UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

A Rasch analysis of the psychometric properties of the High Potential Trait Indicator (HPTi) using a South African sample

by

David Sanele Semmelink

A mini-dissertation submitted in partial fulfilment

of the requirements for the degree

Master of Arts (Research Psychology)

to the

Faculty of Humanities

Department of Psychology

at the

University of Pretoria

SUPERVISOR: Professor David J.F. Maree

August 2021

Declaration

I, **David Sanele Semmelink**, hereby declare this work, submitted in partial fulfilment for the MA Research Psychology degree, is my own. This work has not previously been submitted to any other institution or for any other degree. Where possible, all efforts have been made to credit the works of others and in a manner that meets the standards of the University's code of conduct and the American Psychological Association's (APA) guidelines.

Signature: 

Date: 17 August 2021

Ethics Statement

The Author, whose name appears on the title page of this mini-dissertation, has obtained the applicable research ethics approval for the research described in this work.

The Author declares that he has observed the ethical standards required in terms of the University of Pretoria's code of ethics for research and the policy for responsible research.

Acknowledgements

I am grateful to my parents for the support leading up to this day. Without them I would not be here.

I am highly appreciative to Thomas International and the individuals involved in permitting me to carry out this research.

And, of course, to Prof. David Maree. I would not have even known about Rasch methodology if not for you, and I am very grateful for the guidance you have given me in navigating through this topic.

Abstract

Psychometric assessments commonly form part of the decision-making process in various industries. The High Potential Trait Indicator (HPTi) is one such psychometric assessment; a six-trait personality-based questionnaire designed for targeting senior leadership positions. The psychometric tool began its development in the UK by Ian MaCrae and Adrian Furnham (2014; 2016), and has since been utilized worldwide.

Notwithstanding the history of South Africa and the use of psychometric instruments in the country, such tools are generally required to possess scientific evidence of being reliable and valid. However, given the history, South African legislation requires that all psychological assessments are found to be valid and reliable in order to be used for, and not limited to, their use in human resource related contexts.

A measurement theory, known as Rasch Measurement Theory – and its related analytical techniques, – is regarded as achieving the standards required to be considered the definition of measurement. Therefore, should a psychometric instrument meet certain requirements of the Rasch model, the instrument may be regarded as a tool that meets the definition of measurement.

This study, therefore, evaluates the reliability and validity of the HPTi using Rasch analysis techniques. While there are many criteria to evaluate an instrument on within the family of Rasch analysis techniques, this study focused on the unidimensionality and local independence of each subscale, the fit to the Rasch model, person reliability and separation, and differential item functioning (DIF) as key components to determining elements of the reliability and validity of the HPTi. Secondary data, obtained from copywrite holder, Thomas International, was obtained and used for analysis. The initial sample consisted of 1257 respondents, with further reductions taking place for each of the six traits based on the person fit statistics.

The analysis of reliability found that one of the six traits, *Ambiguity acceptance*, would not be considered reliable, with the other five ranging from a reliability index of .70 to .76, where a score of .70 is the minimum required to be considered reliable.

An item fit analysis revealed 14 of the 78 HPTi items did not fit the Rasch model. Of the 14, 10 underfit the model (more varied than the expected model) and 4 underfit the model (less varied than the expected model).

In analysing the differential item functioning of the instrument, no items were found to be functioning differently between genders, female and male. However, 12 items were evaluated to possess differential item functioning across ethnicities and home language groups. The most severe instance of DIF occurred in trait *Competitiveness,* yet this scale had only one item experiencing DIF. On the other hand, trait *Conscientiousness* contained four items experiencing various severities of DIF.

Key words: *Psychometric properties; High Potential Trait Indicator (HPTi); Rasch model fit; Person reliability; Differential item functioning*

Table of Contents

List of Figures

List of Tables

List of Appendices

Chapter 1:
Introduction and Rationale

1.1. Introduction

Literature has exposed how psychometric instruments measuring latent variables, such as surveys or questionnaires, commonly measure variables non-linearly and at the ordinal level (Bond, Yi & Heene, 2020; Boone, Staver & Yale, 2014; Wright & Mok, 2004). The disproportionate and non-linear ordinal-level measurement tend to be the result of the reliance on Classical Test Theory techniques (Boone et al., 2014). The ordinal-level data derived from such instruments are in violation of the assumption of interval-level data in the parametric statistical procedures often used to analyse such data. Rasch analysis techniques provide, not only a transformation from ordinal to interval level data, but also a robust model to which psychological assessments can fit to improve their psychometric properties (Linacre, 2004; Wright & Mok, 2004).

The Rasch model is typically referred to as the one-parameter logistic model (1PL), a component of a psychometric theory known as Item Response Theory (Bond et al., 2020; Boone et al., 2014; Hambleton & Jones, 1993; Reise & Henson, 2003). The two measurement theories, namely Classical Test Theory (CTT) and Item Response Theory (IRT) will therefore be referred to throughout, with a focus on the Rasch model.

The psychometric instrument titled the High Potential Trait Indicator (HPTi) is a personality assessment developed from a CTT perspective, by Adrian Furnham and Ian MacRae (2016). The assessment has a 7-point Likert-type response scale. It is aimed at measuring six personality traits found to predict the potential for success at the senior leadership level of organisations (MacRae & Furnham). However, according to the test manual (Macrae & Furnham), the development of the HPTi was based on a British sample, further warranting the need for the study within the South African context. Searches for peer-reviewed articles did not produce studies on the HPTi in the South African context. Thus, through a Rasch measurement approach, the findings of this study could determine whether the personality trait measurements of the HPTi are unidimensional and fit the Rasch model, are found to be reliable using Rasch analysis techniques, contain item bias assessed using Rasch measurement techniques, and ultimately contribute to the psychometric body of knowledge. The tool will be discussed further in the methodology section.

The reason for the investigation from a Rasch perspective is multi-faceted. First, Reise and Henson (2003) have indicated that while IRT holds various advantages over CTT techniques, some developers of personality assessments continue to utilize CTT techniques, as professionals are yet to be convinced to forego CTT in favour of the more complex IRT. Therefore secondly, the purpose is to evaluate the psychometric properties of the personality assessment based on a model which is able to transform the inherent ordinal data of psychometric tools to interval level data – a key benefit asserted by Rasch techniques (Wright & Mok, 2004). This transformation would make the data obtained by the assessment more useful for conventional parametric statistical analyses (Wright & Mok, 2004), amongst other benefits. Finally, the researcher hopes to achieve the outlined requirements by Wright and Mok (2004) for a measure which can construct inferences from observation. These requirements are that the "measurement model must: (a) produce linear measures, (b) overcome missing data, (c) give estimates of precision, (d) have devices for detecting misfitting items/persons, and (e) the parameters of the object being measured and of the measurement instrument must be separable," achievable only through Rasch measurement techniques (Wright & Mok, 2004, p. 4).

Thus, if the HPTi meets the aims outlined below, the tool will have evidence to show that it is a reliable instrument that satisfies the requirements of being a robust measurement instrument, as explained by Wright and Mok (2004) above.

1.2. Aim of the Study

This study aims to analyse the psychometric properties of the High Potential Trait Indicator (HPTi) from a Rasch measurement approach, an evaluation not yet conducted for this instrument. The study would be quantitative, as the researcher aims to investigate the statistical properties of the personality questionnaire from a Rasch perspective. This includes the psychometric properties: reliability and validity; as well as the statistical *fit* indices to the Rasch model, and differential item functioning.

1.3. Research questions

The following questions are therefore sought to be answered in this study:

- Are the measurements of each trait measured in the HPTi reliable using Rasch person reliability and person separation?
- Are each of the HPTi subscales unidimensional?

- Do the items for each trait measured in the HPTi fit the Rasch model?
- Do any of the items for each trait measured by the HPTi display differential item functioning in groups gender, ethnicity, and first language, as evaluated through Rasch analysis techniques?

1.4. Objectives of the study

From the aim and research questions, the following objectives can be derived:

- Examine the Rasch model fit indices for each trait measured by the HPTi;
- Evaluate the person and item reliability and separation indices of each trait measurement of the HPTi; and,
- Determine the presence of differential item functioning across groups: Gender, ethnicity, and first language using the available Rasch techniques from Winsteps (Linacre, 2020).

1.5. Hypotheses

The following hypotheses are therefore tested in this study:

- HPTi traits, *Conscientiousness*, *Adjustment*, *Curiosity*, *Risk approach*, *Ambiguity acceptance*, and *Competitiveness*, achieve adequate reliability indices as measured by person reliability and separation.
- The HPTi subscales are unidimensional.
- The items of each of the six HPTi traits are found to fit the Rasch model, as measured by item fit statistics: Infit mean-squared and infit z-standardised.
- The items of each of the six HPTi traits do not display instances of differential item functioning, as indicated by the Rasch-Welch and DIF contrast statistics.

Chapter 2:
Literature review

2.1. Chapter introduction

In this chapter, the researcher will provide an argument as to the importance of analysing the psychometric properties of the HPTi using Rasch analysis techniques using a South African Sample.

First, psychometrics will be outlined as a body of Psychology and its uses, before being contextualised in South Africa. The brief history of psychometrics in the South African context discusses the misuse and cultural misalignment of psychological assessments, given the nation's history, which lead to the promulgation for assessments to be culturally fair and unbiased, valid, and reliable to be used in the country.

Three psychometric theories will then be discussed. These are Classical Test Theory (CTT), Item Response Theory (IRT), and Rasch Measurement Theory (RMT). This discussion aims to highlight some of the benefits of analysing a CTT-based assessment using Rasch techniques. A search for literature regarding the Rasch analysis of the HPTi via Google, Google Scholar and the University of Pretoria's online library revealed no related studies, further warranting research of this nature.

2.2. Psychometrics

Psychometrics is the subfield in Psychology referring to the scientific and systematic process in which measurement of psychological constructs are developed. The field ensures the scientific soundness of psychometric assessments through the requirements of technical measurement standards such as reliability and validity (Foxcroft & Roodt, 2018). Coaley (2010, p. 2) classifies the process of psychological assessment as "designed to describe, predict, explain, diagnose and make decisions about" an individual or groups of individuals.

The use of psychometric testing in decision making has become commonplace in various sectors. This includes education, human resources, coaching, forensics, counselling, in medical and clinical applications, and even in sectors relying on psychological assessment for predicting behaviours which may have an influence in financial implications (Arráiz, Bruhn & Stucchi, 2016; Bichi, 2016; Coaley, 2010; Foxcroft & Roodt, 2018). Psychometric tools, therefore, have a growing importance in varying settings globally, as they provide an

objective measurement of psychological constructs not easily observed (Foxcroft & Roodt, 2018).

Consequently, precision is important in the measurability of psychological traits. Where assessments built on a CTT model centre around a person's total test score (Bichi, 2016), assessments built on IRT models are focused at the item level, and person-item interaction. Person-item interaction can be regarded as the relationship between the ability or trait of a person as measured by the instrument and an item response; a feature not present in CTT-based assessments (DeMars, 2010).

2.2.1. Brief history of psychometrics in South Africa

The use of psychometrics in South Africa has had a tainted history, given the nation's history with inequality and biased regimes (Foxcroft, Roodt & Abrahams, 2018; Laher & Cockroft, 2013). Foxcroft et al. illustrate how, in South Africa, psychological assessments were initially brought to the nation post the colonisation of South Africa, and subsequently adapted and standardised for use in the White, Westernised communities (Huysamen, 2002; Van der Vijver & Rothmann, 2004). According to Meiring, Van der Vijver, Rothmann and Barrick (2005), the 1920s saw the rise of cross-cultural issues in psychological testing, yet only in the 1980s did practices of bias, fairness, and discrimination in testing receive attention in line with international standards. Foxcroft et al. continue to illuminate how even tests created in South Africa at the time received the same culturally exclusive treatment, and how tests at the time were used to draw conclusions which were not in favour of the oppressed groups in the country.

In contention to the practices of the time, Biesheuval (1943) went on to challenge the cultural appropriateness of Euro-centric intelligence tests for use in the Black communities. Similarly, Foxcroft et al. (2018) highlight the different cognitive conclusions discovered about an indigenous group – the 'San' – when the measure used to assess the cognitive ability in question of the San group was adapted to be less Euro-centric. The initial study by Porteus (1937) made use of the paper-and-pencil test administration method which the San were arguably not familiar with and therefore performed poorly on, whereas the replication of the study by Reuning and Worley (1973) made use of a wooden model that individuals would tilt in various directions in order to complete the same task, rather than paper-and-pencil. Although directed to minority groups of the USA, Abrahams and Mauer (1999) attributed this

phenomenon to the issue of different groups having not been exposed to the same material used in the testing method, such as paper-and-pencil versus a manoeuvrable wooden object.

A related issue of more recent times may involve the use of computers for testing instead of, for example, paper-and-pencil tests. Tredoux (2013) makes mention of the possible inaccessibility and limited exposure to computers for people from rural and less affluent backgrounds which may result in an aversion and anxiety of the test-taker to computers when they are required to complete a computerized test, adding to the negative impact of bias and fairness of testing.

Furthermore, Taylor (1987) advised that further research on all tests be conducted to identify the existence and extent to which bias exists, to which Abrahams and Mauer (1999) point out that numerous studies have found cognitive tests to be biased when different groups were compared. Abrahams and Mauer highlight a few studies investigating the comparability of results to different groups in the South African population. In Spence's (1982) study, the full South African Personality Questionnaire (SAPQ) designed by Steyn (1977) yielded insufficient alpha coefficients in a group of Black respondents of the questionnaire, even after invalid items were removed from the questionnaire. Also on the SAPQ, while Taylor and Boeyens (1991) found better results in the comparability of constructs between groups, most of the items failed to meet the no-bias criteria set, with the authors giving language the spotlight in their indication for the difficulties of the instrument. Abrahams and Mauer then found in their investigation of the Sixteen Personality Factor Questionnaire (16PF; Cattell, Eber & Tatsuoka, 1992) that Black and White participants reacted differently to at least 50% of items of the 16PF. Further to this discovery, Abrahams and Mauer found in the qualitative extension of their study that these differences are likely due to language.

However, it is not always the case that all psychological assessments are culturally biased. For example, Meiring, Van der Vijver, Rothmann, and Barrick (2005) found that the items for the Fifteen (15FQ+) did not constitute having a major cultural bias. In Heuchart, Parker, Strumf, and Myburg's (2000) study, the NEO Personality Inventory Revised (NEO PI R) discovered a clear five factor model for both Black and White respondents, indicating no evidence for structural bias in the personality assessment. Similarly, a 2017 study by Brouwers, Mostert and Mtshali found the Strengths Use and Deficit Correction Quetionnaire (SUDCO) did not meet the criteria to be bias across language three language groups in the South African Context.

In combating the misuse of psychological tools in the multicultural South African context, the promulgation of the Employment Equity Act 55 of 1998 (EEA, Government Gazette, 1998) demanded a stronger cultural appropriateness of psychological instruments. According to the EEA, psychological assessments such as psychometrics are required to show scientific evidence that they are valid and reliable, applied fairly, and not biased against groups. Further to this, the Health Professions Act 56 of 1974 (Government Gazette, 2006) indicates that a psychologist using a psychological assessment must be familiar with the reliability, validity, standardisations and outcome studies of the assessment, as well as recognise any limits of the assessment with regards to linguistic, cultural, and socio-economic differences. Thus, according to Van der Vijver and Rothmann (2004), assessment practitioners should ensure that current practises are in line with legal demands. One example in doing so, provided by Vijver and Rothmann, is to either develop new instruments designed for use in multicultural environments or validate existing instruments to meet this requirement. The latter will be the focus of the next section focusing on the practice of adapting tests to suit the new context they find themselves in.

2.2.2. Culture and imported tools

Foxcroft, Roodt & Abrahams (2018) indicate that test development has had a history of combating the issues of producing culture-free assessments. While culture-free assessments were attempted, it became apparent that ‘culture-free’ was an unrealistic goal, and instead the aim should be to reduce the influence of culture in testing rather than eradicate it altogether (Foxcroft, Roodt & Abrahams). Tactics used to reduce cultural influences in assessments involved, as far as possible, removing cultural bias in the forms of language use and measuring traits common in different cultures.

This is compounded by the fact that most psychometric tests have historically had routes in the USA and the UK (Foxcroft, Roodt & Abrahams, 2018). This is not untrue of the HPTi, which had been developed in the UK, with further studies conducted primarily in the USA and Canada (MacRae, 2012; Macrae & Furnham, 2014; Macrae & Furnham, 2016; MacRae & Furnham, 2020), which Foxcroft et al. suggests results in the assessment being more appropriate for Enlish-speaking, Westernised individuals. It should be noted at this point that MacRae and Furnham’s (2020) study did include South Africans in their sample but they made up 4.4% of the sample. However, research into the reliability, validity, and bias of the HPTi has not been conducted in the South African context.

2.2.3. Reliability of psychometrics

As discussed previously, one of the key elements of psychological assessments required by the EEA is that a test must show evidence of being reliable. Reliability of a psychometric instrument has been described as "whether an instrument can be interpreted consistently across different situations" (Fields, 2009, p. 11), "how free it is from random error" (Pallant, 2011, p. 6), and the consistency of the scores of a test when obtained under similar circumstances (Price, 2017), giving a similar "result for a person each time it is used…" and "…on different occasions" (Coaley, 2010). Therefore, reliability in psychometric assessments can be defined as how consistent a psychometric instrument is for each respondent, and therefore how resistant the instrument is to random error.

A typical measure of test reliability is internal consistency, assessed by either Cronbach's (1951) coefficient alpha or Kuder-Richardson's various formulae for internal consistency (Pallant, 2011; Price, 2017). Cronbach's alpha is denoted as α and Kuder-Richardson's formula is denoted as KR suffixed by the formula number, such as KR20 for Kuder-Richard's formula 20. It was the case in Spence's (1982) study, mentioned earlier, that the SAPQ yielded an insufficient coefficient alpha in the Black respondent group. This was in reference to the reliability of the SAPQ when administered to the Black respondent group, as measured by Cronbach's alpha.

A commonly used alpha coefficient to describe a sufficiently reliable instrument is stated as .70 and higher (DeVillis, 2017; Kaplan & Saccuzzo, 2017; MacRae & Furnham, 2016; Nunnally, 1978; Pallant, 2020; Yang & Green, 2011). However, this number is arbitrary and susceptible to contention. For example, Smit (1996) recommends a coefficient of .80 for personality questionnaires, with Anastasi and Urbina (1997) agreeing that all standardised measures should have a reliability of .80. Huysamen (1996), however, would impose a stricter coefficient of .85 for measures used in making decisions about individuals. In Peterson's (1994) meta-analysis of Cronbach's alpha, the average alpha coefficient was .77 (n = 4286), however nearly half (49%) were .80 and above. Only 14% reached .90 or higher (Peterson, 1994).

Thus far, the typical measure of reliability discussed has been Cronbach's alpha, which is typical for instruments developed under Classical Test Theory techniques (Crocker & Algina, 2008). Classical Test Theory is a body of psychometric-related theories which will be discussed in the next section of this chapter. However, Rasch reliability statistics, known

as person reliability, may be preferred over a statistic such as Cronbach's alpha, as alpha makes use of raw, nonlinear, and test-dependent data, whereas Rasch reliability statistics make use of transformed, linear, parameter-invariant data (Boone et al., 2014). Person reliability and Rasch analysis techniques will be discussed in more detail in a section further on in this chapter. Furthermore, Cronbach's alpha has been criticised for its misuse as an estimate for internal consistency, as well as it not being the best measure of reliability, with Sijtsma (2009) going as far as to say "The only reason to report alpha is that top journals tend to accept articles that use statistical methods that have been around for a long time such as alpha" (p.118).

2.3. Classical Test Theory (CTT)

CTT is a body of psychometric-related theories, which arose in the early 20th century (Traub, 1997). The rise of CTT can be attributed to Edward Thorndike, Karl Pearson and Charles Spearman in 1904, however, its prevalence was realized by Melvin Novick in his (1966) publication, *The axioms and principal results of classical test theory* (Crocker & Algina, 2008; Traub, 1997). Key to CTT is that an individual measured on a variable has a true score, usually denoted as *T*. However, instruments measuring this variable, and the contexts in which the variable is being measured contain errors, denoted as *E*. Therefore, scores measured by these instruments can only provide the observed score, *X*, consisting of the true score and measurement error, therefore:

$$X = T + E$$

Provided an individual's trait or ability measured does not change between multiple testings, the difference between the individual's observed scores is therefore a result of the error inherent to psychometric testing, including the content of the test and the context of the testing (Croker & Algina, 2008; Magno, 2009). For example, issues such as fatigue, forgetfulness and guessing can influence the difference in an individual's test scores. In other words, the true ability or trait of an individual is not likely to change with each measurement, but rather the changes in scores are subject to the imprecise nature of the assessment tools used to measure them, and context in which the respondent completed the test.

CTT's focus around the total test score limits tests to raw score data, which remains at the ordinal level of measurement (Reise & Henson, 2003). Given the inherent ordinal-level data of Likert-type scales, an exact mathematical relationship between the response options does not exist (Crocker & Algina, 2008). For example, on a personality questionnaire using a

Likert-type scale from *(1) Strongly disagree* to *(7) Strongly agree*, a response of *(6) Agree* is not double the agreement of, say, *(3) Somewhat disagree*. The difference between assigned numbers on a Likert-type scale are therefore not equal and can even measure variables disproportionately, unlike the equal measurements on a meterstick, for example (Boone et al., 2014). Thus, measurement practices such as using Likert-type scales at most produces ordinal scales and not interval scales where the distance between numbers/anchors are equal. Consequently, to make use of such data for parametric statistical techniques violates the requirement of an interval or higher level of measurement (Field, 2009; Pallant, 2020).

CTT techniques do attempt to address and enhance the precision of assessments with statistical techniques such as Cronbach's alpha, examining item difficulty, and item discrimination (Finch, Immekus & French, 2016). However, Smiley (2015) expresses that CTT analyses are group-centred and is reliant on the nature of the group, rather than the specifications of the instrument itself. In other words, CTT analyses are sample dependent, a limitation that is historically acknowledged (Coaley, 2010; Hambelton, 2000; Magno, 2009). This limitation is avoidable if successive samples were representative and not varied across time; a limitation requiring intricate tactics to subdue (Magno, 2009). Moreover, Cronbach's alpha, a statistic used to determine the reliability of an instrument, measures the correlation of the items of the test and is easily manipulated by adjusting the number of items in the assessment. This statistical technique also fails to consider the extent to which a single construct is being measured by the items; a property referred to as *unidimensionality* (Tomyn, Stokes, Cummins & Dias, 2020). However, benefits of CTT based instruments seem to include its simple, practical, affordable nature, relatively weak theoretical assumptions, and the use of relatively simple computations which do not require a strong statistical background or sophisticated software (Fan, 1998; Manapsal, 2017; Smiley, 2015).

Furthermore, Manapsal (2017) discusses three key shortcomings of CTT based assessments. The first is that of person-dependence. This phenomenon occurs when items cannot be removed from the characteristics of those who answered them. In other words, the statistics brought forth from an assessment administered to university students across a region of the country are reflective of the university students who completed those items and not university students across the country. These same statistics are also not reflective of people who are not in a university. The requirement, therefore, is that norms representing differing groups must be created to adequately compare test takers according to which group they are representing. Norms typically require updating to stay relevant which can be disruptive if the

new set of norms creates a different comparative result to the previous norms. For example, if a person was ranked relatively high on a cognitive assessment against a norm group at a particular time, and then receives a lower ranking from the same score but against an updated norm group. The second shortcoming mentioned is that CTT is item-dependent. The person-ability statistic – "the observed score for a given set of items" (Manapsal, 2017, p. 268) – is reflective of the difficulty of the items administered. The score and the norm, therefore, become dependent on the achievement of the sample included in the assessment; and item difficulty is varied depending on the sample of respondents completing the test. Fan (1998) refers to these interactive dependencies as circular dependence. The third limitation mentioned by Manapsal (2017) is the assumption CTT holds, equating the measurement error variance across all test takers, rather than computing the errors per respondent – a technique used in the Rasch model. It should be stated, however, that Lord (1965) mentions the true-score model does maintain its utility, especially when the stronger mathematical models, which provide stronger inferences when successful, fail.

2.4. Item Response Theory (IRT)

IRT possesses stronger assumptions in comparison to CTT, and shifts the analysis from the test to the item. Where analysis of CTT based assessments evaluate the test, analysis of IRT based assessments focus on the probability of endorsing an item – or the probability of getting an item right or wrong (Magno, 2009). It was developed to address the issues surrounding CTT, however, due to its mathematical intensity, has become a topic for debate for its effectiveness in creating psychological measures (Coaley, 2010). Thurstone conceptualized the model as having characteristics of both the person and item being placed on the same "quantitative latent continuum" (Kean, Bisson, Brodke, Biber & Gross, 2017, p. 97). The placement of items and persons on the same scale (measured in logits) allows for the calculation of the probability of success/endorsement of persons on items with different logit values. This extends to calculating the probability of success/endorsement of persons on items in parallel test forms when the logit values of the person and of the second test form are known (Kean et al., 2017).

The probability of answering correctly or endorsing an item (denoted as i), in IRT, is denoted as P_i, and is a function of a respondents underlying ability or trait, denoted as θ. The expression for calculating one's probability of success as a function of their ability is then expressed as:

$$P_i(\theta) = c_i + (1 - c_i)[1 + e^{-Da_i(\theta - b_i)}]^{-1}$$

where *c* is the parameter referred to as pseudoguessing; *b* is the parameter associated with the difficulty of the item; and *a* is the item discrimination parameter; and *D* is the scaling constant of 1.7. In a graphical representation of these parameters, *c* is the y-intercept, *a* is the slope of the graph at the location of *b* which is the point on the scale which represents a respondent's ability (Hambleton & Jones, 1993). According to Hambleton and Jones, these parameters are what form the three accepted logistic models in IRT, namely, the one-parameter (1PL), two parameter (2PL), and three-parameter logistic (3PL) models. When *a*, *b*, and *c* are not constrained, the 3PL model is employed. When the *c* parameter is constrained to 0, then the 2PL model is being used. Whereas when the *c* parameter is constrained to 0, and *a* is fixed, the 1PL model is utilized (Hambleton & Jones, 1993; Kean et al., 2017).

Key to its advantages over CTT, IRT-based instruments are capable of having interval-level measurement, item-level parameters, parameter invariance, and are more focused on the unidimensionality of a latent variable (Hambleton & Jones, 1993; Kean et al., 2017). Item-level parameters allow for each item to evaluate different positions/intensities of the latent construct. Item-level parameters also aid in the development of more valid and precise measures and estimates of reliability, and more comparable results from respondents whom have taken more varied amounts and types of items (Kean et al., 2017). Parameter invariance concerns the inferential generalizability of the person measures and item calibrations across populations, time, or contexts (Rupp & Zumbo, 2006; Smith, 2004b). In other words, parameter invariance indicates the identical nature of the parameter values in different populations or across different conditions. The parameter values are therefore not influenced by the population or the condition in which the measure was taken, if parameter invariance is achieved. This is evaluated through different calibration samples (Rupp & Zumbo, 2006).

2.5. Rasch Measurement Theory (RMT)

2.5.1. RMT and IRT

Rasch Measurement Theory (RMT) – named after its developer, Georg Rasch (1960) – operates on the 1PL level of IRT. Because of this, Rasch is synonymous with 1PL IRT (Finch et al., 2016). However, the two theories developed in separate areas of the world around the same time (Lord & Novick, 1968; Rasch, 1960). However, a key difference between the two psychometric paradigms is that while IRT requires the model to fit the data,

RMT prescribes a model to which the data must fit, with pseudo-chance and item discrimination parameters built into the analysis (Petrillo, Cano, McLeod & Coon, 2015). Boone et al. (2014, p. 220) provide an adequate summary of the function of the Rasch Model, stating that "the Rasch model is a definition of measurement. If persons and/or items do not fit the model, then those items and/or persons are not contributing to useful measurement."

The basis of IRT and RMT instruments is the transformation of item-person probabilities into log odds units (or *logits*), which mathematically places the item and the person statistics on an interval-scale continuum. However, unlike IRT, RMT-based instruments are calibrated using Rasch analysis techniques to fit a strict model. Building questionnaires with data that fit this model permits the opportunity to revise and reduce the test items that have statistically shown to fit a model which Wright and Mok (2004) state is the only model able to satisfy the five model requirements of measurement. These requirements, as stated by Wright and Mok (2004, p. 4), are that the "measurement model must: (a) produce linear measures, (b) overcome missing data, (c) give estimates of precision, (d) have devices for detecting misfitting items/persons, and (e) the parameters of the object being measured and of the measurement instrument must be separable." The items which fit the Rasch model are identified through specific analyses in RMT that then call for the researcher/developer to re-evaluate the items to include in the model (Bond et al., 2020; Wright and Mok, 2004).

To satisfy the requirements of measurement from a Rasch perspective and psychometric evaluation, the analyses to be conducted in this study are therefore the fit to the Rasch model, differential item functioning, and internal reliability, which will be detailed below and in the methodology section. Importantly, these areas of analysis should be conducted on each trait being measured to meet the requirement of unidimensionality, if an instrument measures more than one trait.

2.5.2. Rasch Model Fit

Fundamentally, Rasch analyses involve the fit of the data for each item as well as the entire test to the Rasch model (Heffernan, Maidment, Barry & Ferguson, 2019). The fit of data to a model is emphasized by Wright & Mok (2004), stressing that while most social scientists speak of finding a model to fit the collection of data, in scale construction the opposite is in fact true. That is, the researchers should produce data adequate for measurement models, to which it is mentioned on numerous occasions by Boone et al. (2014)

that Rasch is accepted as the *definition of measurement*. It is therefore important to monitor the process in which one fits the data to the measurement model; a feat provided by Rasch analysis techniques (Wright & Mok, 2004).

The Rasch technique for monitoring model fit is the fit analysis, and is indicated by the fit residual. The fit residual is described by Heffernan et al. (2018, p. 331) as "a standardized summation of the differences between observed and predicted scores for an item." Elaboration of the statistics required to fit the Rasch model will be discussed in the methodology chapter; however, two key aspects are evaluated when reviewing the fit to the Rasch model. These are person fit and item fit. Within person and item fit are evaluations of infit and outfit. Both infit and outfit provide the same conceptual information and are based on chi-square fit statistics, although infit statistics tend to be more weighted towards expected performance, whereas outfit is not weighted, making outfit more susceptible to outliers. For this reason, Bond et al. (2020) suggest that the infit statistics be more relied upon. However, Boone et al. (2014) recommend that the outfit statistics – and more specifically the outfit mean square – be examined first, as they recognize this to be the recommendation found in the Winsteps manual (Linacre, 2012).

In RMT, fit describes the adequacy to which the data conforms to the Rasch model (Boone et al., 2014). Person and item data may therefore fit or misfit the Rasch model. Person and/or items found to not adequately fit the model are either reviewed to understand their misfit or are discarded from the assessment model.

Person fit occurs when the respondents perform as expected in the assessment. Using an ability test as an example, a person who was able to answer the more difficult items correctly should answer the easier items correctly on the ability test. Respondents who conform to this expectation would have *fit* values suggesting adequate fit to the model. In contrast, a respondent who answers the difficult items correctly yet responds incorrectly to easier items conflicts with the model and would therefore be flagged as a misfitting person. Similarly, a respondent who correctly answers easier items and incorrectly answers more difficult items would be expected of a respondent with a lower ability estimate and is therefore *fitting* the model. Likewise, a respondent whose ability estimate is low yet answers a difficult item correctly would be flagged as a *misfit* and should be examined. Upon review, this respondent may have guessed the correct item, although this may just be one explanation. According to Wright and Masters (1982) assessing the person fit of an instrument can provide

information regarding the validity of the instrument. Wright and Stone (1979) state that it is reasonable to deem misfitting person data as unsuitable to determine either person measures or for use in item calibration.

Similar principles are applied to item fit. Once again, misfitting items should be examined to determine why they may be behaving the way they are. This may bring about an important issue when looking into measures applied across groups of respondents. This issue is referred to as *differential item functioning* (DIF) and occurs when items do not function equally across subgroups such as gender or ethnicity (American Educational Research Association, [AERA], 2014).

2.5.3. Differential Item Functioning (DIF)

AERA (2014) classifies Differential Item Functioning (DIF) as a form of validity, more specifically, validity with evidence based on internal structure; while Boone et al. (2014) mention DIF analysis using Rasch techniques to be a form of construct validity. A valid psychometric assessment is one which measures the construct it aims to measure and does so well, and is one of two key properties of psychometrics (Roodt & de Kock, 2018b).

DIF, a type of item bias, exists when disparate groups, such as gender, have a differing likelihood of endorsing (or succeeding on) an item despite possessing an indistinguishable level of the underlying trait (Zumbo, 1999; Heffernan et al., 2019). Boone et al. (2014) elaborate that DIF is not simply the task of finding items that have produced different measures across groups, but rather the evaluation of whether items define the measurement scale in the exact manner for different groups. The latter explanation of DIF is clarified by Boone et al. (2014) using the analogy of a meterstick to measure the heights of males and females. In their example, on average, males are taller than females. The difference in this finding is controlled by the fact that the meterstick does not change how it measures across gender, which therefore exhibits no change in the definition of height between males and females, and therefore does not elicit DIF. Psychological assessments should perform like the meterstick in that differences between groups should be explained by real differences and not due to the items defining the measure differently across groups.

Although a DIF analysis can be conducted using CTT techniques such as an analysis of variance for each item between groups (Heffernan et al., 2019), Rasch analysis, unlike CTT analyses, (a) conduct a DIF analysis using data transformed to interval-level, and (b) calculate item difficulty estimates (which are measured at the interval level) for the

subgroups in question. In the latter point, when item difficulty estimates vary between the subgroups there is evidence that DIF may be present (Bond et al., 2020). Furthermore, while DIF has various methods of being analyzed, a programme for conducting Rasch analyses such as Winsteps (Linacre, 2020) can produce graphic representations of potential DIF. The first is a hierarchical representation (Wright map), where the *location* of the item (its difficulty/endorsability) differs between groups. The second is called a *common item link plot,* which provides a graphic representation of which items, when comparing two groups, may present with DIF (Bond et al., 2020).

2.5.4. Internal reliability in Rasch measurement

Further to validity, another key psychometric property is reliability, which refers to the consistency to which an assessment measures what it proposes to measure (Roodt & de Kock, 2018a). Typical to CTT, tests for the internal consistency reliability of a psychometric assessment are, familiarly, Cronbach's alpha and Kuder-Richardson Formula-20 (KR-20). Boone et al. (2014) reiterate the shortfall of using these common methods of testing internal consistency, as these make use of raw, nonlinear, and test-dependent data. Instead, for inferences to be made beyond the limits of the test, Rasch reliability indices tend to be more conservative and less misleading (Linacre, as cited in Boone et al., 2014). For this, two reliability indices are provided through Rasch analysis: Person reliability, and item reliability. There exists two values of reliability in Rasch analysis (person and item) since there exists two samples, the persons answering the assessment and the set of items included in the assessment (Bond et al., 2020). Already this differs to conventional tests of internal consistency, providing only estimates of test reliability within the sample (Boone et al., 2014).

Further to the calculation of person and item reliabilities, Rasch analysis provides person and item separation statistics, to which Boone et al. (2014) report to be important additions to the instrument evaluation process. They note that the separation values are a "signal-to-noise ratio in the data" (p. 222) and, without providing a complex explanation, indicate that separation values theoretically range from zero to infinity, where higher values are preferred. This preference may be found in the formula for separation where true person variance is the numerator and error variance is the denominator. This suggests that higher separation values are marked by higher person variance than error variance, and therefore greater differences in person scores than error scores. This is confirmed in Boone et al.

(2014) where the authors reveal that lower person separation indices suggest that the instrument may not be sensitive enough to separate higher performers from lower performers (or traits); and that low item separation indices indicate that the person sample may not be "large enough to confirm the item difficulty hierarchy of the instrument" (p. 227). Much like the conventional procedures for evaluating internal consistency when refining instruments, the reliability and separation indices provided from Rach analysis can also be used to compare an instrument in its iterative forms (Boone et al., 2014).

2.6. Chapter summary

The chapter began with a brief introduction into the field of psychometry, discussing the growing use of psychometric assessments which necessitates the precision of these tools. This was followed by a brief history into the use and misuse of psychological assessments in the South African context. This portion of the chapter focussed on how psychometric tests were not ensured to be culturally adequate for a large period of time, although research into the topic was mounting. Finally, in 1998, it was mandated that the use of psychometric tools is prohibited unless there is scientific evidence to show that the tool is valid, reliable, and culturally fair and unbiased. It is then necessary to investigate the psychometric properties of HPTi, which is an assessment developed in the UK, in the South African context, a study not yet conducted, to evaluate the reliability, validity, and bias.

Three psychometric theories are then discussed: CTT, IRT, and RMT. Classical Test Theory (CTT), the most widely adopted theory to base psychometric instrument development, is then detailed with the advantages and disadvantages of the Theory. Item Response Theory (IRT) follows the discussion on CTT, highlighting the stronger assumptions and computations involved in the probabilistic-model of IRT. One such computation is the transformation of ordinal-level data, that CTT produces at best, to interval-level data, which satisfies an assumption of parametric statistical procedures.

However, IRT techniques have three parameter logistic models – 1PL, 2PL, and 3PL – to which the research, developer, or analyst is required to choose from to fit the data. Rasch Measurement Theory (RMT), on the other hand, illustrates that data should be fitting a model that best produce linear measures, overcome missing data, provide estimates of precision, contain devices for detecting the misfit of items or persons, and parameter-separation; to which the Rasch model has been deemed the definition of such measurement. Therefore, making use of the 1PL IRT-model, data fitting the Rasch model achieves person and item

measures fitting what is currently referred to as the definition of measurement. Rasch analysis techniques determine the fit to this model through Rasch Model Fit indices, and is argued to provide better estimates for reliability through person and item reliability and separation indices instead of Cronbach's alpha and Kuder-Richardson formulae, as well as better estimates for the evaluation of differential item functioning – a type of bias which can be used in evaluating the construct validity of a psychometric instrument.

Chapter 3:
Methodology

3.1. Chapter introduction

The purpose of the study is to analyse the psychometric properties of the High Potential Trait Indicator (HPTi) from a Rasch analysis approach. The study is therefore quantitative, as the researcher aims to investigate the psychometric properties of the HPTi by analysing the personality questionnaire using Rasch analysis techniques. This would include the reliability and validity of the assessment once analysed using Rasch analysis techniques.

This chapter talks through the various steps taken to conduct the Rasch analyses, beginning with the way in which the data was obtained, followed by an explanation of the HPTi instrument and the psychometric properties ascertained during its development, before discussing the steps towards conducting the relevant Rasch analysis techniques.

3.2. Data collection

The copyright holder of the relevant HPTi data – Thomas International Ltd. – had been contacted to request the use of their data. Thomas International agreed to the study, and provided the researcher with secondary data. The dataset contains completed HPTi data in raw form.

To remove the potential for individuals in the dataset from being identifiable, data of this nature was removed from the dataset before being made available to the researcher. An anonymous identifier had been provided per respondent to assist in the preparation of the data, such as removing necessary datapoints of respondents who had completed the assessment on more than one occasion.

The dataset contains demographic information such as gender, ethnicity, home language, year of birth, and provincial location at the time of completion, as well as additional work-related information such as employment status, industry sector, job function, managerial level, and highest level of education. Respondents who had not provided responses to the demographic questions were removed from the dataset. This deletion was performed to ensure that all the respondents included in the analysis have points of reference to the general population and can be included in the comparative analyses such as the differential item functioning analysis. The date of completion for each respondent was not provided. For this reason, coupled with the absence of respondents' age in the dataset, a

descriptive analysis of age at the time of completion cannot be conducted, and will be substituted by an analysis of their birth year.

The dataset includes the raw data, with scores from 1 (Strongly disagree) to 7 (Strongly agree) of individuals who have completed the HPTi and have indicated their voluntary participation in further research. Respondents have the opportunity to indicate their voluntary participation in further research once they have completed the assessment, and after the initial voluntary request to complete the Thomas International research questions. It is at the end of these research-related questions where individuals can indicate their willingness to participate in further research. Only data of respondents who completed the HPTi through the South African division of the organisation, and had indicated their voluntary participation in further research was obtained.

Negatively phrased items were reverse scored and subsequently suffixed with the letter 'r'. For example, item number 3 (named, 'q3' in the dataset) required reverse scoring, and was subsequently renamed to 'q3r'. Since the content of the items were not provided, Thomas International supplied the items that required reverse scoring.

According to their website (www.thomas.co) Thomas International Ltd. is a global provider of psychometric assessments. The assessment is used within the organisation, through third-parties such as other organisations trained to use their assessments, and directly upon request from an individual. Their assessments are targeted at, but not limited to, processes involving recruitment, retention and development (thomas.co). The organisation's privacy notice (2019) explains that they collect and process data for research purposes. The additional data they collect for research purposes is strictly voluntary and does not impact the initial purpose of persons completing their assessments; this data will only be used for research purposes.

3.3. The instrument

3.3.1. Background of the High Potential Trait Indicator

The HPTi is a self-reporting online questionnaire, which measures six personality traits: *Conscientiousness*, *Adjustment*, *Curiosity*, *Risk Approach*, *Ambiguity Acceptance*, and *Competitiveness* (MacRae & Furnham, 2016). The instrument comprises of 78 items – 13 per trait – of which respondents are required to rate from *Strongly disagree (1)* to *Strongly agree*

(7) per item. The items of the HPTi are phrased in the first-person. Examples of items were not provided due to copyright restrictions.

Reliability and validity studies indicated that the HPTi obtained sufficient internal consistency in its current form with thousands of working professionals across 50 countries; acceptable structural validity using structural equation modelling (SEM); significantly correlated with some aspects of the Big Five Factor Model measuring similar traits; significantly correlated with various other aspects of certain assessments such as the Hogan Development Survey (HDS), Trait Emotional Intelligence Questionnaire (TEIQue); and demonstrated sufficient predictive validity (MacRae & Furnham, 2016).

3.3.2. Definitions of the six traits

According to MacRae and Furnham (2020), *Conscientiousness* is a higher order personality trait in the Five Factor Model. It is comprised of industriousness, self-control, responsibility, order, traditionalism, and virtue. This trait has been found to have a moderate correlation with job success and other job metrics (Barrick et al, 2001; MacRae & Furnham).

Adjustment is described by MacRae and Furnham (2020) as emotional resilience to stress and positive affect, and is the inverse of trait neuroticism in the Five Factor Model. Higher levels of *Adjustment* were found to be associated with better teamwork and higher performance, while lower levels were associated with low job satisfaction and subjective well-being (Judge & Locke, 1992; MacRae & Furnham, 2020).

Curiosity is synonymous with the openness trait of the Five Factor Model. The trait is characterised as being open to new ideas and experiences, as well as being creative, reflective, innovative (MacRae & Furnham, 2020). *Curiosity* and openness were associated with job satisfaction, trainability and learning outcomes (Barrick et al, 2001; Judge et al, 1999; Linden et al, 2010; MacRae & Furnham).

Risk approach is defined as how an individual handles challenging, difficult, or threatening situations (MacRae & Furnham, 2016). It is the mitigation of negative, threat-based emotions that cause a strong drive to avoid that situation, restricting the potential range of responses to avoidance (MacRae & Furnham, 2020).

Ambiguity acceptance is a measure of how an individual perceives and processes unfamiliarity and that which is not clear (MacRae & Furnham, 2020). Herman et al (2010) suggest that tolerance for ambiguity involves unfamiliarity, change, challenging perspectives,

and valuing diversity. High-fliers and senior leadership are thought to require a tolerance for, and adaption to ambiguity due to the need to make sense of and incorporate multiple streams of mixed information to make effective decisions (Keenan & McBain, 1979; MacRae & Furnham, 2020; McCall, 1997).

Finally, MacRae and Furnham (2020) describe *Competitiveness* as a dimension that drives self-improvement and the desire for success. In a study of sales performance, Wang and Netemeyer (2002) found competitiveness to be a significant predictor of performance.

3.3.3. Internal consistency reliability

Reliability in psychometric assessments can be defined as how consistent a psychometric instrument is for each respondent, and therefore how resistant the instrument is to random error (Coaley, 2010; Fields, 2009; Pallant, 2020; Price, 2017). Internal consistency reliability is the relationship items have with each other, and is typically measured in personality assessments by Cronbach's alpha (α), a correlate of the items amongst themselves. Sources indicate an adequate internal consistency reliability obtains at least α = .70, with an ideal range between .70 and .90 (DeVillis, 2017; Kaplan & Saccuzzo, 2017; MacRae & Furnham, 2016; Nunnally, 1978; Pallant, 2020; Yang & Green, 2011).

According to MacRae and Furnham (2016), the HPTi, overall, obtained an alpha of .87 in its final revision during development, indicating adequate internal consistency reliability. Furthermore, each individual trait measured achieved sufficient internal reliability: *Conscientiousness* (α = .77), *Adjustment* (α = .78), *Curiosity* (α = .80), *Risk Approach* (then *Courage*, α = .72), *Ambiguity Acceptance* (α = .74), *Competitiveness* (α = .74) (MacRae & Furnham, 2016, 2020).

3.3.4. Correlational studies

In comparison to the Five Factor Model (FFM), three HPTi traits, theoretically similar to the three FFM traits, were found to display convergent validity. HPTi *Adjustment* had a significantly strong negative correlation with FFM *Neuroticism* (r = -.64, p < .001). According to MacRae and Furnham (2016), low levels on *Adjustment* indicate more experiences of negative emotions. FFM *Neuroticism* measures more experiences of negative emotions on the higher end of the scale (Rothmann & Coetzer, 2003). Therefore, the negative correlation between HPTi *Adjustment* and FFM *Neuroticism* indicates that more experiences of negative emotions on one assessment correlate with a greater experience of negative

emotions on the other. *Curiosity* on the HPTi had a significant moderate positive correlation with FFM *Openness* (r = .46, p < .001). Finally, *Conscientiousness* on the HPTi showed significant strong positive correlations with FFM *Conscientiousness* (r = .70, p < .001).

Correlations amongst other constructs were found between the HPTi and FFM measures ranging from small to large. HPTi *Risk approach* (then called *Courage*) significantly correlated negatively with FFM *Neuroticism* (r = -.52, p < .001), positively with FFM *Extraversion* (r = .40, p < .001), and positively with FFM *Conscientiousness* (r = .38, p < .001). *Competitiveness* had significant but relatively low correlations across FFM traits, negatively with FFM *Neuroticism* (r = -.23, p < .001), positively with FFM *Extraversion* (r = .26, p < .001), and negatively with FFM *Agreeableness* (r = -.22, p < .001).

HPTi *Ambiguity Acceptance* ranged from small to moderate significant correlations with FFM traits, negatively with FFM *Neuroticism* (r = -.46, p < .001), positively with FFM *Extraversion* (r = .22, p < .001), positively with FFM *Openness* (r = .27, p < .001), and positively with FFM *Agreeableness* (r = .13, p < .05).

Overall emotional intelligence, as measured on the Trait Emotional Intelligence Questionnaire (TEIQue) on 244 UK Armed Forces, had a significant large correlation with HPTi *Adjustment* (r = .72, p < .001), positive moderate correlation with HPTi *Curiosity* (r = .38, p < .001), positive moderate correlation with HPTi *Ambiguity acceptance* (r = .34, p < .001), positive large correlation with HPTi *Conscientiousness* (r = .59, p < .001), positive large correlation with HPTi *Risk approa*ch (then called *Courage*) (r = .58, p < .001), and positive small correlation with HPTi *Competitiveness* (r = .14, p < .001).

All HPTi traits were found to have some predictive validity when correlated with a Work success measure (N = 244, UK Armed Forces) (MacRae & Furnham, 2016): HPTi *Adjustment* (r = .30, p < .001), HPTi *Risk approach* (then named *Courage*) (r = .34, p < .001), HPTi *Competitiveness* (r = .26, p < .001), HPTi *Ambiguity acceptance* (r = .21, p < .001), HPTi *Curiosity* (r = .25, p < .001), and HPTi *Conscientiousness* (r = .49, p < .001).

3.4. Data analysis

The primary data analyses were conducted using Winsteps (2020) using the Rasch rating scale model. The descriptive statistics were calculated through a combination of Microsoft Excel and Winsteps.

3.4.1. Person fit analysis

Rasch analysis provides two fit statistics for persons and items: Infit and Outfit statistics. Winsteps (2020) provides several Infit and Outfit statistics, namely mean-square (MNSQ), which is a chi-square calculation, and z-standardized (ZSTD) which is a t-statistic (Boone et al, 2014). While there may exist some debate about which statistic to evaluate first, the researcher worked with Boone et al.'s recommendation that person fit should be evaluated on outfit statistics, and Bond et al.'s (2020) recommendation that item fit is evaluated using the infit statistics. The reasons for the use of outfit statistics for person fit is explained in this section, while the use of infit statistics for item fit is elaborated on in the Item Fit section of this chapter.

3.4.1.1. Fit statistics

Wright and Linacre (1994) suggest that for high-stakes multiple-choice questionnaires a reasonable MNSQ range between 0.80 and 1.20 should be considered. However, Boone et al. (2014) indicate that a MNSQ range between 0.50 and 1.50 is acceptable, while Bond et al. (2021) provide a MNSQ range of 0.75 and 1.30. Given the various suggested ranges and that Bond et al.'s (2020) range lies between that of Wright and Linacre (1994) and Boone et al. (2014), for the purposes of this study, Bond et al.'s recommended fit statistics range of 0.75 and 1.30 were used. These values can be interpreted as follows: Values above a mean-square statistic of 1.30 is more than 30% more varied in the observed data than in the expected Rasch model and is therefore underfitting the model, and values below 0.75 is more than 25% less varied in the observed data than in the expected Rasch model and therefore overfitting the model (Bond et al., 2020). Should the MNSQ statistics lie outside the recommended range, the ZSTD is then consulted. Bond et al. (2020) indicate that ZSTD statistics greater than the absolute value of 2 is considered statistically significantly misfitting. In other words, where a mean-square statistic is equal to 1.75 with a z-standardized statistic equal to 2.00, the observed data is 75% more varied than the expected data and which did not occur as a result of chance (i.e., statistically significant).

3.4.1.2 Outfit for person fit

The choice of using the outfit statistics for person fit is encouraged by the notion that outfit statistics have more power over infit statistics when analysing person fit (Smith, 2004a), and that outfit statistics are sensitive to unexpected observations (Bond et al., 2020; Boone et al., 2014; Smith, 2004a). According to Ziljstra, van der Ark and Sijtsma (2011),

unexpected observations, such as outliers, can result in a bias and negatively impact statistical results including, but not limited to, Cronbach's alpha and the validity coefficient, justifying their exclusion, and therefore warranting the use of the outlier-sensitive fit statistic for person fit.

According to Curtis (2004) and Mousavi and Cui (2020), addressing misfitting person data is a topic that seems to be under looked. From their accounts, Curtis, and Mousavi and Cui indicate that item fit statistics are mostly scrutinized; yet from their studies, Curtis, and Mousavi and Cui found that including misfitting persons can invalidate item parameters, and have therefore recommended removing misfitting persons. Wright and Stone (1979) confirm this in their statement that misfitting persons can be deemed unsuitable for measurement and item calibration, while Wright and Masters (1982) assert that assessing the person fit of an instrument can provide information regarding the validity of the instrument. In Phillips' (1986) study, it was mentioned that the removal of misfitting persons eliminates respondents with anomalous responses. Deletion of misfitting persons also improved the fit to the Rasch model in Phillips' study.

The extent to which misfitting persons should be removed, however, is unclear. Therefore, in this study, datapoints found to contain misfitting persons in the initial analysis for each HPTi trait were removed based on the outfit statistics: $0.75 \geq MNSQ \geq 1.30$ and $ZSTD \geq |2|$. The sample of each HPTi scale following the removal of misfitting persons will be referred to as the first iteration, and will be the sample used for the primary analyses. For comparative purposes, the descriptive statistics reliability indices are provided for the initial sample and the sample following the removal of misfitting persons per trait. The sample of the 1st iteration, following the removal of misfitting persons, is also described in the preliminary analysis section, adjacent to the initial sample (see **Table 1**).

3.4.2. Descriptive statistical analysis

Descriptive statistics were examined using Microsoft Excel to outline the trends in the responses of the sample group, and demographic information of the dataset. The demographic information was described as frequencies of the initial and iterated sample for each HPTi scale. The frequencies are presented as the number of observations and the respective percentage of the sample (see Table 1) and were conducted for gender, ethnicity, home language, and provincial location.

The age demographic could not be analysed because data for the variable were not provided. Respondents' year of birth was provided, however, the date at which they completed the HPTi was not, therefore, age of the respondents could not be analysed. Instead, the year of birth of the respondents was described as the year of birth in the middle of the distribution (median), the most frequent year of birth in the sample (mode), the youngest and oldest participants in terms of their year of birth, and the range of years from the youngest to the oldest (see Table 2).

Following the demographic statistics, the raw response trends were described for each of the HPTi traits for the initial and iterated sample. The distribution of each trait was analysed using the arithmetic mean, standard deviation (Std. Dev.), median, minimum (Min.), maximum (Max.), skewness and kurtosis (see Table 3). Similarly, in Table 4, the distribution of the responses of the logit scores of each trait were outlined using the same method as the raw response trend. The raw responses are the scores garnered directly from scale of the instrument, whereas logit scores are the log-odd unit scores calculated from IRT analyses Bond et al, 2020; Boone et al, 2014).

The data then underwent the formal testing of Rasch analysis using Winsteps (Linacre, 2020), a statistical programme dedicated to conducting Rasch analysis techniques. During this analysis stage, the key psychometric properties were assessed: Item fit, Person/item reliability and separation, and differential item functioning (Boone et al., 2014).

3.4.3. Item Fit

Following the removal of misfitting persons for each HPTi trait, item fit statistics were estimated per trait. All items were tabulated for each HPTi trait with their respective fit interpretation: Fit, Overfit or Underfit. Misfitting items are based on the infit statistics where $0.75 \geq \text{MNSQ} \geq 1.30$ and $\text{ZSTD} \geq |2|$. The researcher encourages those responsible for the development of the HPTi to investigate the misfitting items further for the potential causes of misfit.

According to Bond et al. (2020), infit statistics tend to be the focus for item-fit evaluation. Bond et al. refer to infit as the "information-weighted indicator of misfit" (p.56). This is due to the infit statistics' higher sensitivity to performances where persons are located closer to the item difficulties, as opposed to the outfit statistics higher sensitivity to outlier performances – i.e. persons are located further from the item difficulty because, for example, the respondent with a high ability answered an item with a low difficult incorrectly. Bond et

al. argue that, while outfit statistics for item fit should not be ignored, deviant infit statistics are more cause for concern than outfit. For this reason, infit statistics will be employed for the evaluation of the item fit of the HPTi.

3.4.4. Unidimensionality and local dependence

Unidimensionality and local independence are two interrelated conditions required for Rasch measurement (Fan & Bond, 2019; Heffernan et al., 2019).

Unidimensionality refers to the measurement of a single construct (or latent trait, or dimension). For example, the trait extraversion is a single construct. A scale measuring extraversion alone is therefore unidimensional. A scale that measures extraversion and anxiety, then, is multidimensional (measuring more than one dimension), and not unidimensional. In Rasch measurement, the requirement, then, is that each latent trait be measured one at a time (Fan & Bond) and is therefore unidimensional. Perfect unidimensionality, however, is not a realistic expectation. Instruments are then required to have a close approximation to unidimensionality. The unidimensionality of an instrument is estimated through a principal component analysis of the residuals (PCAR) (Fan & Bond), available in software such as Winsteps (Linacre, 2020). From the PCAR, contrasts with eigenvalues at or greater than two indicate the possibility of the scale possessing more than one dimension, whereas eigenvalues less than two are regarded as insignificant (Bond et al.). Furthermore, items underfitting the Rasch model provide additional concerns to the threat of unidimensionality (Fan & Bond).

Local independence is the condition in which an individual's responses to an item is not affected by their response to any other items (Fan & Bond). For example, an item regarding reading several times a week would affect, or be affected by, an item regarding reading once a week. In such a case, the items are dependent rather than independent, and therefore violate the condition of local independence. However, like unidimensionality, it is unrealistic to expect perfect independence. An estimate of the correlation between item residuals is therefore required to determine whether there are items that are significantly dependent to each other (Fan & Bond, 2019; Linacre, 2020). According to Linacre, positive correlations of .7 is the beginning of concern for dependency. Furthermore, items overfitting the Rasch model provide additional concerns to the threat of local independence (Fan & Bond).

Unidimensionality of each HPTi scale will therefore be evaluated through PCAR in Winsteps. From which, scales which demonstrate contrasts with eigenvalues greater than or equal to two will be considered in violation of unidimensionality. Following the flagging of such scales, items that load highly, in terms of the PCAR, and found to underfit the model may then be flagged as causing a threat to the unidimensionality of the scale.

Local independence of each HPTi scale will also be evaluated through Winsteps, which provides a correlation between item residuals. From these estimates, positive correlations between items of .7 or higher, and presenting as overfitting the model, will be flagged as potentially threatening item local independence.

3.4.5. Person and item reliability and separation

Person and item reliability of the HPTi was analysed through Winsteps (Linacre, 2020). Boone et al. (2014) indicate that the person/item reliability statistics produced by Winsteps are interpreted similar to traditional internal reliability indices. Therefore, using conventional guidelines a person/item reliability index of .70 or higher is considered acceptable (DeVillis, 2017; Kaplan & Saccuzzo, 2017; Nunnally, 1978; MacRae & Furnham, 2016; Pallant, 2020; Yang & Green, 2011). However, Rasch reliability statistics may be preferred over a statistic such as Cronbach's alpha, as alpha makes use of raw, nonlinear, and test-dependent data, whereas Rasch reliability statistics make use of transformed, linear, parameter-invariant data (Boone et al., 2014). Winsteps provides both Rasch reliability estimates and Cronbach's alpha, both of which will be provided for comparative purposes. To remain consistent with MacRae and Furnham's (2016) report, a reliability coefficient of .70 will be considered adequate for this study.

Separation indices, indicating the number of distinct strata of persons and items, are noted to be acceptable when estimated to be at least 1.50 (Duncan, Bode, Lai, & Perera, 2003; Tennant & Conaghan, 2007). Duncan et al. elaborate that a person separation index of 2.00 is regarded as a good reliability index, and 3.00 to be excellent.

The HPTi was therefore evaluated on the reliability indices of each trait by examining whether the person and item reliability indices produced in Winsteps (2020) of each trait reaches .70; and whether the person and item separation indices produced in Winsteps (2020) reach a minimum of 1.50 per trait. As mentioned in the Fit section, the reliability will be produced twice: Once, before removing misfitting persons, and again after the misfitting persons are removed – as determined above.

3.4.6. Differential item functioning

Differential item functioning (DIF) is mentioned in the literature section to be an evaluation of how congruently the items of a measure define a construct between certain groups (Boone, et al., 2014), and was examined using Winsteps (Linacre, 2020).

The Winsteps user-guide (Linacre, 2020) states that a DIF analysis is conducted through Winsteps Table 30. In this table, average measures of the relevant groups (e.g. males and females) are presented in logits, and are compared against each other. Boone et al. (2014) indicate that DIF may be present in comparisons with a significant p-value ($p < .05$) of the Rasch-Welch statistic. Linacre (2020) substantiates that in addition to a significant difference between groups, an effect size equal to or greater than .64 is considered moderate to large, whereas between .43 and .64 is considered slight to moderate. Below .43 is considered negligible and insufficient to flag items as having DIF present. A DIF effect size in a Rasch analysis is provided in Winsteps Table 30.1 as 'DIF contrast'.

DIF was therefore analysed across the items of each HPTi trait and between the relevant subgroups by means of a significant difference in the Rasch-Welch statistic and a sufficiently large DIF contrast. A p-value less than .05 in Winsteps Table 30.1 indicated a significant difference, and an effect size of at least .43 indicated by DIF contrast in Winsteps Table 30.1. These items are highlighted as recommended for further examination, as the examination for the reasons for DIF is beyond the aim of this study, and cannot be investigated further since the researcher does not have the content of the HPTi items.

The subgroups for the DIF analysis are gender, ethnicity and language. However, due to the differences in sample sizes between the European languages (English and Afrikaans) and African languages (Sesotho, isiXhosa, isiZulu, Setswana, Sepedi, Xitsonga, isiNdebele, and siSwati), the African languages were collapsed to form the group 'African Languages'. The resultant language groups are thus African languages (n = 326, 26%), Afrikaans (n = 370, 29%) and English (n = 558, 44%) This is not intended to reduce the differences in the African languages, and may warrant further investigation with larger samples in the individual African languages.

3.5. Chapter summary

This chapter outlines the methodology of the study. In this chapter, it is presented that the study is quantitative in nature, as it incorporates statistical procedures from the primarily quantitative philosophy of Rasch measurement theory. Data for this study is secondary data obtained from Thomas International Ltd., and includes individuals who completed the HPTi through the South African division who have indicated their voluntary consent to have their HPTi data used in further studies. The HPTi is a personality questionnaire comprised of six subscales, and has been found to be valid and reliable in its development phase in the UK. The dataset contains the variables: unique identifier, HPTi raw data, year of birth, gender, ethnicity, home language, provincial location, highest educational qualification, employment status, managerial status, industry sector working in and current job function.

The method for the data analysis is as follows: First, a person fit analysis was conducted to identify persons who responded in a manner that does not contribute to the remainder of the study. Misfitting persons are identified where outfit MNSQ is greater than or equal to 1.30 or less than or equal to 0.75, with a ZSTD greater than the absolute value of 2. These persons are then removed from the sample.

Second, the initial sample is described, with the resultant sample following the removal of misfitting persons provided in tabular form for each HPTi trait. This includes the size of the sample, and the frequencies of demographics: gender, ethnicity, home language, and provincial location. Following the frequencies will be the median, mode, minimum and maximum of the years of birth of the respondents since the age of the respondents at the time of completion could not be calculated. The descriptive analysis continues by providing the means, standard deviations, median, minimum, maximum, skewness, and kurtosis of the responses for each of the six subscales in the initial sample and the iterated sample f each subscale; followed by the same analysis using person measures obtained from Winsteps instead of raw response scores.

Third, reliability estimates are analysed for each trait of the HPTi on the initial and iterated data. The reliability estimates are Cronbach’s alpha, person reliability, person separation, and item separation. From this stage, the initial dataset was no longer used in analysis.

Fourth is the item fit analysis in which the items of each iterated subscale were analysed on their fit to the Rasch model. Items with infit MNSQ greater than or equal to 1.30

and ZSTD greater than the absolute value of 2 are identified and regarded as underfitting, whereas items with infit MNSQ less than or equal to 0.75 and ZSTD greater than the absolute value of 2 are regarded as overfitting the Rasch model. All items are tabulated with their respective infit and outfit statistics.

Fifth, the unidimensionality and local independence of the HPTi scales are evaluated in Winsteps using PCAR and residual correlations.

Finally, an evaluation of differential item functioning was conducted on the iterated dataset and across the six HPTi subscales. The groups compared were gender, ethnicity, and first languages. Items in this analysis with p-value less than 0.05 and DIF contrast, as provided in Winsteps, greater than or equal to .43 were tabulated with these statistics and their comparison groups' average person measures and standard errors, as well as the interpretation of the DIF contrast (effect size). DIF contrasts of .43 to .64 are regarded as having an slight to moderate effect size, whereas DIF contrasts of .64 and higher are moderate to large effect sizes (Linacre, 2020).

Chapter 4:
Results

4.1. Chapter introduction

In this chapter, the preliminary analysis section will describe the sample of the obtained dataset. Recalling that the data had steps to remove incomplete datapoints, and that further data were removed following from a Rasch analysis technique, the sample is described from the removal of incomplete datapoints and how the resultant 'complete' dataset differs to the sample following on from the Rasch analysis technique – referred to as the 1st iteration.

Subsequent to the preliminary analysis, the primary analysis provides the results of the Rasch analysis techniques per trait of the HPTi. The structure of the primary analysis follows each trait, presenting the results of objectives 1 (Rasch model fit) and 2 (Reliability) together, then furnishing the results of objective 3 (Differential item functioning), before progressing on to objectives 1 to 3 of the next HPTi trait. This structure will begin with trait *Conscientiousness* and conclude with the *Competitiveness* scale.

The chapter will then conclude with the key findings of the analyses, to be focussed on in the final chapter, the Discussion chapter.

4.2. Preliminary analysis

The preliminary analysis section of this study outlines the descriptive statistics of the initial sample's demographic information followed by the person fit analysis in which misfitting persons are removed from the dataset per HPTi subscale. Following the removal of misfitting persons, a summary of the demographic distribution per subscale is provided in **Table 1** and **Table 2**.

Descriptive statistics of response trends of the HPTi traits are then reported for the initial sample and the iterated dataset after the misfitting persons were removed. The response trends will first be reported in the raw aggregated scores followed by the IRT person measures. The section will also present the data from the initial complete dataset and through the 1st iteration of each trait following the removal of misfitting persons in each trait.

The tables have shortened headings for the six subscales, where 'Con.' Is the *Conscientiousness* scale, 'Adj.' is the *Adjustment* scale, 'Cur.' Is the *Curiosity* scale, 'RA' is

the *Risk approach* scale, ‘AA’ is the *Ambiguity acceptance* scale, and ‘Cmp.’ Is the *Competitiveness* scale.

4.2.1. Descriptive statistics of the initial sample demographic information

The initial dataset, after removing incomplete data, consisted of 1257 respondents, of which 684 (54.4%) were female and 573 (46.6%) were male; approximating the gender distribution in the South African population – female (51.2%), male (48.8%; Statistics South Africa, 2019). Unlike the gender distribution, however, the ethnicity of the respondents did not reflect the South African ethnic distribution. **Table 1** below illustrates that the White ethnic group was the largest of the sample with 577 respondents (45.9%), followed by the Black African group with 380 (30.2%), the Coloured group (n = 184, 14.6%), Asian/Indian group (n = 106, 8.4%), and the group not otherwise specified (n = 10, 0.8%). Similar to the White ethnic group, English was the most widely reported home language (n = 558, 44.4%), followed by Afrikaans (n = 370, 29.4%), isiXhosa (n = 76, 6.0%), isiZulu (n = 59, 4.7%), Setswana (n = 49, 3.9%), Sesotho (n = 48, 3.8%), Sepedi (n = 47, 3.7%), Tshivenda (n = 20, 1.6%), Xitsonga (n = 16, 1.3%), siSwati (n = 9, 0.7%), French (n = 3, 0.2%) and isiNdebele (n = 2, 0.2%). Finally, with regards to the provincial residential location of the respondents at the time of completing the assessment, nearly half were in Gauteng (n = 586, 46.6%), followed by the Western Cape (n = 313, 24.9%), the Free State (n = 123, 9.8%), KwaZulu-Natal (n = 112, 8.9%), Eastern Cape (n = 80, 6.4%), Mpumalanga (n = 15, 1.2%), Limpopo (n = 13, 1.0%), the North West (n = 10, 0.8%), and the Northern Cape (n = 5, 0.4%).

Neither the age of the respondents nor the date of completion of the respondents were provided. However, the year of birth of the respondents were obtained. **Table 2** provides a summary of the birth years of the respondents. It is evident that the oldest respondent in the initial sample was born in 1945, whereas the youngest was born in 1999. The most frequent year of birth in the initial sample was 1985, while the median year of birth was 1982. However, unlike the median, most frequent year of birth and the most recent year of birth, the oldest year of birth changes to 1952 in the iterated *Adjustment*, *Curiosity*, *Risk Approach*, and *Competitiveness* sample; a difference of 7 years. Obtaining the iterated sample will be discussed in the next section.

4.2.2. Person fit analysis

A person fit analysis was conducted through Winsteps to identify persons who responded in a manner that does not contribute to the remainder of the study. Misfitting persons are identified where outfit MNSQ is greater than or equal to 1.30, or less than or equal to 0.75, with a ZSTD greater than the absolute value of 2. These persons are then removed from the sample to obtain six iterated samples.

For the subscale *Conscientiousness*, a total of 110 (8.8% of the initial 1257) misfitting persons were detected and subsequently removed, of which, 95 were underfitting (OUT.MSNQ ≥ 1.30) and 15 overfitting (OUT.MNSQ ≤ 0.75). Trait *Adjustment* had total of 166 (13.2%) misfitting persons were detected and removed from the trait Adjustment, of which 74 were underfitting and 56 overfitting. A total of 161 (12.8%) misfitting persons were detected and subsequently removed from the trait *Curiosity*, of which, 92 were underfitting and 69 were overfitting. *Risk Approach*, contained a total of 251 (20%) misfitting persons, of which 136 were underfitting and 115 were overfitting, and were subsequently removed from the dataset. Subscale *Ambiguity Acceptance* had a total of 224 (17.8%) persons were found to be misfitting, with 111 underfitting and 113 overfitting. Finally, trait *Competitiveness* contained a total of 224 (17.8%) misfitting persons, of which 109 were underfitting and 115 overfitting.

Table 1 below illustrates the frequencies of the demographic information for the initial sample (‘All’, n = 1257), the iterated *Conscientiousness* sample (‘Con.’, n = 1147), the iterated *Adjustment* scale (‘Adj.’, n = 1091), the iterated *Curiosity* scale (‘Cur.’, n = 1096), the iterated *Risk approach* scale (‘RA’, n = 1006), the iterated *Ambiguity acceptance* scale (‘AA’, n = 1033), and the iterated *Competitiveness* scale (‘Cmp’, n = 1033).

Table 1: Demographic Statistics of the Initial Sample, and 1st Iterations of Each Trait

	All	Con.	Adj.	Cur.	RA	AA	Cmp.
	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
Total	1257	1147	1091	1096	1006	1033	1033
Gender							
Male	573 (45.6)	518 (45.2)	493 (45.2)	505 (46.1)	465 (46.2)	472 (45.7)	473 (45.8)
Female	684 (54.4)	629 (54.8)	598 (54.8)	591 (53.9)	541 (53.8)	561 (54.3)	560 (54.2)
Ethnicity							
Black African	380 (30.2)	333 (29.0)	324 (29.7)	329 (30.0)	300 (29.8)	311 (30.1)	287 (28.8)
Coloured	184 (14.6)	168 (14.6)	161 (14.8)	161 (14.7)	145 (14.4)	144 (13.9)	149 (14.4)
Asian/Indian	106 (8.4)	95 (8.3)	93 (8.5)	91 (8.3)	83 (8.3)	89 (8.6)	83 (8.0)
White	577 (45.9)	543 (47.3)	505 (46.3)	507 (46.3)	471 (46.8)	481 (46.6)	496 (48.0)
Other	10 (0.8)	8 (0.7)	8 (0.7)	8 (0.7)	7 (0.7)	8 (0.8)	8 (0.8)

Table 1: Demographic Statistics of the Initial Sample, and 1st Iterations of Each Trait

	All	Con.	Adj.	Cur.	RA	AA	Cmp.
	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
Home language							
English	558 (44.4)	511 (44.6)	490 (44.9)	491 (44.8)	444 (44.1)	460 (44.5)	459 (44.4)
Afrikaans	370 (29.4)	347 (30.3)	317 (29.1)	321 (29.3)	303 (30.1)	302 (29.2)	319 (30.9)
Sesotho	48 (3.8)	45 (3.9)	43 (3.9)	40 (3.6)	38 (3.8)	43 (4.2)	36 (3.5)
isiXhosa	76 (6.0)	68 (5.9)	60 (5.5)	66 (6.0)	63 (6.3)	59 (5.7)	58 (5.6)
isiZulu	59 (4.7)	52 (4.5)	53 (4.9)	48 (4.4)	44 (4.4)	46 (4.5)	44 (4.3)
Setswana	49 (3.9)	44 (3.8)	45 (4.1)	46 (4.2)	38 (3.8)	42 (4.1)	38 (3.7)
Sepedi	47 (3.7)	35 (3.1)	41 (3.8)	39 (3.6)	37 (3.7)	38 (3.7)	38 (3.7)
Xitsonga	16 (1.3)	13 (1.1)	15 (1.4)	15 (1.4)	15 (1.5)	13 (1.3)	14 (1.4)
isiNdebele	2 (0.2)	2 (0.2)	1 (0.1)	1 (0.1)	2 (0.2)	0 (0.0)	1 (0.1)
siSwati	9 (0.7)	8 (0.7)	8 (0.7)	9 (0.8)	9 (0.9)	9 (0.9)	8 (0.8)
Tshivenda	20 (1.6)	19 (1.7)	17 (1.6)	19 (1.7)	12 (1.2)	19 (1.8)	16 (1.5)
French	3 (0.2)	3 (0.3)	1 (0.1)	1 (0.1)	1 (0.1)	2 (0.2)	2 (0.2)

Table 1: Demographic Statistics of the Initial Sample, and 1st Iterations of Each Trait

	All	Con.	Adj.	Cur.	RA	AA	Cmp.
	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
Provincial location							
Eastern Cape	80 (6.4)	72 (6.3)	65 (6.0)	72 (6.6)	62 (6.2)	66 (6.4)	62 (6.0)
Free State	123 (9.8)	113 (9.9)	101 (9.3)	104 (9.5)	93 (9.2)	100 (9.7)	95 (9.2)
Gauteng	586 (46.6)	536 (46.7)	515 (47.2)	509 (46.4)	466 (46.3)	481 (46.6)	494 (47.8)
KwaZulu-Natal	112 (8.9)	101 (8.8)	99 (9.1)	97 (8.9)	88 (8.7)	91 (8.8)	87 (8.4)
Limpopo	13 (1.0)	12 (1.0)	11 (1.0)	11 (1.0)	9 (0.9)	8 (0.8)	10 (1.0)
Mpumalanga	15 (1.2)	12 (1.0)	12 (1.1)	12 (1.1)	13 (1.3)	13 (1.3)	14 (1.4)
North West	10 (0.8)	8 (0.7)	9 (0.8)	10 (0.9)	7 (0.7)	7 (0.7)	8 (0.8)
Northern Cape	5 (0.4)	4 (0.3)	3 (0.3)	4 (0.9)	5 (0.5)	4 (0.4)	3 (0.3)
Western Cape	313 (24.9)	289 (25.2)	276 (25.3)	277 (25.3)	263 (26.1)	263 (25.5)	260 (25.2)

Table 2 below contains the median, mode, minimum (oldest), maximum (youngest), and range of birth years for the initial sample, the iterated *Conscientiousness* sample, the iterated *Adjustment* scale, the iterated *Curiosity* scale, the iterated *Risk approach* scale, the iterated *Ambiguity acceptance* scale, and the iterated *Competitiveness* scale.

Table 2: Descriptive Statistics of the Years of Birth of Respondents Across Each Iterated Trait

	All	Con.	Adj.	Cur.	RA	AA	Cmp.
Median	1982	1982	1982	1982	1982	1982	1982
Mode	1985	1985	1985	1985	1985	1985	1985
Oldest	1945	1945	1952	1952	1952	1945	1952
Youngest	1999	1999	1999	1999	1999	1999	1999
Range	54	54	47	47	47	54	47

4.2.3. Response trend

The response trends of the sample per HPTi trait aggregated scores are depicted in **Table 3** below. Similar to the demographics, each trait has been provided with statistics for the initial sample and the iterated samples.

The trait with the highest average is observed to be *Conscientiousness* (M = 6.00, SD = 1.31), and with the highest skewness of -1.66, suggestive of a negatively skewed distribution (Groeneveld & Meeden, 1984), although there is no separation between the mean and median; whereas the lowest average trait can be seen to be *Adjustment* (M = 2.55, SD = 1.60). After iteration, the trend remains similar with *Conscientiousness* the highest trait (M = 6.01, SD = 0.54), and *Adjustment* the lowest (M = 2.51, SD = 0.72). Both *Conscientiousness* and *Adjustment* are found to have a less pronounced variation and skewness in the iterated sample when compared to the initial sample.

Table 3: Descriptive Statistics of the Response Scores of Each HPTi Trait, Initial and Iterated

HPTi Trait	Sample	Mean	Std Dev.	Median	Min.	Max.	Skewness	Kurtosis
Conscientiousness	Initial	6.00	1.31	6.00	1.00	7.00	-1.66	2.60
	1st iteration	6.01	0.54	6.00	4.08	7.00	-0.46	-0.04
Adjustment	Initial	2.55	1.60	2.00	1.00	7.00	0.99	0.09
	1st iteration	2.51	0.72	2.46	1.00	5.31	0.48	0.05
Curiosity	Initial	5.59	1.44	6.00	1.00	7.00	-1.11	0.69
	1st iteration	5.62	0.62	5.62	3.54	7.00	-0.30	-0.03
Risk Approach	Initial	5.23	1.58	6.00	1.00	7.00	-0.89	0.01
	1st iteration	5.23	0.66	5.31	3.31	6.92	-0.22	-0.14
Ambiguity Acceptance	Initial	3.90	1.93	4.00	1.00	7.00	0.03	-1.25
	1st iteration	3.88	0.75	3.85	1.85	6.15	0.13	-0.09
Competitiveness	Initial	4.06	1.91	4.00	1.00	7.00	-0.05	-1.16
	1st iteration	4.07	0.80	4.08	1.69	6.54	0.06	-0.22

The overall trend is similar in **Table 4** below, which illustrates the same descriptive statistics as above but using person measures obtained from Rasch calculations. Where **Table 3** above used the aggregated scores to calculate the distribution of each trait, the measures below are based on the person measures (measured in logits) obtained from Winsteps (Linacre, 2020) before calculating the descriptive statistics.

From **Table 4**, *Conscientiousness* remains the trait with the highest average measure (Initial sample, $M = 1.30$, $SD = 0.77$; 1st iteration sample, $M = 1.31$, $SD = 0.79$). However, the skewness has increased and is now in the opposite direction, suggesting that the sample is now positively skewed with more respondents on the lower end of the distribution.

Similarly, *Adjustment* remains the lowest average measure (Initial sample, $M = -0.82$, $SD = 0.57$; 1st iteration sample, $M = -0.85$, $SD = 0.57$), with the skewness considered to be negatively skewed (-1.56 and -1.66, respectively), indicating that sample contains more respondents on the higher end of the distribution).

Table 4: Descriptive statistics of the Rasch measure scores of each HPTi trait

HPTi Trait		Mean	Std Dev.	Median	Min.	Max.	Skewness	Kurtosis
Conscientiousness	Initial	1.30	0.77	1.16	0.00	4.89	2.06	6.19
	1st iteration	1.31	0.79	1.09	0.00	4.89	2.05	5.89
Adjustment	Initial	-0.82	0.57	-0.76	-4.82	0.63	-1.56	6.59
	1st iteration	-0.85	0.57	-0.76	-4.82	0.63	-1.66	7.17
Curiosity	Initial	1.17	0.72	1.06	-0.33	5.47	1.65	5.76
	1st iteration	1.20	0.73	1.06	-0.33	5.47	1.73	6.00
Risk Approach	Initial	0.61	0.48	0.58	-0.46	3.82	1.43	5.12
	1st iteration	0.61	0.48	0.58	-0.35	3.82	1.62	6.22
Ambiguity Acceptance	Initial	-0.07	0.37	-0.10	-1.28	1.30	0.19	0.66
	1st iteration	-0.09	0.38	-0.10	-1.28	1.30	0.14	0.53
Competitiveness	Initial	0.06	0.40	0.07	-1.61	1.89	0.24	1.70
	1st iteration	0.07	0.39	0.07	-1.61	1.89	0.13	1.20

4.3. Primary analysis

In the primary analysis section, the results of the statistical procedures of the main objectives will be presented. Starting with *Conscientiousness* through to *Competitiveness*, the first two objectives – item fit to the Rasch model and Reliability – will be covered, followed by the third objective, Differential item functioning.

The first Rasch analysis technique conducted involved the fit of items to the Rasch model. The Person Fit sections outline the number of misfitting persons for each HPTi scale which are removed from the initial dataset to form the 1st iterated dataset of the corresponding trait. This is followed by the results of the reliability indices – Cronbach's alpha, person reliability, and person separation – presenting the reliability indices before and after the removal of the misfitting persons. Lastly, before the third objective of the study, the results of the item fit to the Rasch model are provided.

Finally in this section, the results of the differential item functioning analysis are presented. In each HPTi scale, DIF results are provided for differences between gender groups, – male and female – ethnic groups, – Black African, Asian/Indian, Coloured, and White – and first language groups, - African Languages, Afrikaans, and English.

4.3.1. Conscientiousness

4.3.1.1. Reliability

The internal consistency reliability indices were obtained through the Winsteps software. In addition to person and item reliability, Winsteps produces the traditional Cronbach's alpha.

It can be observed in **Table 5** that the Cronbach's alpha of the *Conscientiousness* scale approaches the accepted cut-off of .70 in the initial sample (α = .68). Similarly, the person reliability index also falls slightly short (person reliability = .65). Following the removal of the misfitting persons, both reliability indices achieve sufficiency (α = .71, person reliability = .70). A similar trend is observed for person separation. The initial sample falls short of the accepted 1.50 cut-off (person separation = 1.37), yet achieves sufficiency following the removal of the misfitting persons (person separation = 1.53). The *Conscientiousness* scale is therefore found to be sufficiently reliable in its internal consistency. Finally, the item reliability index of 1.00 (1.00 after person removal) and

separation of 15.08 (16.10 after person removal) for the *Conscientiousness* scale indicates that the sample of people was large enough to estimate the item parameters.

Table 5: Conscientiousness – Reliability Indices

	Initial sample	1st iteration sample	Difference
Cronbach's alpha	0.68	0.71	0.03
Person reliability	0.65	0.70	0.05
Person separation	1.37	1.53	0.16
Item separation	15.08	16.10	1.02

4.3.1.2. Unidimensionality and local independence

Through a principal components analysis of residuals (PCAR), the unidimensionality of the *Conscientiousness* subscale was examined. **Figure 1** shows the PCAR results, in which the unexplained variance in the 1st contrast is less than 2 (1st contrast = 1.68). This indicates that the *Conscientiousness* subscale of the HPTi is not measuring more than one dimension.

Figure 1: Conscientiousness PCAR

```
     Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information unit
                                      Eigenvalue   Observed   Expected
Total raw variance in observations     =      20.7828 100.0%         100.0%
  Raw variance explained by measures   =       7.7828  37.4%          42.3%
    Raw variance explained by persons  =       2.5364  12.2%          13.8%
    Raw Variance explained by items    =       5.2464  25.2%          28.5%
  Raw unexplained variance (total)     =      13.0000  62.6% 100.0%   57.7%
    Unexplned variance in 1st contrast =       1.6794   8.1%  12.9%
    Unexplned variance in 2nd contrast =       1.4861   7.2%  11.4%
    Unexplned variance in 3rd contrast =       1.3260   6.4%  10.2%
    Unexplned variance in 4th contrast =       1.2070   5.8%   9.3%
    Unexplned variance in 5th contrast =       1.0605   5.1%   8.2%
```

The largest standardized residual correlations were analysed to evaluate the local independence of the items of the scale. No item pairs were found to be above the correlation of .70, ranging from -.25 to .32. This indicates that none of the items of this scale are in violation of local independence.

4.3.1.3. Item fit

Infit statistics were examined for the items of the HPTi trait, *Conscientiousness*. A total of two of the 13 items were found to fall outside the range between 0.75 and 1.30. Error! Reference source not found. below lists the items and indicate whether they are considered fitting, underfitting, or overfitting.

Table 6: Conscientiousness – Item Fit Indices and Item Measures

Item	Measure	IN.MNSQ	IN.ZSTD	OUT.MNSQ	OUT.ZSTD	Status
q57r	-0.03	1.94	9.90	1.69	9.66	Underfit
q75r	0.70	1.25	5.47	1.28	5.47	Fit
q69	1.09	1.16	3.88	1.27	5.80	Fit
q51r	0.10	1.13	2.23	1.18	3.01	Fit
q33	0.63	1.08	1.74	1.19	3.61	Fit
q15	0.51	1.08	1.62	1.18	3.34	Fit
q3r	-0.92	1.02	0.37	0.92	-1.17	Fit
q27r	-0.07	0.98	-0.40	0.99	-0.16	Fit
q39	0.08	0.83	-3.19	0.88	-2.12	Fit
q21	-1.11	0.82	-2.72	0.63	-5.54	Fit
q45	-0.41	0.77	-3.81	0.70	-5.12	Fit
q9	-0.37	0.76	-4.08	0.74	-4.47	Fit
q63	-0.21	0.66	-6.30	0.69	-5.56	Overfit

Error! Reference source not found. above indicates that item q57r underfits the model (IN.MNSQ = 1.94, INZSTD = 9.90) and may be further supported as underfitting by the outfit statistics (OUT.MNSQ = 1.69, OUT.ZSTD = 9.66). The infit mean-squared suggests that item 57 is 94% more varied than the expectation of the Rasch model.

Items q63 was found to be overfitting the model (IN.MNSQ = 0.66, and IN.ZSTD = -6.30). This indicates that q63 is 34% less varied than the Rasch model. Similar to q57r, item q63 has supporting outfit statistics (OUT.MNSQ = 0.69, OUT.ZSTD = -5.56).

4.3.1.4. Differential item functioning

An analysis of differential item functioning of the *Conscientiousness* scale reveals that a significant difference between males and females occurred on five of the 13 items.

However, the practical significance of these differences – as measured by the DIF contrast – are far below .43 and are therefore regarded as inconsequential. **Table 7** below lists these 5 items with their corresponding statistics.

Table 7: Conscientiousness – DIF on Gender groups

	Mean (S.E.)				
Item	Females	Males	p-value	DIF contrast	Effect size
q9	-0.45 (.06)	-0.28 (.06)	.019	-.18	Negligible
q27r	0.01 (.04)	-0.18 (.05)	.001	.20	Negligible
q45	-0.52 (.06)	-0.28 (.06)	.001	-.24	Negligible
q63	-0.11 (.05)	-0.34 (.06)	.000	.23	Negligible
q69	1.05 (.03)	1.13 (.03)	.048	-.08	Negligible

DIF inspection across ethnicities found seven instances where DIF may be present, warranting further investigation into the items presented in **Table 8** below.

Table 8 depicts seven instances in which DIF may be present across ethnic groups. It is evident that of those items listed, q21 appears to have the most instances in which DIF is present between ethnicities, with four of the seven occurrences (57%). Other instances of statistically significant differences were present, however, upon the evaluation of their DIF contrasts their effect sizes were found to be negligible.

Table 8: Conscientiousness – DIF on Ethnicity groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q9	Asian/Indian	-.56 (.16)	Black African	-.11 (.07)	.001	-.45	Slight to moderate
q21	Asian/Indian	-1.66 (.26)	Coloured	-1.01 (.14)	.025	-.64	Moderate to large
q21	Asian/Indian	-1.66 (.26)	White	-.89 (.07)	.001	-.77	Moderate to large
q21	Black African	-1.52 (.13)	Coloured	-1.01 (.14)	.022	-.50	Slight to moderate
q21	Black African	-1.52 (.13)	White	-.89 (.07)	.000	-.63	Slight to moderate
q27r	Asian/Indian	.19 (.11)	Black African	-.23 (.07)	.035	.43	Slight to moderate
q45	Black African	-.66 (.09)	White	-.23 (.05)	.000	-.43	Slight to moderate

Table 9: Conscientiousness – DIF on Language groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q9	African languages	-.07 (.07)	English	-.50 (.06)	.000	.43	Slight to moderate
q9	African languages	-.07 (.07)	Afrikaans	-.47 (.07)	.000	.40	Approaching slight
q21	African languages	-1.47 (.13)	English	-1.04 (.08)	.002	.43	Slight to moderate
q21	African languages	-1.47 (.13)	Afrikaans	-.97 (.09)	.002	.50	Slight to moderate

Finally, the DIF analysis across first language groups, African languages, English, and Afrikaans was conducted for trait *Conscientiousness*. The analysis revealed differences between the African languages group and English and Afrikaans on items 9 and 21 (see **Table 9**), yet no difference between the English and Afrikaans language groups on the same items.

4.3.2. Adjustment

4.3.2.1. Reliability

The *Adjustment* measure of the HPTi achieved sufficient internal consistency reliability coefficients, measured by Cronbach's alpha, with .75 before person removal and .76 after. The Rasch measure for internal consistency reliability – person reliability – indicates that the *Adjustment* measure just misses the cut-off of .70 before person deletion (person reliability = .69), and achieves sufficient reliability post-person deletion (person reliability = .73). Expectedly then, the person separation index of trait Adjustment only just falls short of the recommended cut-off of 1.50 prior to person deletion with 1.49, and meets the requirement post-removal with 1.63.

Following person removal based on misfitting persons, discussed in the prior subsection, the *Adjustment* scale of the HPTi therefore meets the criteria to be regarded as having internal reliability.

Table 10: Adjustment – Reliability Indices

	Initial sample	1st iteration sample	Difference
Cronbach's alpha	0.75	0.76	0.01
Person reliability	0.69	0.73	0.04
Person separation	1.49	1.63	0.14
Item separation	16.39	17.37	0.98

4.3.2.2. Unidimensionality and local independence

The unidimensionality of the *Adjustment* subscale was examined. **Figure 2** shows the PCAR results, in which the unexplained variance in the 1st contrast is less than 2 (1st contrast = 1.80). This indicates that the *Adjustment* subscale of the HPTi is not measuring more than one dimension.

Figure 2: Adjustment PCAR

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	21.5738	100.0%		100.0%
Raw variance explained by measures =	8.5738	39.7%		42.6%
Raw variance explained by persons =	1.0198	4.7%		5.1%
Raw Variance explained by items =	7.5539	35.0%		37.5%
Raw unexplained variance (total) =	13.0000	60.3%	100.0%	57.4%
Unexplned variance in 1st contrast =	1.7976	8.3%	13.8%	
Unexplned variance in 2nd contrast =	1.5617	7.2%	12.0%	
Unexplned variance in 3rd contrast =	1.2172	5.6%	9.4%	
Unexplned variance in 4th contrast =	1.1185	5.2%	8.6%	

The largest standardized residual correlations were analysed to evaluate the local independence of the items of the scale. No item pairs were found to be above the correlation of .70, ranging from -.24 to .37. This indicates that none of the items of this scale are in violation of local independence.

4.3.2.3. Item fit

The outfit statistics were examined for the items of the HPTi trait, *Adjustment*. One of the 13 items were found to be fall outside the range between .75 and 1.30. **Table 11** below lists the items and indicate whether they are considered fitting, underfitting, or overfitting.

Table 11: Adjustment – Item Fit Indices and Item Measures

Item	Measure	IN.MNSQ	IN.ZSTD	OUT.MNSQ	OUT.ZSTD	Status
q31	0.38	1.43	6.84	1.28	4.30	Underfit
q37	-0.21	1.21	4.66	1.25	4.82	Fit
q67	-0.47	1.20	4.85	1.27	5.70	Fit
q55r	-0.48	1.20	4.83	1.25	5.49	Fit
q19	-0.68	1.14	3.52	1.17	3.89	Fit
q7	0.91	1.13	1.84	0.91	-1.32	Fit
q13	0.93	1.05	0.79	0.92	-1.15	Fit
q49	-0.61	0.98	-0.64	1.10	2.40	Fit
q61	0.05	0.93	-1.57	1.00	-0.03	Fit
q43	0.48	0.93	-1.22	0.81	-3.27	Fit
q25r	-0.08	0.92	-1.87	1.04	0.71	Fit
q73	0.09	0.91	-1.85	0.89	-2.21	Fit
q1	-0.30	0.71	-8.00	0.72	-6.85	Overfit

Table 11 above indicates that most items in trait *Adjustment* adequately fit the Rasch model based on Bond et al.'s (2020) criteria. Item q31, however, is found to be underfitting the model (IN.MNSQ = 1.43, IN.ZSTD = 6.84), indicating that this item 43% more varied than the Rasch model. Item q1 is found to be overfitting the model (IN.MNSQ = 0.71, IN.ZSTD = -8.00) with supporting outfit statistics (OUT.MNSQ = 0.72, OUT.ZSTD = -6.85). The infit mean-squared of item q1 suggests that this item is 29% less varied than the Rasch model expects.

4.3.2.4. Differential item functioning

An analysis of differential item functioning of the *Adjustment* trait reveals that a significant difference between males and females occurred on six of the 13 items. However, the examination of the DIF contrast suggests that the differences on these items between females and males are not large enough to infer a practical significance. **Table 12** below lists these 6 items with their corresponding statistics.

Table 12: Adjustment – DIF on Gender groups

	Mean (S.E.)				
Item	Females	Males	p-value	DIF Contrast	Effect size
q13	1.00 (.06)	0.83 (.06)	.030	.17	Negligible
q25r	0.03 (.03)	-0.21 (.04)	.000	.24	Negligible
q31	0.29 (.04)	0.49 (.05)	.008	-.20	Negligible
q43	0.55 (.05)	0.40 (.05)	.002	.16	Negligible
q49	-0.67 (.03)	-0.54 (.03)	.003	-.13	Negligible
q73	0.03 (.03)	0.16 (.04)	.038	-.13	Negligible

A DIF analysis across ethnic groups found two instances in which DIF may be present, suggesting a need for further investigation into the items presented in **Table 13** below.

The two instances in which DIF may be present across ethnic groups in trait *Adjustment* are items q7 and q49. Although both are seen to have a slight effect size, the common ethnic group appears to be the Black African group.

Finally, the investigation of DIF across the first language groups revealed that item q7 had a significant difference between the African Language group and the Afrikaans group (see **Table 14** below). This finding may be related to the difference between the Black African and White ethnic groups on item q7 in **Table 14**.

Table 13: Adjustment – DIF on Ethnicity groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q7	Black African	1.21 (.10)	White	.75 (.05)	.039	.46	Slight to moderate
q49	Black African	-.45 (.04)	Asian/Indian	-.88 (.07)	.000	.43	Slight to moderate

Table 14: Adjustment – DIF on Language groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q7	African Languages	1.22 (.10)	Afrikaans	.75 (.07)	.004	.47	Slight to moderate

4.3.3. Curiosity

4.3.3.1. Reliability

HPTi trait *Curiosity* was found to be reliable (initial α = .75, and 1st iteration α = .77). This is supported by the adequate person reliability (.73 before removal, and .76 after removal) and separation indices (1.62 before removal, and 1.79 after removal). Trait *Curiosity* of the HPTi is then regarded as being a reliable measure.

Table 15: Curiosity – Reliability Indices

	Initial	After removal	Difference
Cronbach's alpha	0.75	0.77	0.02
Person reliability	0.73	0.76	0.03
Person separation	1.62	1.79	0.17
Item separation	21.32	20.84	-0.48

4.3.3.2. Unidimensionality and local independence

The unidimensionality of the *Curiosity* subscale was examined. **Figure 3** shows the PCAR results, in which the unexplained variance in the 1st contrast is greater than 2 (1st contrast = 2.34). This is indicative of a scale that may be measuring more than one dimension. Inspection of the PCAR loadings revealed three highest and three lowest contributors to the 1st contrast. Items q30 (.63), q42 (.58) and q72 (.58) had the highest loadings. Items q78r (-.55), q18r (-.54) and q66r (-.50) had the lowest loadings.

Figure 3: Curiosity PCAR

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	24.5460	100.0%		100.0%
Raw variance explained by measures =	11.5460	47.0%		50.2%
Raw variance explained by persons =	3.6853	15.0%		16.0%
Raw Variance explained by items =	7.8607	32.0%		34.2%
Raw unexplained variance (total) =	13.0000	53.0%	100.0%	49.8%
Unexplned variance in 1st contrast =	2.3395	9.5%	18.0%	
Unexplned variance in 2nd contrast =	1.5343	6.3%	11.8%	
Unexplned variance in 3rd contrast =	1.1669	4.8%	9.0%	
Unexplned variance in 4th contrast =	1.0983	4.5%	8.4%	
Unexplned variance in 5th contrast =	1.0529	4.3%	8.1%	

The largest standardized residual correlations were analysed to evaluate the local independence of the items of the scale. No item pairs were found to be above the correlation of .70, ranging from -.30 to .32. This indicates that none of the items of this scale are in violation of local independence.

4.3.3.3. Item fit

The outfit statistics were examined for the items of the HPTi trait, *Curiosity*. A total of four of the 13 items fell outside the range of .75 and 1.30. **Table 16** below lists the items with their relevant fit status.

Table 16: Curiosity - Item Fit Indices and Item Measures

Item	Measure	IN.MNSQ	IN.ZSTD	OUT.MNSQ	OUT.ZSTD	Status
q66r	1.18	1.37	8.37	1.40	8.81	Underfit
q78r	1.02	1.35	7.87	1.41	8.88	Underfit
q36	0.51	1.30	6.34	1.28	5.83	Underfit
q48	0.13	1.12	2.53	1.19	3.80	Fit
q18r	0.88	1.10	2.33	1.14	3.30	Fit
q54	-0.23	0.98	-0.30	1.00	0.07	Fit
q30	-0.95	0.90	-1.78	0.81	-3.37	Fit
q60	0.31	0.89	-2.53	0.94	-1.21	Fit
q24	-1.02	0.86	-2.53	0.80	-3.44	Fit
q72	-0.41	0.80	-4.08	0.80	-3.91	Fit
q6	-0.06	0.77	-5.22	0.77	-5.08	Fit
q12	-0.56	0.77	-4.69	0.74	-5.17	Fit
q42	-0.81	0.76	-4.64	0.72	-5.30	Fit

Table 16 above indicates that items q66r (IN.MNSQ = 1.37, IN.ZSTD = 8.37), q78r (IN.MNSQ = 1.35, IN.MNSQ = 7.87), and q36 (IN.MNSQ = 1.30, IN.ZSTD = 6.34) underfit the Rasch model. These three items are more varied than the expected model by 37%, 35% and 30%, respectively. Two of the three underfitting items had outfit statistics above the 1.30 threshold. These were items q66r (OUT.MNSQ = 1.40, OUT.ZSTD = 8.81) and q78r

(OUT.MNSQ = 1.41, OUT.ZSTD = 8.88). No items were found to overfit the Rasch model based on infit statistics.

4.3.3.4. Differential item functioning

The analysis of the HPTi trait *Curiosity* for differential item functioning, unlike the two preceding traits, had only one of the 13 items showing a significant difference between females and males. However, like the other traits, the difference obtained a negligible effect size, indicating no practical differences between genders on the item (see **Table 17** below).

Table 17: Curiosity – DIF on Gender groups

	Mean (S.E.)				
Item	Females	Males	p-value	DIF Contrast	Effect size
q6	.01 (.04)	-.14 (.05)	.031	.15	Negligible

An inspection of DIF between ethnicities, however, found 8 instances in which DIF may be present (see **Table 18** below).

Table 18: Curiosity – DIF on Ethnicity groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q12	Black African	-.85 (.08)	White	-.40 (.05)	.000	-.44	Slight to moderate
q24	Black African	-1.49 (.11)	Coloured	-.88 (.11)	.001	-.62	Slight to moderate
q24	Black African	-1.49 (.11)	White	-.84 (.06)	.000	-.66	Moderate to large
q24	Asian/Indian	-1.25 (.18)	White	-.84 (.06)	.017	-.42	Approaching slight
q36	Black African	.27 (.06)	Asian/Indian	.67 (.09	.000	-.40	Approaching slight
q78r	Black African	1.40 (.04)	Asian/Indian	.86 (.09)	.000	.54	Slight to moderate
q78r	Black African	1.40 (.04)	Coloured	.91 (.06)	.000	.49	Slight to moderate
q78r	Black African	1.40 (.04)	White	.84 (.04)	.000	.55	Slight to moderate

Table 19: Curiosity – DIF on Language groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q12	African languages	-.86 (.09)	Afrikaans	-.43 (.06)	.000	-.43	Slight to moderate
q24	African languages	-1.47 (.12)	Afrikaans	-.90 (.08)	.004	-.57	Slight to moderate
q24	African languages	-1.47 (.12)	English	-.91 (.06)	.001	-.56	Slight to moderate
q78r	African languages	1.41 (.05)	Afrikaans	.96 (.04)	.000	.45	Slight to moderate
q78r	African languages	1.41 (.05)	English	.83 (.04)	.000	.58	Slight to moderate

Table 18 above reveals that differential item functioning may be present in items q12, q24, q36 and q78r. The latter, q78r, appears to be functioning differently to the Black African group from the other ethnic groups – Asian/Indian, Coloured, and White – with a slight to moderate effect. However, within item q24, it appears that the item functions differently to the Black African group with a higher practical significance than the Coloured and White groups, and that there is also DIF present – albeit to a lesser degree – between the Asian/Indian and White ethnic groups.

A DIF analysis was finally conducted across first language groups, with the statistical and practically significant items listed in **Table 19**. Five instances were found to potentially possess differential item functioning across first language groups. These instances involve items q12, q24 and q78r, similar to the findings across ethnic groups. All instances had a slight to moderate effect size, with the common language group being the African Language group. It may then be necessary to investigate the causes of difference in item functionality for these three items, and the possible review of the items.

4.3.4. Risk Approach

4.3.4.1. Reliability

An analysis of the *Risk Approach* trait's reliability found the measure to just reach sufficiency. Cronbach's alpha before misfitting person were removed was calculated to be .69, marginally falling short of the cut-off value, whereas the Rasch person reliability and separation indices fell further short of adequacy at .64 and 1.33, respectively. After removal of misfitting persons, however, both traditional and Rasch measures of reliability slightly surpassed the adequate reliability threshold (α = .71, person reliability = .71, separation = 1.56).

Table 20: Risk Approach – Reliability Indices

	Initial	1st iteration sample	Difference
Cronbach's alpha	0.69	0.71	0.02
Person reliability	0.64	0.71	0.07
Person separation	1.33	1.56	0.23
Item separation	15.08	14.29	-0.79

4.3.4.2. Unidimensionality and local independence

The unidimensionality of the *Risk Approach* subscale was examined. **Figure 4** shows the PCAR results, in which the unexplained variance in the 1st contrast is less than 2 (1st contrast = 1.62). This indicates that the *Risk Approach* subscale of the HPTi is not measuring more than one dimension.

Figure 4: Risk Approach PCAR

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	19.8170	100.0%		100.0%
Raw variance explained by measures =	6.8170	34.4%		36.7%
Raw variance explained by persons =	1.8069	9.1%		9.7%
Raw Variance explained by items =	5.0101	25.3%		27.0%
Raw unexplained variance (total) =	13.0000	65.6%	100.0%	63.3%
Unexplned variance in 1st contrast =	1.6217	8.2%	12.5%	
Unexplned variance in 2nd contrast =	1.5506	7.8%	11.9%	
Unexplned variance in 3rd contrast =	1.3532	6.8%	10.4%	
Unexplned variance in 4th contrast =	1.2618	6.4%	9.7%	
Unexplned variance in 5th contrast =	1.0759	5.4%	8.3%	

The largest standardized residual correlations were analysed to evaluate the local independence of the items of the scale. No item pairs were found to be above the correlation of .70, ranging from -.24 to .20. This indicates that none of the items of this scale are in violation of local independence.

4.3.4.3. Item fit

The item outfit statics of the *Risk Approach* trait of the HPTi were analysed to determine the fit to the Rach model, and are listed below with their relevant fit status. A total of 2 items were found to breach the range between .75 and 1.30, however, some items boarder on the limits.

Table 21: Risk Approach – Item Fit Indices and Item Measures

Item	Measure	IN.MNSQ	IN.ZSTD	OUT.MNSQ	OUT.ZSTD	Status
q74r	0.01	1.51	9.55	1.55	9.52	Underfit
q8r	0.53	1.40	9.17	1.45	9.53	Underfit
q44r	-0.22	1.29	5.09	1.29	4.93	Fit
q50	0.22	1.19	4.27	1.30	6.08	Fit
q32r	-0.09	1.18	3.47	1.20	3.70	Fit
q2	0.35	0.96	-1.08	1.04	0.90	Fit
q20	0.69	0.87	-3.58	0.93	-1.87	Fit
q56	-0.76	0.86	-2.35	0.76	-4.30	Fit
q26	0.43	0.83	-4.66	0.84	-3.99	Fit
q62r	-0.23	0.80	-4.09	0.78	-4.29	Fit
q14	-0.11	0.79	-4.71	0.79	-4.24	Fit
q38	-0.46	0.79	-3.99	0.79	-3.99	Fit
q68	-0.36	0.78	-4.33	0.77	-4.34	Fit

Table 21 above reveals that items q74r (IN.MNSQ = 1.51, IN.ZSTD = 9.55) and q8r (IN.MNSQ = 1.40, IN.ZSTD = 9.17) underfit the Rasch model. These two items are more varied than the expected model by 51% and 40%, respectively. Outfit statistics for items q74r (OUT.MNSQ = 1.55, OUT.ZSTD = 9.52) and q8r (OUT.MNSQ = 1.45, OUT.ZSTD = 9.53) provide further support for these items underfitting the Rasch model. No items were found to overfit the model.

4.3.4.4. Differential item functioning

An analysis of differential item functioning of the *Risk Approach* scale revealed a significant difference between males and females on 4 of the 13 items. However, the practical significance of these differences can be regarded as negligible.

Table 22: Risk Approach – DIF on Gender groups

Item	Mean (S.E.)		p-value	DIF Contrast	Effect size
	Females	Males			
q44r	-.26 (.04)	-.16 (.04)	.050	-.10	Negligible
q56	-.76 (.05)	-.73 (.06)	.005	-.03	Negligible
q62r	-.26 (.04)	-.19 (.04)	.033	-.07	Negligible
q68	-.40 (.04)	-.30 (.05)	.027	-.10	Negligible

The DIF inspection across ethnicities found three instances where DIF may be present, all of which occurred on one item in the *Risk Approach* scale, item q74r (see **Table 23**) below.

Table 23 indicates that DIF is likely present in item q74r on the *Risk Approach* scale. It can be observed that the common ethnic group where DIF may be present is the Coloured ethnic group. When compared to all other ethnic groups in this sample, the practical significance ranges from slight to large on item q74r, which could warrant further investigation into this item.

Finally, for the DIF analysis across first language groups, no instances were found to meet the criteria indicative of DIF being present. However, one instance was found to approach the criteria with a DIF contrast of .41. This instance occurred in item q50 across the African Languages and Afrikaans language groups.

Table 23: Risk Approach – DIF on Ethnicity groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q74r	Coloured	-.54 (.09)	Asian/Indian	-.06 (.09)	.011	-.48	Slight to moderate
q74r	Coloured	-.54 (.09)	Black African	.04 (.05)	.000	-.58	Slight to moderate
q74r	Coloured	-.54 (.09)	White	.13 (.03)	.000	-.67	Moderate to large

Table 24: Risk Approach – DIF on Language groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q50	African languages	.45 (.04)	Afrikaans	.05 (.04)	.000	.41	Approaching slight

4.3.5. Ambiguity Acceptance

4.3.5.1. Reliability

The reliability indices of the *Ambiguity Acceptance* trait were found to boarder on sufficiency. Before the removal of misfitting persons, neither Cronbach's alpha nor person reliability and separation meet the accepted cut-offs (α = .68, person reliability = .65, separation = 1.35). After misfitting persons are removed, only Cronbach's alpha reaches sufficiency (α = .71), while the Rasch measures of reliability boarder on being adequate but do not quite meet the criteria (person reliability = .69, separation = 1.49).

Table 25: Ambiguity Acceptance – Reliability indices

	Initial	1st iteration sample	Difference
Cronbach's alpha	0.68	0.71	0.03
Person reliability	0.65	0.69	0.04
Person separation	1.35	1.49	0.14
Item separation	23.24	22.01	-1.23

4.3.5.2. Unidimensionality and local independence

The unidimensionality of the *Ambiguity Acceptance* subscale was examined. **Figure 5** shows the PCAR results, in which the unexplained variance in the 1st contrast is less than 2 (1st contrast = 1.96). Although the *Ambiguity Acceptance* subscale did not obtain a 1st contrast eigenvalue of 2, the value approaches the recommended cutoff value and may therefore benefit from further investigation of its unidimensionality.

Figure 5: Ambiguity Acceptance PCAR

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	24.4978	100.0%		100.0%
Raw variance explained by measures =	11.4978	46.9%		47.5%
Raw variance explained by persons =	2.0276	8.3%		8.4%
Raw Variance explained by items =	9.4702	38.7%		39.1%
Raw unexplained variance (total) =	13.0000	53.1%	100.0%	52.5%
Unexplned variance in 1st contrast =	1.9570	8.0%	15.1%	
Unexplned variance in 2nd contrast =	1.5637	6.4%	12.0%	
Unexplned variance in 3rd contrast =	1.3376	5.5%	10.3%	
Unexplned variance in 4th contrast =	1.1698	4.8%	9.0%	
Unexplned variance in 5th contrast =	1.0922	4.5%	8.4%	

The largest standardized residual correlations were analysed to evaluate the local independence of the items of the scale. No item pairs were found to be above the correlation of .70, ranging from -.27 to .24. This indicates that none of the items of this scale are in violation of local independence.

4.3.5.3. Item fit

The outfit statistics of the items of the *Ambiguity Acceptance* trait of the HPTi were examined. Four of the 13 items exceeded the range of 0.75 and 1.30. **Table 26** below lists the items with their relevant fit status.

Table 26: Ambiguity Acceptance – Item Fit Indices and Item Measures

Item	Measure	IN.MNSQ	IN.ZSTD	OUT.MNSQ	OUT.ZSTD	Status
q5r	-0.64	1.51	9.90	1.45	8.31	Underfit
q77	-0.65	1.37	7.53	1.41	7.66	Underfit
q29r	0.18	1.26	6.32	1.27	6.21	Fit
q65r	0.08	1.10	2.65	1.12	3.07	Fit
q17r	-0.45	1.08	1.94	1.06	1.41	Fit
q11r	1.06	1.05	0.90	1.00	0.09	Fit
q35	-0.31	0.98	-0.58	0.96	-0.96	Fit
q47	-0.84	0.96	-0.76	0.97	-0.58	Fit
q53r	0.35	0.91	-2.22	0.94	-1.36	Fit
q59r	0.42	0.82	-4.71	0.82	-4.42	Fit
q23r	0.09	0.80	-5.70	0.81	-5.23	Fit
q41r	0.10	0.74	-7.76	0.74	-7.19	Overfit
q71r	0.61	0.67	-8.38	0.65	-8.29	Overfit

Table 26 above indicates that most items of the *Ambiguity Acceptance* trait fit the Rasch model. However, items q5r (IN.MNSQ = 1.51, IN.ZSTD = 9.90) and q77 (IN.MNSQ = 1.37, IN.ZSTD = 7.53) were found to underfit the model. The infit mean-squared indices suggest that these two items are 51% and 37% more varied than the model, respectively. The outfit statistics of q5r (OUT.MNSQ = 1.45, OUT.ZSTD = 8.31) and q77 (OUT.MNSQ = 1.41, OUT.ZSTD = 7.66) are also indicative that these items underfit the model.

Items q41r (IN.MNSQ = 0.74, IN.ZSTD = -7.76) and q71r (IN.MNSQ = 0.67, IN.ZSTD = -8.38) were found to overfit the Rasch model. The infit mean-squared indices indicate that these two items are 26% and 33% less varied than the expected model, respectively. The outfit statistics of items q41r (OUT.MNSQ = 0.74, OUT.ZSTD = -7.19) and q71r (OUT.MNSQ = 0.65, OUT.ZSTD = -8.29) are also indicative of these two items overfitting the model.

4.3.5.4. Differential item functioning

A DIF analysis, conducted on the HPTi trait *Ambiguity Acceptance* between gender groups, resulted in two statistically significant instances on items q41r and q65r of the 13-item scale. However, these instances had negligible practically significant differences.

Table 27: Ambiguity Acceptance – DIF on Gender groups

	Mean (S.E.)				
Item	Females	Males	p-value	DIF Contrast	Effect size
q41r	.16 (.03)	.04 (.03)	.015	.12	Negligible
q65r	.12 (.03)	.02 (.03)	.038	.10	Negligible

An inspection of DIF across ethnicities found three occurrences in which DIF may be present. All three instances were found in item q11r. The first instance between the Black African group and Asian/Indian group narrowly missed the criteria, and is therefore considered to be approaching a slight effect size. The second instance between the Black African and Coloured ethnic group meets the criteria to be considered a slight to moderate effect size. The third occurrence between the Black African and White ethnic groups was found to have a moderate to large practical significance. This result may therefore warrant an investigation into the functioning of item q11r between the Black African ethnic group and the other ethnic groups.

Finally, the DIF analysis across the first language groups discovered two instances in which DIF may be present. Both occurrences exist also on item q11r between the African Languages group and the Afrikaans group with a slight to moderate effect size, and between the African Languages and English group with a moderate to large effect size. This evidence indicates that there may be a differential functioning on item q11r between the Black African

group whose first language is an African language and those who are not Black African and whose first language is not an African language.

Table 28: Ambiguity Acceptance – DIF on Ethnicity groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q11r	Black African	1.54 (.08)	Asian/Indian	1.12 (.10)	.008	.42	Approaching slight
q11r	Black African	1.54 (.08)	Coloured	1.06 (.08)	.000	.48	Slight to moderate
q11r	Black African	1.54 (.08)	White	.86 (.04)	.000	.69	Moderate to large

Table 29: Ambiguity Acceptance – DIF on Language groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q11r	African languages	1.57 (.08)	Afrikaans	.98 (.05)	.000	.59	Slight to moderate
q11r	African languages	1.57 (.08)	English	.93 (.04)	.000	.64	Moderate to large

4.3.6. Competitiveness

4.3.6.1. Reliability

Cronbach's alpha before and after misfitting persons were removed suggest that the measure is reliable (α = .75 before, and .74 after). However, the person reliability and separation indices are indicative of a reliable measure after misfitting persons are removed, where the reliability is .69 and separation is 1.48 before removal, and .72 and 1.61, respectively, after removal of misfitting persons.

Table 30: Competitiveness – Descriptives, Person Fit Exclusions, and Reliability Indices

	Initial	1st iteration sample	Difference
Cronbach's alpha	0.75	0.74	-0.01
Person reliability	0.69	0.72	0.03
Person separation	1.48	1.61	0.13
Item separation	26.54	26.44	-0.10

4.3.6.2. Unidimensionality and local independence

The unidimensionality of the *Competitiveness* subscale was examined. **Figure 6** shows the PCAR results, in which the unexplained variance in the 1st contrast is less than 2 (1st contrast = 1.63). This indicates that the *Competitiveness* subscale of the HPTi is not measuring more than one dimension.

Figure 6: Competitiveness PCAR

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units

	Eigenvalue	Observed		Expected
Total raw variance in observations =	23.1963	100.0%		100.0%
Raw variance explained by measures =	10.1963	44.0%		44.4%
Raw variance explained by persons =	2.0061	8.6%		8.7%
Raw Variance explained by items =	8.1902	35.3%		35.6%
Raw unexplained variance (total) =	13.0000	56.0%	100.0%	55.6%
Unexplned variance in 1st contrast =	1.6294	7.0%	12.5%	
Unexplned variance in 2nd contrast =	1.5322	6.6%	11.8%	
Unexplned variance in 3rd contrast =	1.4117	6.1%	10.9%	
Unexplned variance in 4th contrast =	1.1836	5.1%	9.1%	
Unexplned variance in 5th contrast =	1.1505	5.0%	8.9%	

The largest standardized residual correlations were analysed to evaluate the local independence of the items of the scale. No item pairs were found to be above the correlation of .70, ranging from -.22 to .16. This indicates that none of the items of this scale are in violation of local independence.

4.3.6.3. Item fit

The outfit statistics were inspected for the items of the HPTi trait, *Competitiveness.* Only one of the 13 items was discovered to fall outside the 0.75 and 1.30 range. **Table 31** below lists the items and their respective fit statuses.

Table 31: Competitiveness – Item Fit Indices and Item Measures

Item	Measure	IN.MNSQ	IN.ZSTD	OUT.MNSQ	OUT.ZSTD	Status
q76	0.45	1.36	8.51	1.35	7.81	Underfit
q58	0.47	1.15	3.77	1.16	3.70	Fit
q46r	0.19	1.13	3.46	1.13	3.21	Fit
q10	-2.24	1.08	1.18	1.00	0.02	Fit
q40r	0.25	1.05	1.38	1.09	2.25	Fit
q28	0.00	1.04	0.94	1.03	0.74	Fit
q34	0.08	0.99	-0.20	0.99	-0.29	Fit
q64r	0.28	0.99	-0.14	1.00	-0.01	Fit
q52r	0.69	0.92	-1.98	0.94	-1.25	Fit
q22	0.28	0.89	-3.00	0.89	-2.98	Fit
q70	-0.40	0.89	-2.70	0.87	-3.05	Fit
q4	-0.17	0.84	-4.33	0.82	-4.62	Fit
q16	0.12	0.81	-5.47	0.86	-3.68	Fit

Table 31 above reveals that item q76 (IN.MNSQ = 1.36, IN.ZSTD = 8.51) underfits the Rasch model. The infit mean-squared suggests that this item is 36% more varied than the expected model. The outfit statistics for item q76 (OUT.MNSQ = 1.35, OUT.ZSTD = 7.81) also indicate that the item is underfitting the model. No items were found to overfit the Rasch model.

4.3.6.4. Differential item functioning

The analysis of differential item functioning of the *Competitiveness* scale across the gender group found five of the 13 items to be statistically significant between females and males. However, all five instances had negligible effect sizes.

Table 32: Competitiveness – DIF on Gender groups

	Mean (S.E.)				
Item	Females	Males	p-value	DIF Contrast	Effect size
q16	.07 (.03)	.18 (.03)	.038	-.11	Negligible
q28	.11 (.03)	-.13 (.03)	.000	.24	Negligible
q34	.00 (.03)	.16 (.03)	.000	-.16	Negligible
q40r	.19 (.03)	.32 (.03)	.016	-.13	Negligible
q46r	.25 (.03)	.13 (.03)	.015	.11	Negligible

One item – item q10 – was discovered to have among the highest practical differences between groups, with DIF contrasts ranging from |.75| to |1.00| between ethnic groups and first language groups.

With regards to ethnic groups, item q10 had a moderate to large effect size between the White ethnic group and the Asian/Indian group, and the White ethnic group and Black African ethnic group.

In terms of first languages, moderate to large effect sizes were found between the African Languages group and the Afrikaans first language group, and the African Languages group and English first language group for item q10. Investigation into the functioning of item q10 may therefore be justified, as this item appears to have the highest levels of practical significance across all DIF analyses.

Table 33: Competitiveness – DIF on Ethnicity groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q10	White	-1.90 (.07)	Asian/Indian	-2.65 (.23)	.000	.75	Moderate to large
q10	White	-1.90 (.07)	Black African	-2.91 (.14)	.000	1.00	Moderate to large

Table 34: Competitiveness – DIF on Language groups

Item	Group 1	Mean (S.E.)	Group 2	Mean (S.E.)	p-value	DIF contrast	Effect size
q10	African languages	-2.93 (.15)	Afrikaans	-2.00 (.09)	.000	-.93	Moderate to large
q10	African languages	-2.93 (.15)	English	-2.13 (.08)	.000	-.79	Moderate to large

4.4. Chapter conclusion

In this chapter, results of the preliminary and primary analyses were provided. The preliminary analysis described the initial sample of 1257 respondents, and the samples of each of the iterated versions of the HPTi traits after the removal of the misfitting persons, as well as describing the response patterns in classical test format and Rasch measurement format.

The primary analysis of the study had each iterated HPTi trait undergo a series of Rasch analysis techniques to determine item fit to the Rasch model, person and item reliability and separation, and differential item functioning; ultimately assessing the reliability and validity of the HPTi from a Rasch measurement perspective.

In terms of reliability, only one trait, *Curiosity*, was found to have a sufficient alpha, person reliability and person separation index before and after the removal of misfitting persons. On the other hand, the iterated sample *Ambiguity acceptance* was the only scale not to meet the criteria – albeit by .01 – for the person reliability index, and person separation, yet achieved a sufficient Cronbach's alpha. Traits *Adjustment* and *Competitiveness*, like *Curiosity*, met the criteria on the Cronbach's alpha measure of reliability before and after person removal. However, unlike *Curiosity*, fell short of the criteria based on their person reliability and separation indices before removal of misfitting persons, but achieved sufficiency after removal. Interestingly, *Competitiveness* was the only scale in which Cronbach's alpha decreased after person removal. *Conscientiousness* and *Risk approach* followed similar patterns to each other in which they both fell short of the mark in alpha, person reliability and person separation before removal, and achieved sufficiency after the removal of misfitting persons.

The item fit analysis of the HPTi traits found that 14 of the 78 items (18%) were misfitting the model; of which 10 were underfitting, and four were overfitting. In the *Conscientiousness* scale, one of the 13 items – item q57r – underfit the Rasch model, and one item – q63 – overfit the model. Similarly, the *Adjustment* scale had two misfitting items, of which q1 overfit the model and q31 underfit the mode. Trait *Curiosity* was found to have three misfitting items, in which items q78r, q66r, and q36 were found to underfit the model, with no overfitting items. Likewise, the *Risk approach* scale had two underfitting items – items q74r and q8r – with no items overfitting the model. In the *Ambiguity acceptance* scale,

items four items were found to misfit the Rasch model. Items q5r and q77 were regarded as underfitting the model, while items q41r and q71r overfit the model. The *Competitiveness* scale was found to have only one item, q76, as underfitting the Rasch model, with no items overfitting the model.

Finally, a differential item functioning analysis was conducted across all six HPTi traits across gender, ethnicity, and first language groups to determine the presence of items that may not be congruently defined or interpreted between these groups. DIF was not present between females and males across the six traits.

In terms of the items with the highest effect size, item q10 in the *Competitiveness* scale had been found to have a DIF contrast as high as 1.00, with the lowest instance for q10 having a DIF contrast of .75. However, unlike item q10, the next two items with the highest effect sizes are also items that are misfitting the Rasch model. Item q74r of *Risk approach* is an underfitting item which has a slight to large effect size between ethnic groups, with the largest (|.67|) between the Coloured ethnic group and the White ethnic group. At the same length, item q21 of *Conscientiousness* is an overfitting item which has a slight to large effect size between ethnic groups, with the largest DIF contrast (|.77|) between the Asian/Indian ethnic group and the White ethnic group. However, DIF also appears to be evident between the African Languages group and the English and Afrikaans language groups.

Though other instances indicating a possible presence of DIF exists to a smaller degree on the DIF contrast measure and in instances other than what appears to be the trend, the trend seems to be that the main differences occur between the Black African ethnic group and the other ethnic groups, and the African Language group and the other first language groups, Afrikaans and English.

Chapter 5:
Discussion

5.1. Introduction

The purpose of this study was to evaluate, through Rasch analysis techniques and with a South African sample, the psychometric properties of the High Potential Trait Indicator, including the fit of the items to the Rasch model using fit statistics, and the presence of differential item functioning within the instrument. A first for this assessment with a South African sample. The discussion section focuses on the results obtained following the removal of the misfitting persons, and concludes with possible limitations of the study.

5.2. Outline of results

5.2.1. Person-fit

The person-fit analysis was conducted first to identify responses provided in a manner that is either too predictable or unpredictable, and therefore not fitting the Rasch model (Bond et al., 2020; Boone et al., 2014) and potentially invalidating other estimates (Curtis, 2004; Mousavi & Cui, 2020). The result of the person-fit analysis saw between 8% (*Conscientiousness*) and 20% (*Risk approach*) of the total sample as misfitting and subsequently removed.

5.2.2. Descriptive statistics

The descriptive statistics in **Table 4** reveal that traits *Conscientiousness*, *Adjustment*, *Curiosity*, and *Risk Approach* could not be considered normally distributed, with skewness indices situated outside the accepted range of -1 to 1 (Groeneveld & Meeden, 1984). Among the four traits *Adjustment* was negatively skewed while the other three were positively skewed. The remaining traits, *Ambiguity acceptance* and *Competitiveness* had skewness indices located near zero, indicative of a symmetrical distribution (Howell, 2014).

Similarly, the average person measures, measured in logits, for traits *Conscientiousness*, *Adjustment*, *Curiosity*, and *Risk Approach* are situated further from zero than the less skewed traits *Ambiguity acceptance* and *Competitiveness*. Item measure averages are constrained to zero in Rasch analyses. Average person measures on an instrument that markedly vary from the average item measure of zero are indicative of the measure either being too easy to agree with or too difficult to agree with, with regards to

rating scales (Bond et al., 2020; Boone et al., 2014). The matching of the person and item measure means is related to the phenomenon known as ‘targeting’. Targeting refers to the matching distribution of items to the range of abilities or characteristics of the sample, where a well-targeted instrument ideally has a mean and standard deviation close to that of the sample (Bond et al., 2020; Boone et al., 2014). Therefore, an instrument developed for administrating to the general population should be targeted towards the characteristics of the general population. In reference to evaluating a measuring instrument, the recommendation is to adjust the items accordingly, however, the difference could be due to the nature of the respondents in the sample. For example, if a scale that has been well calibrated to where the person and item measures share a mean of zero is administered to a team of competitive athletes, it may not be unexpected to find that specific sample’s average person measure on *Competitiveness* to be above the average item measure. The person measure averages for traits *Ambiguity acceptance* and *Competitiveness* were relatively close to the item average of zero, indicating that these traits were neither too easy to endorse nor too difficult, which is reflected as so in their minimum and maximum values. *Conscientiousness*, on the other hand, with a mean of 1.3 logits has a difference of 1.3 logits to the item mean. This is indicative of respondents finding the items easier to agree with overall, and is reflected in a minimum value equal to the constrained item average, and a high maximum value. Trait *Curiosity* has a similar outcome, and the opposite is true for trait *Adjustment* which may contain items that are more difficult to agree with, with a mean person measure of -0.85 and a minimum and maximum of -4.82 and 0.63, respectively.

5.2.3. Reliability

Item reliability indices for all traits demonstrated high reliability, indicating that the sample size was large enough to accurately estimate the item difficulty (Boone et al., 2014; Linacre, 2020). On the other hand, the person reliability indices mostly bordered on the accepted cut-off of .70, ranging from .69 (*Ambiguity acceptance*) to .76 (*Curiosity*). Person reliability, according to Linacre (2020), is largely dependent on the dispersion of the characteristics of the sample – in other words, a sample with varying degrees of the trait being measured, – the length of the instrument, the number of response options per item, and the targeting of the sample and items. Some traits’ reliability indices are potentially being impacted by the inadequate targeting, evident in the large differences between the average person scores in traits *Conscientiousness*, *Adjustment*, *Curiosity*, and *Ambiguity acceptance*, and the constrained item measure average of zero. A recommendation provided by Boone et

al. (2014) for addressing this matter is to revise items to either make them easier or more difficult to endorse, depending. On the other end, the scales that appear well-targeted – *Ambiguity acceptance*, and *Competitiveness* – may have their respective reliability indices impacted by the inability to adequately separate the higher scorers on the trait from those who have lower person measures of that trait being, resulting in a lower person separation index and therefore reliability index.

Given the results, it may be difficult, especially for trait *Ambiguity acceptance*, to defend the reliability of the HPTi subscales, with *Curiosity* being the most defensible. Interestingly, the results are not dissimilar to the alpha coefficients in MacRae and Furnham (2020). In their article, MacRae and Furnham found trait *Curiosity* to be the most reliable (α = .80, n = 779), while *Courage* (known as *Risk approach* in the HPTi, α = .72), *Ambiguity* (*Ambiguity acceptance* in the HPTi, α = .74), and *Competitiveness* (α = .74) approach an inadequate reliability index.

5.2.4. Item-fit

Item fit statistics evaluated how well the items of each HPTi trait conformed to the Rasch model, with outfit mean-squared statistics greater than 1.30 indicating that the item is 30% less predictable (more varied) than the model, and the same statistic under 0.75 indicating that the item is 25% more predictable (less varied) than the model expects (Bond et al., 2020). Although Linacre (2020) and Bond et al. (2020) make use of different ranges for fit statistic cut-offs, Linacre provides an interpretation of what different ranges mean. For example, Linacre stipulates that a mean-square greater than 1.5 but less than 2.0 as being unproductive for instrument construction but not yet corruptive, with a similar interpretation for fit values less than 0.5, with the add-on that lower values lead to the exaggeration of reliability indices. The interpretation can be transferred to the more conservative cut-offs of 1.30 and 0.75, respectively, used in this study. Regardless of the cut-offs chosen, the recommendation is not to merely discard misfitting items, but instead to flag as possibly requiring further investigation.

Therefore, in trait *Conscientiousness*, it was found that item q57r was found to be more varied than the expected model but to the degree that it does not degrade the instrument. Item q63, on the other hand may be more predictable (less varied) than is expected, and are therefore considered overfitting.

For *Adjustment*, item q31 was found to be underfitting, while item q1 was found to be overfitting the Rasch model.

Curiosity observed three underfitting items: q78r, q66r and q36; and no overfitting items. However, having observed a PCAR eigenvalue above two, items q78r and q66r were found to be contributing highly to a possible second dimension – as evaluated by the PCAR loadings. The eigenvalue of the *Curiosity* subscale being greater than two suggests that more than one underlying dimension is being measured. Furthermore, it may be items q78r and q66r threatening the unidimensionality of this subscale, as evidenced by the high loadings of these items and their underfitting nature (Fan & Bond, 2019).

The subscale, *Risk approach*, found two underfitting items present: q74r and q8r. Meanwhile, trait *Ambiguity acceptance* observed two underfitting items: q5r and q77, as well as two overfitting items: q41r and q71r. The unidimensionality of *Ambiguity acceptance* appeared to be threatened by item q77, as it this item was both underfitting as well as contributing highly to the 1st contrast in the PCAR. The 1st contrast of this HPTi trait approached the significant eigenvalue of two.

Competitiveness had only one misfitting item, item q76 which was evaluated to be underfitting the Rasch model. According to Tennant and Conaghan (2007), however, fit to the Rasch model can be influenced by item bias such as differential item functioning (DIF).

5.2.5. Differential Item Functioning (DIF)

DIF is an evaluation of how congruently the items of a measure define a construct between certain groups (Boone, et al., 2014), and is important in cross-cultural settings (Tennant et al., 2004). Across all scales, while statistically significant DIF was found between gender groups the findings were not practically significant, as measured by DIF contrast, where .43 to .64 is considered a slight to moderate effect size, and .64 and greater are considered a moderate to large effect size (Linacre, 2020). This may indicate that the traits measured by the HPTi do not contain items whose definitions are interpreted differently between men and women. Therefore, none of the items of the HPTi could be considered biased towards either males or females.

However, when it came to ethnic groups and first language groups DIF was discovered on a few items, with varying severity, across the scales. With regards to severity of DIF contrast, item q10 in the *Competitiveness* scale obtained the highest DIF contrast of all

the items, with 0.75 and 1.00 between the White ethnicity and Asian/Indian ethnicity, and the White ethnicity and Black African ethnicity, respectively; and -0.93 and -0.79 between the African Languages first language group and Afrikaans first language group, and African Languages first language group and English first language group, respectively. This is followed by item q21 in the *Conscientiousness* scale with a DIF contrast of -0.77, -0.64, -0.63, and -0.50 when comparing the Asian/Indian ethnicity to the White ethnicity groups, Asian/Indian ethnicity to the Coloured ethnicity groups, Black African to White ethnicity groups, and Black African to Coloured ethnicity groups, respectively. Item q11r of the *Ambiguity Acceptance* scale was also picked up to contain DIF with a DIF contrast of 0.69 when comparing the Black African ethnicity group to the White ethnicity group, and 0.64 when comparing African Languages first language groups to the English first language groups. From the *Risk Approach* scale, item q74r is the next most severe item experiencing DIF. This item had a DIF contrast of -0.67 between the Coloured and White ethnicity groups. Finally, with regards to the severity of DIF contrast, item q24 of the *Curiosity* scale had a DIF contrast of -0.66 when comparing the Black African ethnicity group with the White ethnicity group, and -0.62 when comparing the Black African ethnicity group with the Coloured ethnicity group. The *Adjustment* scale, although containing slight to moderate DIF contrasts, was the only scale not to have moderate to large DIF contrasts. Furthermore, while *Competitiveness* had the item with the largest DIF contrast, the scale only had one item experiencing DIF. This item (q10) also experienced severe DIF between multiple ethnicity and first language groups. On the other hand, *Adjustment*, while not experiencing DIF contrasts as high as *Competitiveness* had more items with some evidence of DIF. This, however, is minor in comparison to the scales measuring *Conscientiousness* and *Curiosity*, with 4 and 3 items, respectively, of which one has a DIF contrast above |0.64|, and 12 and 9 instances of a DIF contrast above |0.43|, respectively.

The findings of item-level bias are not dissimilar with typical findings of personality questionnaires in the South African context, in which bias between African and European ethnic groups and language groups are usually found and recommended for further investigation (Abrahams & Mauer, 1999; Boeyens, 1991; Spence, 1982) It may therefore be required that the relevant HPTi items are re-examined in the South African context to reduce the item-level bias.

5.2.6. Limitations

The current study is not without limitations. First, with respect to the methodology, is the use of secondary data obtained from an organisation whose use in psychometric tools is largely in the recruitment sector (thomas.co). This falls into the disadvantage of secondary data research expressed by Boslaugh (2007), to which secondary data is often collected for purposes other than that of the research question using the secondary data. This further links to the second limitation which is inherent in self-reporting personality assessments, especially used in decision-making, in which respondents may respond in a way that distorts or misrepresents them (Coaley, 2010). It is therefore not unimaginable to have obtained data with some responses skewed towards the purposes the HPTi was originally administered for, such as applications to employment. To address these two limitations, further research is encouraged in which respondents are randomly selected, the administration is standardized, and the purpose of completing the assessment is exclusively for research and not for, say, the application process for employment.

5.2.7. Chapter conclusion

In evaluating the psychometric properties of the HPTi in the South African context using Rasch analysis techniques, it had been found that the six scales of the HPTi bordered on the lower end of the reliability cut-off, ranging from .69 (*Ambiguity Acceptance*) on the Rasch person reliability index to .76 (*Curiosity*). The fit to the Rasch model of the items of the HPTi found that there are items that over-fit and items that under-fit the Rasch model, indicative of items that are less and more varied than the expected Rasch model, respectively. These items are not considered detrimental to the measurement itself but unproductive to the calibration of the test. Several items across all scales were found to display some degree of differential item functioning. The most severe observation of DIF was in item q10 of *Competitiveness,* and the least overall severity in the *Adjustment* scale with no effect sizes greater than 0.64. *Conscientiousness* had the most incidents of DIF with four items displaying a combined count of 12 DIF contrasts above 0.43. None of the DIF cases with a large enough effect size were found between the gender group comparison, but rather between different ethnicities and between first language groups.

The limitations of this study were explained to be the use of secondary data, to which the data may not have originated to address the research questions of this study. This issue is further exacerbated by the fact that the assessment is largely used in, but not limited to, the

recruitment process of organisations, to which the respondents of this self-reported questionnaire have the capacity to respond in a manner that may be desirable to the organisation but not a true reflection of the respondent. Therefore, further research into the psychometric properties of the HPTi using Rasch analysis techniques should be conducted in a setting in which the process is optimized for the most accurate responses from respondents.

References

Abrahams, F., & Mauer, K. F. (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.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Arráiz, I., Bruhn, M., & Stucchi, R. (2016). Psychometrics as a tool to improve credit information. *The World Bank Economic Review, 30*(1), 67-76. https://doi.org/10.1093/wber/lhw016

Barrick, M. R., Mount, M. K., Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*(1), 9-30. Doi: 10.111/1468-2389.00160.

Bichi, A. A. (2016). Classical test theory: An introduction to linear modelling approach to test and item analysis. *International Journal for Social Studies, 2*(9), 27-33. https://www.researchgate.net/publication/317012320

Bond, T. G., Yi, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4rd ed.). New York, NY: Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. New York, NY: Springer. doi:10.1007/978-94-007-6857-4.

Brouwers, S. A., Mostert, K., & Mtshali, S. V. (2017). Bias and equivalence of the Strengths Use and Deficit COrrection Quiestionnaire in a South African context. *South African Journal of Human Resource Management, 15*, 1-11. doi:https://doi.org/10.4102/sajhrm.v15i0.889

Cattell, R.B., Eber, H.W. & Tatsuoka, M.M. (1992). Handbookfor the Sixteen Personality Factor Questionnaire. Champaign, lL: Institute for Personality & Ability Testing.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. London, UK: SAGE Publications.

Crocker, L., & Algina, J. (2008). *Introduction to classical test theory and modern test theory*. Mason, OH: Cengage Learning. ISBN-10: 0-495-39591-9

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi:10.1007/bf02310555

DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. New York, NY: Oxford University Press.

DeVillis, R. F. (2017). *Scale development: Theory and application* (4th ed.). Los Angeles, CA: Sage.

Duncan, P. W., Bode, R. K., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The Stroke Impact Scale. *Archives of Physical Medicine and Rehabilitation, 84*(7), 950-963. doi:10.1016/S0003-9993(03)00035-2.

Employment Equity Act, number 55 (1998). Government Gazette, 400 (19370), Cape Town, 19 October.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-374.

Fan, J., & Bond, T. (2019). Applying the Rasch measurement model in language assessment: Unidimensionality and local independence. In Aryadoust & Raquel (Eds.), *Quantitative data analysis for language assessment volume 1: Fundamental techniques* (pp. 81-176). Abingdon, UK: Routledge.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, UK: Sage Publications.

Finch, W. H., Immekus, J. C., & French, B. F. (2016). *Applied psychometrics using SPSS and AMOS*. Charlotte, NC: Information Age Publishing.

Foxcroft, C., & Roodt, G. (2018). An overview of assessment: Definition and scope. In Foxcroft & Roodt (Eds.), *An introduction to psychological assessment in the South African context* (5th ed., pp. 3-11). Cape Town, South Africa: Oxford University press.

Government Gazette, Republic of South Africa, Vol. 400, no. 19370. Cape Town, 19 October 1998.

Government Gazette, Republic of South Africa, no. 29079. Cape Town, 4 August 2006.

Groeneveld, R., A., & Meeden, G. (1984). Measuring skewness and kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician), 33*(4), 391-399. https://doi-org.uplib.idm.oclc.org/10.2307/2987742

Hambelton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical care, 38*(9), 60-65. DOI: 10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Health Professions Act, Act 56 of 1974. Health Professions Council of South Africa.

Heffernan, E., Maidment, D. W., Barry, J. G., & Ferguson, M. A. (2019). Refinement and validation of the Social Participation Restrictions Questionnaire: An application of Rasch analysis and traditional psychometric analysis techniques. *Ear & Hearing, 40*(2), 328-339.

Herman, J. L., Stevens, M. J., Bird, A., Mendenhall, M., & Oddou, G. (2010). The Tolerance for Ambiguity Scale: Towards a more refined measure for international management research. *International Journal of Intercultural Relations, 34*(1), 58-65. https://doi.org/10.1016/j.ijintrel.2009.09.004

Howell, D. C. (2014). *Fundamental Statistics for the Behavioural Sciences* (8th ed.). Wadsworth, CA: Cengage Learning.

Huysamen, G. K. (1996). *Psychological measurement: An introduction with South African examples* (3rd rev. ed.). Pretoria: Van Schaik.

Huysamen, G. K. (2002). The relevance of the APA standards for educational and psychological testing for employment testing in South Africa. *South African Journal of Psychology, 32*, 26-33.

IBM Corp. (2017). *IBM SPSS® Statistics for Windows* (Version 25.0) [Computer Software]. Armonk, NY: IBM Corp.

Judge, T. A., & Locke, E. A. (1992). The effect of dysfunctional thought processes on subjective well-being and job satisfaction. *Journal of Applied Psychology, 78*(3), 475-490. Doi:10.1037/0021-9010.78.3.475

Judge, T. A., Higgins, C. A., Thoresan, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the lifespan. *Personnel Psychology, 52*(3), 621-652. https://doi.org/10.1111/j.1744-6570.1999.tb00174.x

Kapalan, R. M., & Saccuzzo, D. P. (2018). *Psychological testing: Principles, applications, & issues* (9th ed.). Boston, MA: Cengage Learning.

Kean, J., Bisson, E. F., Brodke, D. S., Biber, J., & Gross, P. H. (2017). An introduction to Item Response Theory and Rasch analysis using the Eating Assessment Tool (EAT-10). *Brain Impairment, 19*(1), 91-102.

Keenan, A., & McBain, G. D. M. (1979). Effects of Type A behaviour, intolerance of ambiguity, and locus of control on the relationship between role stress and work-related outcomes. *Journal of Occupational and Organizational Psychology, 52*(4), 277-285. https://doi.org/10.1111/j.2044-8325.1979.tb00462.x

Laher, S., & Cockcoft, K. (2013). Contextualising psychological assessment in South Africa. In Laher & Cockcroft (Eds.), *Psychological assessment in South Africa: Research and applications*. Johannesburg, South Africa: Wits University Press.

Linacre, J. M. (2004). Estimation methods for Rasch measures. In Smith & Smith (Eds.), *Introduction to Rasch Measurement* (pp. 25-47). Maple Grove, MN: JAM Press.

Linacre, J. M. (2020). *Winsteps*® (Version 4.5.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. https://www.winsteps.com/

Linden, D., Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of the big five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*, 315-327. https://doi.org/10.1016/j.jrp.2010.03.003

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30*(3), 239-270.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Charlotte, NC: Information Age Publishing

MacRae, I. (2012). *Success, Potential and Validating a Measure of High Flying Personality Traits in Organisations*. Unpublished Master's Dissertation, London, UK: University College London.

MacRae, I., & Furnham, A. (2014). *High Potential: How to Spot, Manage and Develop Talented People at Work*. London: Bloomsbury Publishing.

MacRae, I., & Furnham, A. (2016). *High Potential Traits Inventory: Leadership Capacity Testing Manual*. High Potential Psychology Ltd.

MacRae, I., & Furnham, A. (2020). A psychometric analysis of the High Potential Trait Inventory (HPTI). *Psychology, 11*, 1125-1140).

Magno, C. (2009). Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1-11.

Manapsal, J. P. (2017). A comparative study on Classical Test Theory and Rasch Model in Comprehensive Mental Ability Test (CMAT): A pilot test. *International Journal of Sciences: Basic and Applied Research, 33*(2), 266-283. https://www.researchgate.net/publication/326126602_A_Comparative_Study_on_Classical_Test_Theory_and_Rasch_Model_in_Comprehensive_Mental_Ability_Test_CMAT_A_Pilot_Test.

McCall, M. W. (1997). *High Flyers: Developing the Next Generation of Leaders.* Boston, MA: Harvard Business School.

Meiring, D., Van der Vijver, A. J. R., Rothmann, S., & Barrick, M. R. (2005). Construct, item, and method bias of cognitive and personality tests in South Africa. *South African Journal of Industrial Psychology, 31*(1), 1-8.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1-18.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Pallant, J. (2011). *SPSS survival manual: A step by step guide to data analysis using SPSS* (4th ed.). Crows Nest, Australia: Allen & Unwin.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*(2), 381-391.

Petrillo, J. P., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health, 18,* 25-34.

Phillips, S. E., (1986). The effects of the deletion of misfitting persons on vertical equating via the Rasch model. *Journal of Educational Measurement, 23*(2), 107-118.

Porteus, S.D. (1937). Primitive intelligence and environment. New York: Macmillan

Price, L. R. (2017). *Psychometric methods: Theory into practice*. New York, NY: The Guilford Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Reuning, H. & Wortley, W. (1973). Psychological studies of the Bushmen. *Psychologia Africana, Monograph Supplement, 7*(113), 1−113.

Rothmann, S., & Coetzer, E. P. (2003). The big five personality dimensions and job performance. *SA Journal of Industrial Psychology, 29*(1), 68-74.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*(2), 93-103.

Roodt, G., & de Kock, F. (2018a). Reliability: Basic concepts and measures. In Foxcroft & Roodt (Eds.), *Introduction to psychological assessment in the South African context* (5th ed., pp. 59-68). Cape Town, South Africa: Oxford University Press.

Roodt, G., & de Kock, F. (2018b). Validity: Basic concepts and measures. In Foxcroft & Roodt (Eds.), *Introduction to psychological assessment in the South African context* (5th ed., pp. 69-81). Cape Town, South Africa: Oxford University Press.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Education and Psychological Measurement, 66*(1), 63-84.

doi:10.1177/0013164404273942.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120. doi:10.1007/s11336-008-9101-0.

Smiley, J. (2015). Classical test theory or Rasch: A personal account from a novice user. *Shiken, 19*(1), 16-31.

Smit, G. J. (1996). *Psychometrics: Aspects of measurement*. Pretoria: Kagiso.

Smith, R. M. (2004a). Fit analysis in the latent trait measurement models. In Smith & Smith (Eds.), *Introduction to Rasch Measurement* (pp. 73-92). Maple Grove, MN: JAM Press.

Smith, E. V. (2004b). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In Smith & Smith (Eds.), *Introduction to Rasch Measurement* (pp. 93-122). Maple Grove, MN: JAM Press.

Spence, B. A. (1982). *A psychological investigation into the characteristics of black guidance teachers*. Unpublished Master's dissertation, University of South Africa, Pretoria.

Statistics South Africa (2019, July). *Mid-year population estimates.* Department of statistics, Republic of South Africa.

Steyn, D. W. (1977). The construction and evaluation of a South African personality questionnaire. Unpublished Master's dissertation, University of Pretoria, Pretoria.

Taylor, T. R. (1987). *Test bias: The roles and responsibility of test users and test publishers*. Pretoria: Human Sciences Research Council.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism, 57*(8), 1358-1362. DOI: 10.1002/art.23108.

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J., Slade, A., Lawton, G., Simone, A., Carter, J., Lundgren-Nilsson, Å, Tripolski, M., Ring, H., Biering-Sørensen, F., Marincek, Č, Burger, H., & Phillips, S. (2004). Assessing and Adjusting for Cross-Cultural Validity of Impairment and Activity Limitation Scales through Differential Item Functioning within the Framework of the Rasch Model: ThePRO-ESOR Project. *Medical Care, 42*(1), 137-148. DOI: 10.1097/01.mlr.0000103529.6313277

Tomyn, A. J., Stokes, M. A., Cummins, R. A., & Dias, P. C. (2020). A Rasch analysis of the Personal Well-being Index in school children. *Evaluation & the Health Professions, 43*(2), 1-22. DOI: 10.1177/0163278718819219

Traub, R., E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8-14.

Tredoux, N. (2013). Using computerized and internet-based testing in South Africa. In, Laher & Cockcroft (Eds.) *Psychological assessment in South Africa: Research and applications*. Johannesburg, South Africa: Wits University Press.

Van der Vijver, A. J. R., & Rothmann, S. (2004). Assessment in multicultural groups: The South African case. *South African Journal of Industrial Psychology, 30*(4), 1-7.

Wang, G., & Netemeyer, R. G. (2002). The effect of job autonomy, customer demandingness, and trait competitiveness on salesperson learning, self-efficacy, and performance. *Journal of the Academic Study of Marketing Science, 30*(3), 217-227. https://doi.org/10.1177/00970302030003003

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, Il: MESA Press.

Wright, B. D., & Mok, M. C. M. (2007). An overview of the family of Rasch measurement models. In Smith & Smith (Eds.), *Introduction to Rasch Measurement* (pp. 1-24). Maple Grove, MN: JAM Press.

Wright, B. D., & Stone, M. H., (1979). *Best test design*. Chicago, Il: MESA Press.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient of the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377-392. DOI:10.1177/0734282911406668

Zilstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioural Statistics, 36*(2), 186-212. DOI: 10.3102/1076998610366263

Zumbo, B. D., (1999). *A handbook on the theory and methods of Differential Item Functioning (DIF): Logistic Regression Modelling as a unitary framework for binary and Likert-Type (Ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendices

Appendix A: Permission to use HPTi secondary data from Thomas International

thomas

DATE: 27/4/20

TO WHOM IT MAY CONCERN:

SECONDARY DATA PERMISSION LETTER: DAVID SANELE SEMMELINK (9207075219081)

I hereby confirm that David Sanele Semmelink (9207075219081) is permitted to make use of data obtained through Thomas International Ltd., for the sole purpose of completing his Master of Arts research mini-dissertation at the University of Pretoria. The data requested for this research contains completed and anonymised data of the psychometric instrument: High Potential Trait Indicator (HPTi); as well as biographical information, of individuals who have completed the HPTi through the South African division, Thomas SA. The released dataset contains only individuals who have indicated their voluntary participation in future research projects, an option presented upon the completion of the HPTi.

Please contact our offices should you have any queries:

t +44 20 7257 2815

e stephenc@thomas.co.uk

Regards,

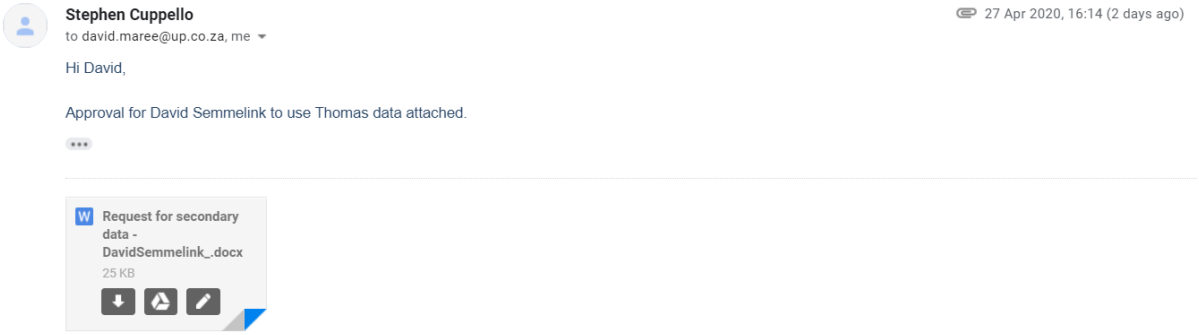
Stephen Cuppello

_____

Stephen Cuppello

Senior Psychologist

Appendix B: Email regarding permission sent to Supervisor



Stephen Cuppello
to david.maree@up.co.za, me

27 Apr 2020, 16:14 (2 days ago)

Hi David,

Approval for David Semmelink to use Thomas data attached.

Request for secondary data - DavidSemmelink_.docx
25 KB

Appendix C: Statement of ethical approval from the Research Committee

22 June 2020

Dear David Sanele Semmelink
Copy to David Maree

Re: MA Proposal

The committee meeting held on the 13th of July 2020 – 20th of July 2020 reviewed your proposal entitled ***"A Rasch analysis of the High Potential Trait Indicator (HPTi): A South African sample."*** and came to the decision that the proposal can move forward once minor changes have been made.

Specific comments made by the committee are as follows:
Both Prof Steyn and Dr. Rawatlal appreciate the effort David has put into the editorial aspects as well as the conceptual depth of the study. No changes are suggested, and they wish David and his supervisor well on this research journey.

Submission to Faculty Ethics approved unconditionally.

If you have any further queries please contact the Chair, Prof David Maree on psychology.rescom@up.ac.za.

Yours sincerely,

Regards

Prof David Maree
Chair: Research Committee
Department of Psychology