Management: Gaining a competitive advantage.* New York: McGraw-Hill/Irwin.

Nunnally, J. (1978). *Psychometric Theory* (2 ed.). New York: McGraw-Hill.

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modelling. *European Journal of Operational Research, 151*, 53-79.

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods formissing data in structural equation modeling. *European Journal of Operational Research, 151*, 53-79.

Ones, D., Dilchert, S., Viswesvaran, C., & Judge, T. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995-1027.

Owen, K. (1991). Test bias: The validity of the Junior Aptitude Tests (JAT) for various population groups in South Africa regarding constructs measured [Abstract]. *South African Journal of Psychology, 21*(2), 112-118.

Palmer, T., & Flanagan, D. (2016). The sustainable company: looking at goals for people, planet and profits. *Journal of Business Strategy, 37*(6), 28-38.

Pervin, A., Cervone, D., & John, O. (2005). Trait approaches to personality: Allport, Eysenck and Cattell. In A. Pervin, D. Cervone, & O. Johan, *Personality: Theory and Research* (pp. 221-250). New Jersey: John Wiley & Sons, Inc.

Pervin, A., Cervone, D., & John, O. (2005). Trait theory: The five factor model; applications and evaluation of the trait approaches to personality. In A. Pervin, D. Cervone, & O. John, *Personality: Theory and Research* (pp. 251-292). USA: John Wiley & Sons, Inc.

Phares, E. (1984). Psychoanalytic Theory I: The Freudian Revolution. In E. Phares, *Introduction to personality* (pp. 62-88). Ohio: Charles E Merril Publishing Company.

Ployhart, R., Lim, B., & Chan, K. (2001). Exploring relations between typical and maximum performance ratings and the five factor model of personality. *Personnel Psychology, 54*(4), 809-843.

Preacher, K., & Coffman, D. (2006). Computing power and minimum sample saize for RMSEA. *Computer software]. Available from http://quantpsy. org.*

Putnick, D., & Bornstein, M. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90.

Ramsay, L., Taylor, N., De Bruin, G., & Meiring, D. (2008). The Big Five personality factors at work: A South African validation study. In J. Deller, *Research contributions to personality at work* (pp. 99-114). Munich, Germany: Rainer Hampp Verlag.

Republic of South Africa (RSA). (1998). *Employement Equity Act 55 of 1998.* Juta Law.

Republic of South Africa (RSA). (2006). *Ethical and professional rules of conduct for practitioners registered under the health professions act (Act 56 of 1974).*

Samuel, D., South, S., & Griffin, S. (2015). Factorial invariance of the Five-Factor Model rating form across gender. *Assessment, 22*(1), 65-75.

Saucier, G. (2008). Measures of the personality factors found recurrently in human lexions. In G. Boyle, G. Matthews, & D. Saklofske, *The SAGE handbook of personality theory and assessment: Personality measurement and testing* (Vol. 2, pp. 29-54). London, UK: SAGE Publications.

Saucier, G. (2009). Semantic and linguistic aspects of personality. In P. Corr, & G. Matthews, *The Cambridge handbook of personality psychology* (pp. 379-399). New York: Cambridge University Press.

Schmidt, F., & Hunter, J. (1998). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings. *Psychological Bulletin, 124*(2), 262-274.

Schmidt, F., Oh, I., & Shaffer, J. (2016, October 17). *The validity and utility of selection methods in Personnel psychology: Practical and theoretical Implications of 100 years.* Retrieved July 3, 2018, from ResearchGate: https://www.researchgate.net/publication/309203898_The_Validity_and_Utility_of_Selection_Methods_in_Personnel_Psychology_Practical_and_Theoretical_Implications_of_100_Years_of_Research_Findings?enrichId=rgreq-ac03d29ee90470b28e65c5b16b199479-XXX&enrichSou

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selectio procedures* (4th ed.). Bowling Green OH: SIOP.

Spangenberg, H., & Theron, C. (2005). Promoting ethical follower behaviour through leadership of ethics: the development and psychometric evaluation of the Ethical Leadership Inventory (ELI). *South African Journal of Business Management, 36*(2), 1-18.

Steenkamp, J., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national research. *Journal of Consumer Research, 25*(June), 28-90.

Stellenbosch University. (2013). Policy for responsible research conduct at Stellenbosch University. Stellenboisch University.

Strydom, H. (2014). Sampling in the quantitative paradigm. In H. Strydom, C. B. Fouche, & C. S. Delport, *Research at Grassroots for Social Sciences and Human Service Professions* (4th ed., pp. 222-235). Pretoria: Vam Schaik Publishers.

Su, R., Rounds, J., & Armstrong, P. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin, 135*(6), 859-884.

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). Boston MA: Allyn and Bacon, Pearson Education.

Tett, R., & Christiansen, N. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology, 60*(4), 967-993.

Theron, C. (2007). Confessions, Scapegoats and Flying Pigs: Psychometric Testing and the Law. *SA Journal of Industrial Psychology, 33*(1), 102-117.

Theron, C. (2011a). *Lecture Series 2: Psychological and Performance Assessment in Personnel Selection.* Unpublished class notes (Industrial Psychology 743), University of Stellenbosch.

Theron, C. (2012). *Research designs in explanatory research.* Stellenbosch: Unpublished class notes (Industrial Psychology 776), Stellenbosch University.

Theron, C. (2012). *Research designs in explanatory research. Unpublished class notes (Industrial Psychology 776).* University of Stellenbosch.

Theron, C. (2016). *Intermediate Statistics and Computer Usage.* Stellenbosch: Unpublished class notes (Industrial Psychology (873), University of Stellenbosch.

Tracey, J., & Hinkin, T. (2010). Contextual factors and cost profiles associated with employee turnover. In C. Enz, *The Cornell School of Hotel Administration handbook of applied hospitality strategy* (pp. 736-753). Los Angeles, CA: SAGE.

Valchev, V., Van de Vijver, F., Nel, A., Rothmann, S., Meiring, D., & De Bruin, G. (2011). Implicit Personality Conceptions of the Nguni Cultural- Linguistic Groups of South Africa. *Cross-cultural Research, 45*(3), 235-266.

Valchev, V., Van de Vijver, F., Nel, J., Rothmann, S., & Meiring, D. (2013). The use of traits and contextual information in free personality descriptions across ethnocultural groups in South Africa. *Journal of Personality and Social Psychology, 104*(6), 1077.

Van Aarde, N., Meiring, D. & Wiernik, B.M. (2017). The validity of the big five personality traits for job performance: Meta-analyses of South African Studies. *International Journal of Selection and Assessment*. 25, 223-239.

Van de Vijver, A., & Poortinga, Y. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*(1), 29-37.

Van de Vijver, A., & Rothmann, S. (2004). Assessment in multicultural groups: The South African case. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 30*(4), 1-7.

Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Thousend Oaks, CA: Sage Publications Inc.

Van de Vijver, F., & Leung, K. (2001). Personality in cultural context: Methodological issues. *Journal of Personality, 69*(6), 1007-1031.

Van de Vijver, F., & Tanzer, N. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*(4), 263-280.

Van de Vijver, F., & Tanzer, N. (2004). Bias and equivalence in cross-cultural assessment: an overview. *European Review of Applied Psychology, 54*, 119-135.

Van Heerden, S. (2011). *Modification, elaboration, and empirical evaluation of the De Goede learning potential structural model.* Unpublished master's research proposal, University of Stellenbosch.

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Vianello, M., Schnabel, K., Sriram, N., & Nosek, B. (2013). Gender differences in implicit and explicit personality traits. *Personality and individual differences, 55*(8), 994 - 999.

Visser, D., & Du Toit, J. (2004). Using the Occupational Personality Questionnaire (OPQ) for measuring broad traits. *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde, 30*(4), 65-77.

Walker, M. (2018). Aspirations and equality in higher education gender in a South African university. *Cambridge Journal of Education, 48*(1), 123-139.

Weiten, W. (2011). *Psychology: Themes and variations* (7th ed.). USA: Thomson Wadsworth.

Wikipedia. (2018, July 20). *Sex and gender distinction.* Retrieved July 21, 2018, from Wikipedia: https://en.wikipedia.org/wiki/Sex_and_gender_distinction

World Economic Forum. (2016). *Executive Summary: The industry gender gap.* Geneva: World Economic Forum.

World Economic Forum. (2016). *The future of jobs: Employment, skills and workforce strategy for the Fourth Industrial Revolution.* Geneva, Swtizerland: World Economic Forum.

Yoon, M., & Kim, E. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavioural research methods, 46*, 1199-1206.

Yoon, M., & Millsap, R. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo Study. *Structural Equation Modeling, 14*(3), 435-463.

Zell, E., Krizan, Z., & Teeter, S. (2015). Global gender differences can be operationalized and tested. *American Psychologist, 70*(1), 10-20.