Test mode equivalence in a South African personality context:

Paper-and-pencil vs computerised testing

Megan Helene Lubbe, Hons BA

Mini-dissertation as partial fulfillment for the degree Magister Artium in Industrial Psychology
at the North-West University, Potchefstroom Campus

Supervisor: Dr J A Nel

September 2012

Potchefstroom

Declaration of originality of research

DECLARATION

I, Megan Helene Lubbe, hereby declare that *Test mode equivalence in a South African personality context: Paper-and-pencil vs computerised testing* is my own work and that views and opinions expressed in this study are those of the author and relevant literature references as shown in the references. I also declare that the content of this research will not be handed in for any other qualification at any other tertiary institution.

_____

MEGAN HELENE LUBBE

SEPTEMBER 2012

COMMENTS

The reader is reminded of the following:

- The editorial style as well as the references referred to in this mini-dissertation follow the format prescribed by the Publication Manual (4th edition) of the American Psychological Association (APA). This practice is in line with the policy of the Programme in Industrial Psychology of the North-West University to use APA in all scientific documents as from January 1999.
- The mini-dissertation is submitted in the form of a research article. The editorial style specified by the South African Journal of Industrial Psychology (which agrees largely with the APA style) is used, but the APA guidelines were followed in constructing tables.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to the following persons for their support and guidance and making the completion of this mini-dissertation possible:

- My Heavenly Father for blessing me with talents and opportunities and for granting me patience and perseverance throughout this challenging process.

- My wonderful husband Adriaan for his loving support, patience, consistent encouragement, and for always believing in me.

- Dr Alewyn Nel, my supervisor, for his guidance, patience, time and effort to support me through this process.

- Tom Larney, for the professional manner in which he conducted the language editing.

- All the collaborators on the SAPI project.

- My parents for their love and support throughout my many years of study.

- The financial assistance of the National Research Foundation (NRF) towards this research is also acknowledged.

TABLE OF CONTENTS

LIST OF TABLES

SUMMARY

Title: Test mode equivalence in a South African personality context: Paper-and-pencil vs computerised testing

Keywords: Personality measurement, computerised testing, paper-and-pencil testing, reliability, validity, test mode equivalence, South African Personality Inventory (SAPI)

The use of computers in testing has increased dramatically in recent years. Since its first introduction in the fields of education and psychological assessment, the popularity of computer-based testing (CBT) has increased to such an extent that it is likely to become the primary mode of assessment in the future. The shift towards CBT, although successful, has raised many practical, ethical and legal concerns. Due to the potential influence of differential access to computers and varying levels of computer familiarity amongst South African test-takers, it has become increasingly important to study the effect of different modes of administration on test performance.

The objective of this study is to determine whether traditional paper-and-pencil and computerised assessment measures will lead to equivalent results when testing facets of personality on a dichotomous (2 point) rating scale. A cross-sectional survey design was used. A non-probability convenience sample was drawn from university students in South Africa (N = 724). The sample included undergraduate students from two higher education institutions in South Africa. A 48 item behaviour questionnaire measuring facets from the soft-heartedness personality construct was administered. Participants completed the questionnaire either in a paper-and-pencil or in a computerised format. Apart from the difference in administration mode the questionnaires were made to look as similar as possible in all other aspects, such as the number of questions per page, colour, numbering of questions, etc. to minimise the possibility of scoring differences due to other factors. The paper-and-pencil and computerised formats were then factor-analysed and subjected to correlation analysis. The two test modes were also tested for reliability using the Cronbach Alpha coefficient. The obtained results were then used to determine whether equivalence exists between the different modes of administration.

Results indicated that the psychometric functioning of the traditional paper-and-pencil test mode is superior to that of the computerised version. The paper-based test consistently outperformed its computer-based counterpart in terms of mean scores, skewness, kurtosis, factor loadings, inter-item correlations and reliability.

Recommendations were made for future research.

OPSOMMING

Titel: Toets-tipe gelykheid in 'n Suid-Afrikaanse persoonlikheidskonteks: Papier-en-potlood teenoor rekenaargebaseerde toetsing.

Sleutelterme: Persoonlikheidsmeting, rekenaargebaseerde toetsing, papier-en-potlood toetsing, betroubaarheid, geldigheid, toets-tipe gelykheid, SAPI

Die gebruik van rekenaars in assessering is drasties aan die toeneem. Sedert die eerste bekendstelling van rekenaars in die velde van onderrig en psigologiese toetsing het die gewildheid van rekenaargebaseerde toetsing so gegroei dat dit in die toekoms moontlik die primêre toetsmetode sal word. Alhoewel die skuif na rekenaargebaseerde toetsing as suksevol beskou kan word, bring dit ook menige praktiese, etiese en wetlike kwessies na vore. As gevolg van die potensiële invloed van ongelyke toegang tot rekenaars en verskillende vlakke van rekenaarvertroudheid onder Suid Afrikaanse toetsnemers, word dit al hoe meer belangrik om die effek van verskillende afneemmetodes op toetsprestasie te bestudeer.

Die doelwit van hierdie studie is om te bepaal of tradisionele papier-en-potlood toetse en rekenaargebaseerde toetse tot gelyke resultate lei wanneer fasette van persoonlikheid op 'n digotome (2 punt) metingskaal getoets word. 'n Dwarssneeopname-ontwerp is gebruik in die studie. 'n Niewaarskynlikheid-geskiktheidsteekproefneming is geneem van 'n aantal universiteitstudente in Suid Afrika (N = 724). Die steekproef sluit voorgraadse studente van twee hoër onderrig institusies in Suid-Afrika in. 'n 48-item gedragsvraelys wat fasette van die saggeaardheid-persoonlikheidskonstruk meet, is afgeneem. Deelnemers het die vraelys of in 'n papier-en-potlood formaat of in 'n rekenaargebaseerde formaat voltooi. Behalwe vir die verskil in afneemmetode is die vraelyste gemaak om so identies as moontlik te lyk in alle ander aspekte, soos aantal vrae per bladsy, kleur, numering van vrae, ens. om te verhoed dat verskille in telling veroorsaak kan word deur ander faktore. Die papier-en-potlood en rekenaargebaseerde weergawes is gefaktor-analiseer en onderwerp aan korrelasie-analise. Die twee toetstipes is ook getoets vir betroubaarheid deur gebruik te maak van die Cronbach Alpha koeffisiënt. Die resultate wat verkry is, is gebruik om vas te stel of toets-tipe gelykheid bestaan tussen die verskillende toetsmetodes.

Die resultate wys dat die psigometriese funksionering van die tradisionele papier-en-potlood toetstipe beter is as dié van die rekenaargebaseerde weergawe. Die papiergebaseerde toets het in terme van mediaantellings, skeefheid, kurtose, faktorbelading en betroubaarheid deurlopend beter gevaar as die rekenaargebaseerde weergawe. Aanbevelings vir toekomstige navorsing word gemaak.

CHAPTER 1

INTRODUCTION

This mini-dissertation focuses on the psychometric equivalence between paper-based and computer-based versions of a South African personality questionnaire. Chapter 1 contains the problem statement and a discussion of the research objectives in which the general objectives and specific objectives are set out. The research method is explained as well as the division of chapters.

1.1 PROBLEM STATEMENT

The use of computers in testing has increased dramatically in recent years (Booth-Kewley, Edwards & Rosenfeld, 1992; Buchanan et al., 2005a; Davies, Foxcroft, Griessel & Tredoux, 2005; Foxcroft & Davies, 2006; Joubert & Kriek, 2009; Murphy & Davidshofer, 2005; Vispoel, Boo & Bleiler, 2001; Wang, Jiao, Young, Brooks, & Olson 2008). Since its first introduction in the fields of education and psychological assessment, the popularity of computer-based testing (CBT) has increased to such an extent that it is likely to become the primary mode of assessment in the future (Davis, 1999; Vispoel et al., 2001; Wang et al., 2008). The popularity of CBT can be attributed to the various unique advantages that this mode of administration holds. Commonly cited advantages of computer-based testing include increased standardisation, reductions in time and cost, increased accuracy in scoring, wider accessibility, more complete and accurate data reports, and the almost instant scoring and interpretation of results (Davies et al., 2005; Mead & Drasgow, 1993; Tippins et al., 2006).

Despite the many advantages that CBT offers over traditional paper-and-pencil testing (PPT), assessment experts, researchers, practitioners and users have raised questions about the comparability of scores between the two modes of administration (Rosenfeld, Booth-Kewley, & Edwards, 1996). Further concerns have been raised regarding the use of CBT where there is unequal access to computers and technology (Foxcroft & Davies, 2006). Various researchers have conducted studies investigating the equivalence between PPT and CBT. A literature review of previous research appears to indicate that measurement equivalence between traditional PPT and CBT is generally established (Bartram & Brown, 2004; Holtzhausen, 2004; Joubert & Kriek, 2009; Pouwer, Snoek, Ploeg, Heine & Brand, 1998; Simola & Holden, 1992; Vispoel et al., 2001; Wang et al., 2008). However, researchers caution that the translation of paper-based questionnaires to a computer-based format represents a significant change in measurement which could affect reliability and result in inequivalent scores (McDonald, 2002; Webster & Compeua, 1996). According to Suris, Borman, Lind and Kasher (2007), these differences can be attributed to the differences in presentation mode and response requirements between the two test modes. Equivalence

between different modes of administration should therefore be proven and not assumed. Section 2c of the International Test Commission's Guidelines (2005:11) states that "where the CBT/Internet test has been developed from a paper-and-pencil version, ensure that there is evidence of equivalence".

Personality measurement in South Africa

According to Foxcroft, Roodt and Abrahams (2005), psychological tests in South Africa (and internationally) were developed in response to a growing public need. The use of psychological tests in industry gained popularity after World War II and the inauguration of the Nationalist Government in 1948. Advances in technology have since impacted greatly on the ease and sophistication of the testing process. Technical innovations such as the high-speed scanner in the 1950's and 1960's increased the use of interest questionnaires and personality tests in particular (Davies et al., 2005).

Psychological assessment measures in South Africa were developed in an environment where large inequalities existed in the distribution of resources between racial groups. According to Foxcroft et al. (2005) it was therefore almost inevitable that the development of psychological assessment reflected the racially segregated society in which it evolved. The majority of personality inventories in use in South Africa are Westernised measures, imported from Europe or the United States and then translated into either English or Afrikaans (Nel, 2008). When such tools are used to assess non-whites without being adapted to the specific population group, issues of bias are often raised. In the past, previously disadvantaged people (Africans, Coloureds and Indians) were not protected from unfair discrimination, which was often worsened by the misuse of tests and test results (Abrahams, 1996). Selection and promotion decisions were often made on the basis of tests which have not been proven to be comparable across different racial and language groups. However, developments in South African labour legislation, and in particular the Employment Equity Act 55 of 1998 (EEA), now compel test validation, specifically in industry (Joseph & Van Lil, 2008). The Employment Equity Act No.55 of 1998 (Section 8) states that: Psychological testing and other similar forms of assessments of an employee are prohibited unless the test or assessment being used (a) has been scientifically shown to be valid and reliable; (b) can be applied fairly to all employees; (c) is not biased against any employee or group (Davies et al., 2005).

In response to the bias and poor item functioning of personality assessment measures used in South Africa (Nel, 2008), researchers are currently in the process of developing the South African Personality Inventory (SAPI), with the aim of fairly and effectively measuring personality across the eleven official language groups. According to Holzhausen (2005:6), fairness is not only a legal issue but also "a social issue where group differences and equitable treatment of all test-takers should be considered and a psychometric issue, where psychological assessments or the decisions derived from them cannot be

considered fair if we cannot rely on the information produced by them". Joseph and Van Lil (2008) state that large inequalities still exist in South Africa's social and economic structure, and that variables such as language, race, social and educational background are therefore likely to influence an individual's test performance. Specific reference should therefore be made to the reliability and validity of psychological tests (Holzhausen, 2005). According to Zieky (2002), "fairness" is a complex concept and can therefore not be proven by a single statistical method. The best way to ensure test fairness is therefore to build fairness into the development, administration and scoring processes. Due to the growing popularity of CBT it is becoming increasingly important to study the effects of the administration mode on the fairness and stability of occupational assessments (Holzhausen, 2005).

Paper-and-pencil versus computerised testing

Paper-and-pencil and computerised assessments are undoubtedly the most popular forms of test administration currently being used in the fields of education and psychometrics. In a paper-and-pencil assessment test takers are required to "make a verbal or written response to a stimulus presented on paper or a verbal stimulus given by the test administrator" (Suris et al., 2007, p98). Paper-and-pencil tests can be administered in individual or group settings (Murphy & Davidshofer, 2005) and are usually administered under standardised conditions in the presence of a proctor or assessment practitioner (Foxcroft, Roodt & Abrahams, 2005). Although PPT remains the most familiar mode of administration in a variety of test settings (Foxcroft & Davies, 2006), researchers, test developers and administrators are seeing a definite shift towards computerisation. According to Bartram and Brown (2004) all the personality inventories commonly used in occupational assessment are being made available in computerised format or on the Internet.

CBT refers to selection instruments that are administered and scored via a computer (Davies et al., 2005; Tippins et al., 2006). CBT can be administered via computer in offline settings, in network configurations, or on the Internet (Wang et al., 2008). Until the 1980's, the role of the computer in testing was restricted mainly to recording answers and computing test scores (Davies et al., 2005). However, the emergence of more advanced computer technology in the 1980s and the possibilities this presented for the design, administration and scoring of tests, led to computers becoming a fundamental part of the testing process (Davies et al., 2005). CBT has become such an integral part of the testing process that researchers are of the opinion that CBT is likely to replace PPT in the future (Davis, 1999; Vispoel et al., 2001; Wang et al., 2008).

The computer-based tests currently in use are mostly computer-adapted versions of existing paper-and-pencil tests (Davies et al, 2005). This is especially true in developing countries such as South Africa.

Many of the computer-based tests being used in South Africa started off as paper-and-pencil tests which were adapted to a computerised format when appropriate technology became available. Examples of such tests include the following: Occupational Personality Questionnaire (OPQ), 15FQ+ Questionnaire, Myers-Briggs Type Indicator (MBTI), Jung Type Indicator (JTI), Customer Contact Style Questionnaire (CCSQ), Occupational Interest Profile (OIP), Occupational Personality Profile (OPP), General and Graduate Reasoning Tests, and Critical Reasoning Test Battery (Davies et al., 2005).

Both paper-and-pencil and computerised assessments have unique advantages and disadvantages. The advantages of PPT include, amongst others, accessibility where there are no computer facilities, standardised testing conditions, and direct supervision (Davies et al., 2005). A proctor or assessment practitioner is present during the testing session to provide instructions, motivate test takers and manage any irregularities which may arise (Griessel, 2005). In individual or small group testing, the assessment practitioner may keep a close record of the test takers' behavior during the assessment session. The information gained from the practitioner's observations is then used to support and better understand the assessment findings (Griessel, 2005). Researchers, however, suggest that a paper-and-pencil format may not be ideal for all testing situations. The use of standard PPT to report sensitive information such as drug use or sexual behaviour has been criticised for various reasons (Bates & Cox, 2008; Bressani & Downs, 2002). Bressani and Downs (2002) is of the opinion that face-to-face or written assessments are more intimidating and that test takers may be fearful of negative reactions. According to Bates and Cox (2008) paper-and-pencil assessments are notorious for eliciting exaggerated and conflicting responses. Bates and Cox (2008) further identify three important pragmatic issues arising from the paper-and-pencil administration mode: (1) paper-and-pencil tests are rigid in terms of question-selection; (2) paper-and-pencil tests are time inefficient; and (3) paper-and-pencil tests are costly with regard to the printing, transporting and processing costs involved. One solution to these pragmatic issues is to make use of computerised or Internet data collection (Bates & Cox, 2008).

CBT has enhanced the efficiency of testing in various ways. According to Davies et al. (2005) advantages of computerised tests include the following: computers allow for objective testing in terms of eliminating the potential biasing effect of the assessment practitioner; printing costs are eliminated and fewer assessment practitioners are needed, making computerised tests cost effective and less labour intensive; computer-based testing allows for rapid feedback to test takers as well as test administrators by providing instant feedback and comprehensive data reports; and computer-based tests also allow for wider distribution and increased accessibility by making tests available via the internet (Davies et al., 2005; Tippins et al., 2006). In addition, data is automated, reducing the risk of transcription and coding errors, and there are no missing data or out of range responses (Cronk & West, 2002; Mead & Drasgow, 1993). In terms of test mode preference, Foxcroft, Watson and Seymore (2004) found that most test takers,

despite their level of computer familiarity, respond favourably to CBT. Similar results were reported by Vispoel et al. (2001). In a study by Vispoel et al. (2001), participants reported that the computerised version was more enjoyable and comfortable and that it was easier and less fatiguing to use, even though it took more time to complete than the paper-based version.

Despite the various advantages of CBT, this mode of administration also holds a number of unique challenges. The increased use of computers in testing has raised many legal and ethical concerns (Davies et al., 2005). As a result, various guidelines regarding the appropriate and ethical practice of CBT have been published. The International Test Commission's Guidelines for Computer-Based and Internet-Delivered Testing (2005) represent the most recent attempt to provide test developers, administrators, publishers and users with such guiding principles (Foxcroft & Davies, 2006). Based on the ITC's guidelines, the Health Professions Council of South Africa has released their own South African Guidelines on Computerised Testing specifically adapted to match the unique legislative and regulatory practices in South Africa (HPCSA, 2012). According to the ITC's Guidelines (ITC, 2005), the major issues related to CBT include computer hardware and software technology, test materials and testing procedure quality, control of the test delivery, test-taker authentication, prior practice, security issues of testing materials, privacy, data protection, and confidentiality. The unsupervised setting of the CBT can also cause additional problems such as cheating, unstandardised test conditions, and poor response rates. When computerised tests involve the Internet, disrupted connections to the Internet may result in testing being interrupted (Davies et al., 2005; Tippins et al., 2006). Furthermore, Foxcroft and Davies (2006), stress the importance of monitoring practitioners' perceptions of CBT in developing countries where PPT still predominates. Due to low levels of computer familiarity, many practitioners may be sceptic about the use of CBT in practice. In a recent study conducted in South Africa, psychologists indicated that they felt threatened by the increasing use of computer-based tests in South Africa, given their own low levels of computer familiarity and lack of training in computer-based testing (Foxcroft, Patterson, Le Roux & Herbst, 2004).

Questions are also arising regarding the fairness with which CBT can be applied with technologically unsophisticated test-takers (Foxcroft, Seymore, Watson & Davies, 2002). Studies have shown that CBT may lead to increased anxiety levels, especially amongst older test-takers where higher levels of computer illiteracy can be found and, in turn, have a negative impact on test performance (Davies et al., 2005; Foxcroft, Watson & Seymore, 2004). Foxcroft et al. (2004) thus concluded that, given the relationship between computer familiarity and anxiety, it was important to offer test takers with low levels of computer familiarity the alternative of doing a paper-based test, or, if they preferred to do the computer-based test, they should be given a more extensive introductory tutorial and be debriefed afterwards. In

such cases where different test modes are used interchangeably, equivalence of the score distributions across modes needs to be established (Vispoel et al., 2001).

Establishing test mode equivalence

As far back as 1968, researchers at the Computer-Assisted Testing Conference in the United States (USA) expressed concerns that medium effect size differences might exist between the means on different test modes (Mead & Drasgow, 1993). Guideline 22 of the ITC's Guidelines on Computer-Based and Internet-Delivered Testing state that when a test has both paper-based and computer-based versions, test developers need to document evidence of their equivalence (ITC, 2005). The American Psychological Association's Guidelines for Computer-Based Tests and Interpretations (1986) also emphasise the importance of score equivalence. The American Psychological Association (APA, 1986:18) defines score equivalence between PPT and CBT as follows:

> Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode.

The HPCSA's South African Guidelines on Computerised Testing further states that it should be shown that the two versions have comparable reliabilities, correlate with each other at the expected level from the reliability estimates, correlate comparably with other tests and external criteria, and produce comparable means and standard deviations or have been appropriately calibrated to render comparable scores (HPCSA, 2012, p.13).

Lievens and Harris (2003) note that initial evidence seems to indicate that measurement equivalence between CBT and PPT is generally established. A review of previous studies investigating the equivalence between PPT and CBT seems to support this view and suggests that computerised psychological assessments can have satisfactory psychometric properties and can measure the same constructs as traditional versions (Bartram & Brown, 2004; Holzhauzen, 2005; Joubert & Kriek, 2009; Mead & Drasgow, 1993; Pouwer et al., 1998; Simola & Holden, 1992; Vispoel, Boo & Bleiler, 2001; Wang et al., 2008). Despite the growing evidence in support of test mode equivalence between scores from PPT and CBT (Buchanan, Ali et al., 2005), equivalence cannot be taken for granted in all cases. According to Kim and Huynh (2008), past research findings on the equivalence of scores from PPT and CBT are inconsistent. A number of studies have shown differences between paper-and-pencil and computerised versions of the same tests in terms of both the score distributions achieved and the psychometric properties of the tests (Buchanan, Ali et al., 2005).

In a meta-analysis of 28 studies, Mead and Drasgow (1993) found no significant effects of test mode on performance for carefully constructed power tests but a substantial effect for speeded tests. Buchanan, Johnson et al. (2005), working with an on-line version of a 5-factor personality inventory, found that the latent structure of the inventory appeared to have changed slightly: a small number of items loaded on factors other than those they had loaded on in the off-line development sample. Buchanan, Ali et al. (2005) compared the equivalence of on-line and paper-and-pencil versions of the prospective memory questionnaire. The PMQ has four-factor analytically derived subscales. In a large sample tested via the Internet, only two factors could be recovered; the other two subscales were essentially meaningless. This demonstration of non-equivalence underlines the importance of computerised test validation. Without examining the psychometric properties of a test, one cannot be sure that a test administered in computerised format actually measures the intended constructs.

According to Buchanan, Johnson et al. (2005), characteristics of the testing medium, such as anonymity and the use of computer-mediated communication, can also result in phenomena such as disinhibition effects or reduced socially desirable responding. Joinson (1999) found lower levels of socially desirable responding in the on-line condition when administering a social desirability questionnaire to students in either on-line or off-line testing conditions. Joinson's finding suggest that individuals are likely to disclose more about themselves in online questionnaires than in face-to-face interviews, and thus may respond more truthfully to personality questionnaires. However, contrasting findings were reported by Rosenfeld et al (1996). In their study, Rosenfeld et al. (1996) found higher levels of impression management in the computer-linked condition. It was concluded that perceiving that one's responses are linked to a larger database may lead to more impression management on computer surveys. Cronk and West (2002) support this finding by stating that the problem of confidentiality needs to be considered in Internet testing. Participants may feel uncomfortable providing information over the Internet, because they believe that others may use the results. They may respond differently than they would if they were certain their responses would be anonymous.

Foxcroft and Davies (2006) further stress the need to not only establish equivalence between PPT and CBT, but also to examine the impact of differential access to computers and technology on test performance. According to Foxcroft et al. (2004), there is evidence indicating that CBT has an adverse impact on the test performance of individuals with lower levels of technological sophistication. Although increased exposure to computer and information technology in the South African society means that the adverse impact on computer-based test performance is probably diminishing (Foxcroft et al., 2004), the possibility should not be overlooked. Guideline 33 of the ITC's Guidelines (2005) thus indicates that alternative methods of testing should be considered in instances where there is unequal access to

computers and technology. In developing countries such as South Africa, it might therefore be preferable to make psychometric instruments available in both paper-and-pencil and computerised formats.

From the problem statement it has become apparent that a need exists for psychometric assessment measures that are valid, reliable and can be applied fairly in the multicultural South African context. One step towards developing such an instrument in South Africa is determining the optimal mode of administration in which the assessment measure should be presented. Research findings suggest that it may be preferable to make the South African Personality Inventory (SAPI) available in both paper-and-pencil and computerised formats. Through the use of interchangeable test modes, future test administrators may ensure that the choice of test mode is aligned with the socio-economic status and educational levels of the testee. Establishing equivalence between PPT and CBT would allow test administrators to use the two modes interchangeably.

The following research questions emerge from the problem statement:

- How is the equivalence between paper-and-pencil and computerised assessment measures conceptualised according to literature?
- To what extent is paper-and-pencil and computerised assessment measures equivalent?
- How does the reliability and validity of paper-and-pencil and computerised assessment measures compare?
- What recommendations can be made for future research?

1.2 RESEARCH OBJECTIVES

The research objectives are divided into a general objective and several specific objectives.

1.2.1 General objective

To determine whether traditional paper-and-pencil and computerised assessment measures will lead to equivalent results when testing facets from the soft-heartedness personality cluster on a dichotomous rating scale.

1.2.2 Specific objectives

The specific objectives of this study are:

- To determine how the equivalence between paper-and-pencil and computerised assessment measures is conceptualised according to the literature.
- To determine equivalence using paper-and-pencil and computerised assessment measures.
- To compare the reliability and validity of paper-and-pencil and computerised assessment measures.
- To make recommendations for future research.

1.3 RESEARCH METHOD

The research method will consist of a literature review and an empirical study (quantitative research).

1.3.1 Literature review

The literature review will be conducted by making use of databases such as *Academic Search Premier, EBSCO Host, SAe Publications, Science Direct, and Emerald Online*. The most recently published relevant articles will be identified. Relevant journals such as *South African Journal of Industrial Psychology, South African Journal of Psychology, Journal of Occupational and Organizational Psychology, Personnel Psychology, Behavior Research Methods, Journal of Personality Assessment, International Journal of Selection and Assessment, and International Journal of Testing* will be consulted in the search. The aim of the literature review will be to explore and understand current issues relating to personality testing in the South African context. The review will also be aimed specifically at investigating existing research regarding the use of computerised assessment measures and the equivalence between paper-and-pencil and computer-based modes of administration. Furthermore, the literature study will motivate the need for establishing test mode equivalence between paper-and-pencil and computerised modes of administration.

1.3.2 Empirical Study

The empirical study consisted of the research design, the research participants and the measuring instruments.

1.3.2.1 Research Design

The study will be quantitative in nature. A quantitative design is used to "predict, describe and explain quantities, degrees and relationships" (Du Plooy, 2002:82). Findings from quantitative studies are

generalised from a sample to the wider population by collecting numerical data. For the purpose of this study, a cross-sectional design will be used, meaning that the sample will be drawn from a population at a single point in time (Du Plooy, 2002). Information collected is used to describe the population at a specific point in time. The primary goal of such research is to assess interrelationships among variables within a population and to describe cause-and-effect relationships between observable and tangible variables (Kerlinger & Lee, 2000; Du Plooy, 2002).

1.3.2.2 Research participants

A combination of quota and convenience sampling will be used. A sample will be drawn from university students in South Africa (N = 700/800). The sample will include undergraduate students from two higher education institutions in South Africa. The sample will include both male and female participants from different race and language groups. Although specific inclusion and exclusion criteria will not be followed, the aim will be to include a diverse group of participants which will be representative of the South African population demographics.

1.3.2.3 Measuring instruments

The study makes use of two parallel measuring instruments consisting of items which measure different facets from the soft-heartedness personality construct. The soft-heartedness questionnaire will be presented in both paper-and-pencil and computerised formats. Various facets from the SAPI have been measured for reliability in two independent studies with university students from various industries (Flattery, 2011; Oosthuizen, 2012; Van der Linde, 2012). The facets which generated the highest levels of reliability were then extracted and used to construct a behaviour questionnaire consisting of 48 items. Because the focus of the study is on determining the item functioning between two different modes of administration and not on measuring the overlap between clusters, sub-clusters or facets of personality, it is acceptable to include only certain facets. The facets to be included are "generous", "compassionate" and "appreciative". The following alpha coefficients were attained in the first study: generous (α = 0,83), compassionate (α = 0,87) and appreciative (α = 0,83) (Flattery, 2011). In the second study, the following alpha coefficients were attained: generous (α = 0,77), compassionate (α = 0,88) and appreciative (α = 0,86) (Van der Linde, 2012). Both questionnaires will be on a dichotomous (two-choice) rating scale, consisting only of "agree" or "disagree" response options. Statements will be included such as *"I share what I have with others"* (generous), *"I am sensitive to other people's feelings"* (compassionate), and *"I value life as it is"* (appreciative).

The paper-and-pencil and computerised formats will be compared in terms of factor-loadings with the various facets of reliability, validity and bias to determine whether equivalence exists between the two modes of administration. Besides the difference in administration mode the questionnaires will be made to look as similar as possible in all other aspects such as number of questions per page, colour, numbering of questions, etc. to minimise the possibility of scoring differences due to other factors.

1.4 RESEARCH PROCEDURE

Both versions of the soft-heartedness questionnaire will be administered to undergraduate students. The paper-and-pencil version will be administered to students at the various universities during their normal lecture hours. The times and dates of the assessment sessions will be pre-arranged with the various subject lecturers. The assessments will take place in a controlled classroom setting. Researchers will be present to provide instructions and to supervise the assessment process. The computerised assessments will also take place under supervised test conditions in computer laboratories at the various universities. The computerised version of the questionnaire will be loaded onto a server (eFundi) and students will be granted access to the questionnaire through the use of passwords. Students will then be requested to schedule a time to complete the questionnaire, using extra marks towards their subject grade as incentive to complete the questionnaire. The data obtained from the two different modes of administration will then be factor-analysed and compared in terms of factor loadings, variance explained, validity, and reliability to determine whether equivalence exists between the different test modes.

1.5 STATISTICAL ANALYSIS

Statistical analysis will be carried out with the help of the SPSS-program (SPSS, 2008). Descriptive statistics, including means, standard deviations, range, skewness and kurtosis, and inferential statistics will be used to analyse the data. Factor analysis will be used to measure item correlations with certain facets of soft-heartedness when using two different modes of administration. Factor analysis will be exploratory. Factor analysis and Cronbach alpha coefficients will further be used to assess the reliability and validity of the measuring instrument (Clark & Watson, 1995). Cronbach alpha coefficients are commonly used to measure the internal consistency of a measure. The alpha coefficient provides an indication of the average correlation among all the items that make up a scale (Pallant, 2007). For the purpose of this study, alpha values higher than 0,70 would indicate acceptable levels of reliability.

1.6 ETHICAL CONSIDERATIONS

Fair and ethical research procedures are vital to the success of this study. The following ethical considerations were met:

Permission to perform assessments was obtained from the various universities. Ethical aspects regarding the research were discussed with all participants prior to assessment. Participation in the research was voluntary and participants were required to sign informed consent forms giving researchers permission to use their results for academic purposes. Electronic versions of the informed consent form were completed prior to the computerised assessment. Participants were informed verbally about the purpose of the study. The data obtained from the study was used for academic research purposes only, to address the item functioning and psychometric properties of the questionnaire. Participants therefore remained confidential, were not analysed on their individual personality constructs, and did not receive feedback on their performance. This was communicated to participants in advance.

1.7 DIVISION OF CHAPTERS

The chapters of this mini-dissertation are presented as follows:

Chapter 1: Introduction, problem statement and objectives
Chapter 2: Research article
Chapter 3: Conclusions, limitations and recommendations

1.8 CHAPTER SUMMARY

This chapter discussed the problem statement and research objectives. The measuring instruments and research method used in this research were explained, followed by a brief overview of the chapters that follow.

REFERENCES

Abrahams, F. (1996). *The cross-cultural comparability of the 16PF*. Unpublished doctoral thesis, Department of Industrial Psychology, University of South Africa.

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.

Bartram, D., & Brown, A. (2004). Online testing: Mode of administration and the stability of the OPQ32i scores. *International Journal of Selection and Assessment, 12*(3), 278-284. doi:101111/j.0965-075X.2004.282_1.x

Bates, S. C. & Cox, J. M. (2008). The impact of computer versus paper-pencil survey, and individual versus group administration, on self-reports of sensitive behaviors. *Computers in Human Behavior, 24*, 903-916.

Booth-Kewley, S., Edwards, J. E., & Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology*, 77, 562-566.

Bressani, R. V. & Downs, A. C. (2002). Youth independent living assessment: Testing the equivalence of web and paper/pencil versions of the Ansell-Casey Life Skills Assessment. *Computers in Human Behavior, 18*, 453-464.

Buchanan, T., Ali, T., Heffernan, T. M., Ling, J., Parrott, A. C., Rodgers, J., & Scholey, A. B. (2005). Non-equivalence of online and paper-and-pencil psychological tests: The case for the prospective memory questionnaire. *Behaviour Research Methods, 37*(1), 148-154.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the internet. *European Journal of Psychological Assessment, 21*(2), 115-127. doi:10.1027/1015_5759.18.1.116

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309-319.

Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers, 34*, 177-180.

Davies, C., Foxcroft, C., Griessel, L., & Tredoux, N. (2005). Computer-based and internet delivered assessment. In C. Foxcroft & G. Roodt (Eds.), *An introduction to psychological assessment in the South African context* (2nd ed.), (pp 153-166) Cape Town: Oxford University Press.

Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. *Behavior Research Methods, Instruments & Computers, 31,* 572-577.

Du Plooy, G.M. (2002). *Communication research: techniques, methods and applications*. Lansdowne: Juta

Flattery, A. (2011). *Developing and validating a hostility, gratefulness and active support measuring instrument.* Unpublished master's dissertation. North-West University, Potchefstroom, South Africa.

Foxcroft, C., Roodt, G., & Abrahams, F. (2005). Psychological assessment: A brief retrospective overview. In C. Foxcroft & G. Roodt (Eds.), *An introduction to psychological assessment in the South African context* (2nd ed.), (pp. 8-23). Cape Town: Oxford University Press.

Foxcroft, C. D., & Davies, C. (2006). Taking ownership of the ITC's guidelines for computer-based and internet-delivered testing: A South African application. *International Journal of Testing, 6*(2), 173-180.

Foxcroft, C. D., Paterson, H., Le Roux, N., & Herbst, D. (2004). *Psychological assessment in South Africa: a needs analysis.* Retrieved from www.hsrc.ac.za/.../1716_Foxcroft_Psychologicalassessmentin%20SA.pdf

Foxcroft, C. D., Seymore, B. B., Watson, A. S. R., & Davies, C. (2002). *Towards building best practice guidelines for computer-based and Internet testing.* Paper presented at the 8th National Congress of the Psychological Society of South Africa, University of the Western Cape, Cape Town, 24-27 September 2002.

Foxcroft, C. D., Watson, A. S. R., & Seymore, B. B. (2004b). Personal and situational factors impacting on CBT practices in developing countries. Paper presented at the 28th International Congress of Psychology, Beijing, China, 8-13 Augustus 2004.

Holtzhausen, G. (2004). *Mode of administration and the stability of the OPQ32n: Comparing internet (controlled) and paper-and-pencil (supervised) administration.* Unpublished master's thesis. University of Pretoria, Pretoria, South Africa.

Health Professions Council of South Africa (2012). *South African Guidelines on Computerised Testing.* Retrieved from http://www.hpcsa.co.za/downloads/psychology/Form_257.pdf

International Test Commission (2005). *International guidelines on computer-based and Internet delivered testing.* Retrieved from http://www.intestcom.org

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers, 31*, 433-438.

Joubert, T., & Kriek, H. J. (2009). Psychometric comparison of paper-and-pencil and online personality assessments in a selection setting. *South African Journal of Industrial Psychology, 35(*1), 1-11. doi:10.4012/sajip.v35i1.727

Joseph, L., & Van Lill, B. (2008). Investigating subscale differences among race and language groups on the Occupational Personality Profile. *South African Journal of Psychology, 38*(3), 501-514.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioural research (4th Ed.).* London: Wadsworth. Kim, D., & Huynd, H. (2008). Computer-based and paper-and-pencil administration mode effects on a statewide end-of-course English test. *Educational and Psychological Measurement, 68*(4), 554-570.

King, W. C., & Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80*, 643-651. doi:10.1037/0021-9010.80.6.643

Lahola, P. (2009). *South African statistics, 2009*. Pretoria: StatsSA.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458. doi:10.1037/0033-2909.114.3.449

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Upper Saddle River, NJ: Pearson Education Inc.

Nel, J. A. (2008). *Uncovering personality dimensions in eleven different language groups in South Africa: An exploratory study*. Unpublished doctoral dissertation. North-West University, Potchefstroom, South Africa.

Oosthuizen, T. H. (2012). *Developing and validating a measuring instrument for the relationship harmony personality cluster.* Unpublished master's dissertation. North-West University, Potchefstroom, South Africa.

Pallant, J. (2007). SPSS Survival Manual (3rd ed.). New York, USA: Open University Press.

Pek, J. (2008). *A brief introduction to CEFA*. Retrieved from www.unc.edu/~pek/CEFAQuickStart.pdf

Pouwer, F., Snoek, F. J., Van Der Ploeg, H. M., Heine, R. J., & Brand, A. N. (1998). A comparison of the standard and computerised versions of the Well-Being Questionnaire (WBQ) and the Diabetes Treatment Satisfaction Questionnaire (DTSQ). *Quality of Life Research, 7*, 33-38. doi:10.1023/A:1008832821181

Rosenfeld, P., Booth-Kewley, S., & Edwards, J. E. (1996). Responses to computer surveys: Impression management, social desirability, and the Big Brother syndrome. *Computers in Human Behavior, 12*(2), 263-274.

Siloma, S. K., & Holden, R. R. (1992). Equivalence of computerized and standard administration of Piers-Harris Children's Self-Concept Scale. *Journal of Personality Assessment, 58*(2), 287-294. doi:10.1207/s15327752jpa5802_8

SPSS Inc. (2008). *SPSS 16.0 for Windows*. Chicago, IL: SPSS Inc.

Surís, A., Borman, P. D., Lind, L., & Kashner, T. M. (2007). Aggression, impulsivity, and health functioning in a veteran population: Equivalency and test-retest reliability of computerized and paper-and-pencil administrations. *Computers in Human Behavior, 23*, 97-110.

Tippins, N. T., Beary, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., et al. (2006). Unproctored internet testing. *Personnel Psychology, 59*(1), 189-225. doi:10.1111/j.1744-6570.2006.00909.x

Van der Linde, P. (2012). *South African Personality Inventory: Developing amiability, egoism and empathy scales for a soft-heartedness measuring instrument.* Unpublished master's dissertation. North-West University, Potchefstroom, South Africa.

Vispoel, W. P., Boo, J., & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg Self-Esteem Scale: A comparison of psychometric features and respondent preferences. *Educational and Psychological Measurement, 61*(3), 461-474. doi:10.1177/00131640121971329

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5-24.

Webster, J., & Compeau, D. (1996). Computer-assisted versus paper-and-pencil administration of questionnaires. *Behavior Research Methods, Instruments & Computers, 28*, 567-576.

Zieky, M. (2002). *Ensuring the fairness of licensing tests.* CLEAR Exam Review, Vol xii, 1, p20.

CHAPTER 2

RESEARCH ARTICLE

TEST MODE EQUIVALENCE IN A SOUTH AFRICAN PERSONALITY CONTEXT: PAPER-AND-PENCIL VS COMPUTERISED TESTING

M. H. LUBBE

J. A. NEL

ABSTRACT

The general objective of this study was to determine whether traditional paper-and-pencil and computerised assessment measures will lead equivalent results when testing facets from the soft-heartedness personality cluster on a dichotomous rating scale. A non-probability, convenience sample was drawn from undergraduate university students from two higher education institutions in South Africa (N = 724). The participants varied according to racial and cultural backgrounds. Participants completed either a paper-based (N = 344) or a computer-based (N = 380) version of the same personality scale. Scores obtained from the two test modes were then compared by means of factor analysis, correlation analysis as well as reliability analysis in order to determine whether equivalence exists between the two test modes. It was concluded that the psychometric functioning of the traditional paper-and-pencil test mode is superior to that of its computerised counterpart. These results suggest that it may be preferable to administer the South African Personality Inventory in a paper-based format.

The importance of personality measurement for the prediction of academic and job performance has grown considerably in recent years (La Grange & Roodt, 2000; Van der Walt, Meiring, Rothman, & Barrick, 2002). Although considerable doubt surrounded the importance of personality testing in the past, the relationship between personality traits and job performance has since been well researched and confirmed by numerous researchers (Bartram, 2004; Murphy & Bartram, 2002). In organisations worldwide, personality tests are commonly used to aid in job-related decision-making, selection and classification processes (Goodstein & Lanyon, 1999; Holtzhausen, 2005; Van der Merwe, 2002). Personality measurement also plays an important role outside of the organisational setting in fields such as psychological health, education and research. Due to the nature and consequence of decisions based on personality assessment, the need exists for instruments that accurately measure personality traits and provides valid, reliable and unbiased results for all test takers.

Paper-and-pencil and computerised assessments are undoubtedly the most popular forms of test administration currently being used in the fields of education and psychometrics. However, improvements in computer technology along with increased affordability and access are resulting in a rapid shift towards computerisation (Leeson, 2006; Mills, Potenza, Fremer, & Ward, 2002; Pomplun, Frey, & Becker, 2002). The popularity of computer-based testing can be attributed to the various unique advantages that this mode of administration holds. Commonly cited advantages of CBT include increased standardisation, reductions in time and cost, increased accuracy in scoring, wider accessibility, more complete and accurate data reports, and the almost instant scoring and interpretation of results (Bugbee & Bernt, 1990; Cronk & West, 2002; Davies, Foxcroft, Griessel, & Tredoux, 2005; Foxcroft & Roodt, 2005; Goldberg & Pedulla, 2002; Lancaster & Mellard, 2005; Mead & Drasgow, 1993; Mills et al., 2002; Pomplum et al., 2002; Tippins, Beaty, Drasgow, Gibson, Pearlman, Segall, & Sheperd, 2006; Wise & Plake, 1989).

However, despite the various advantages of computer-based testing, this mode of administration also holds a number of unique challenges. As the use of computers in testing rapidly spreads through both the public and private sectors of society, many practical, legal and ethical concerns are raised (Davies et al., 2005; Leeson, 2006). Concerns are especially being raised regarding the use of computer-based testing where there is unequal access to technology and varying levels of computer familiarity (Barak, 2003; Bennett et al., 2008; Foxcroft & Davies, 2006; Goldberg & Pedulla, 2002). Decades after the abolishment of apartheid, the after-effects of the societal segregation during this historic period remain visible in South Africa. The "gap between those who do and do not have access to computers and the Internet" is referred to as the digital divide (Van Dijk, 2006, p. 178). Van Dijk (2006) demonstrates that while the digital divide is closing in developed countries, the gap is still growing in developing countries such as South Africa. The effect of computer unfamiliarity and computer anxiety on test performance amongst technological unsophisticated test takers therefore remain a major concern in computer-based testing

(Davies et al., 2005; Foxcroft & Davies, 2006; Foxcroft, Seymore, Watson, & Davies, 2002; Foxcroft, Watson, & Seymore, 2004). As a result of such concerns, tests are frequently being made available in both paper-based and computer-based formats to allow test takers to choose their preferred mode of administration. Where different modes of administration are used interchangeably, concerns with regard to test mode equivalence are often raised (APA, 1986; HPCSA, 2012; ITC, 2005; Mead & Drasgow, 1993; Rosenfeld, Booth-Kewley, & Edwards, 1996).

The American Psychological Association (APA, 1986) defines score equivalence between paper-based and computer-based testing as follows:

> Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode (p. 18).

Initial evidence suggests that test mode equivalence is generally established between paper-based and computer-based modes of administration (Arce-Ferrer & Guzman, 2009; Bartram & Brown, 2004; Holtzhausen, 2005; Joubert & Kriek, 2009; Pouwer, Snoek, Ploeg, Heine & Brand, 1998; Salgado & Moscoso, 2003; Simola & Holden, 1992; Vispoel, Boo, & Bleiler, 2001; Wang, Jiao, Young, & Brooks, 2008). However, with numerous studies demonstrating opposing results (Bugbee & Bernt, 1990; Clariana & Wallace, 2002; Mazzeo, Druesne, Raffield, Checketts, & Muelstein, 1991; Pomplun et al., 2002; Russell, 1999), it becomes imperative that equivalence be proven and not assumed.

Another important consideration in the measurement of personality is to identify the optimal response scale that will lead to the most valid and reliable measures of the construct in question (see Vorster, 2011). Response scales are generally broken down into two formats: dichotomous and polytomous rating scales. Dichotomous (or two-point) scales contain two scale categories, while polytomous scales contain three or more scale categories (Alwin, 1992; Comrey & Montag, 1982**;** Cox, 1980). Both scale formats have unique advantages and disadvantages. The choice of scale format therefore often depends on the context in which it will be applied. Polytomous scales, on the one hand, are generally associated with higher levels of reliability, validity and discriminating power (Alwin, 1992; Comrey, 1988; Netemeyer, Bearder, & Sharma., 2003). This is based on the assumption that test takers are better able to select their particular trait and make more accurate ranking decisions than on a dichotomous scale. However, scales containing several response categories may become confusing and difficult to interpret (Busch, 1993; Chen, Lee & Stevenson, 1995). Due to the almost undetectable differences between some of the response

options presented, test takers may find it difficult to select the answer that most accurately represents them.

Dichotomous scales, on the other hand, are commonly associated with greater ease of use and perceptibility for respondents. Two-point scales are also considered to be sufficiently reliable and account for an adequate amount of variance (Busch, 1993; Chen et al., 1995; Jacoby & Mattel, 1971; Netemeyer et al., 2003). In studying the optimal response categories for the SAPI, Vorster (2011) found the two-point scale to be superior to a five-point scale in terms of psychometric functioning and reliability.

The current study forms part of the greater SAPI project (see Nel et al., in press; Valchev et al., in press). The SAPI, an acronym for South African Personality Inventory, is a project that aims to develop an indigenous personality measure for all eleven official language groups in South Africa. The general objective of the study is to determine whether traditional paper-and-pencil and computerised assessment measures will lead to equivalent results when testing facets from the soft-heartedness personality cluster on a dichotomous rating scale. Establishing equivalence is an important part of the test development process as it will allow future test administrators to use to use the different modes of administration interchangeably.

Psychological assessment in South Africa

South Africa is synonymous with diversity. Comprising eleven official language groups and characterised by a diverse array of cultural, racial and socio-economic groups, the South African demographic proposes unique challenges for test adapters and developers in the country (Foxcroft, 2004). One of the major stumbling blocks concerning the use of psychological tests in South Africa is the complexity of tests which may be used across a diversity of linguistic and cultural backgrounds (Foxcroft, 2004; Huysamen, 1996; Nel, 2008). Of foremost concern are the implications of possible discrimination, resulting in the need for test validation within its applied context (Meiring, Van de Vijver, Rothmann, & Barrick, 2005).

Early psychological test use in South Africa was in line with international trends (Van de Vijver & Rothmann, 2004; Foxcroft & Roodt, 2005). The early 1900's saw tests being imported internationally and applied for all population groups without concern for the adaptation or standardisation of norms (Foxcroft, 1997). However, as researchers showed a heightened interest in the educability and trainability of black South Africans in the 1940's and 1950's (Foxcroft & Roodt, 2005), issues of cross-cultural applicability started to arise. Due to the segregation of different race, language and culture groups before the 1994 democratic election, little need existed for "common" measuring instruments. As a result, separate psychological tests were initially constructed for different race and language groups. However,

despite a sound number of tests being developed for white test takers, significantly fewer instruments were developed for blacks, coloureds, and Indians (Owen, 1991). Due to the critical shortage of tests developed for certain cultural groups, the norm was to apply Westernised measures which have been standardised for white test takers amongst other population groups (Foxcroft, 1997).

Further impacting on discriminatory test use in the country was the non-existence of legislation to protect previously disadvantaged people from unfair test practices (Foxcroft & Roodt, 2005). Particularly affected were the educational and organisational sectors where selection and promotion decisions were often made on the basis of tests which have not been proven to be comparable across different race and language groups (Abrahams, 1996). On par with international developments, the late 1980's saw an attention shift towards certain aspects of fairness, bias, and discriminatory practices in psychological assessment (Meiring et al., 2005). Socio-political advances led to the elimination of job reservation on racial grounds and the initiation of racially mixed schools. As a result, industry and educational authorities began to insist on common and unbiased instruments that could be applied fairly to all test takers regardless of race or culture (Claassen, 1995). The need for culturally fair instruments put pressure on test developers to give serious thought to the issue of test bias and to develop psychometric tests with norms not constructed along racial lines (Claassen, 1995; Owen, 1991).

Critical shortages in test development skills in South Africa have, however, resulted in an inclination towards test adaptation as opposed to test development (Foxcroft, 2004; Foxcroft, Paterson, Le Roux & Herbst, 2004). The majority of tests currently in use in South Africa are therefore Westernised measures that have been translated, adapted and normed for the South African context (Foxcroft & Roodt, 2005; Nel, 2008). However, test adaptation often raises concerns about bias, inequivalence and cultural relevance (Foxcroft, 2004). A number of researchers are of the opinion that all cultural groups are not equally represented in the development and standardisation of instruments currently in use in South Africa (Abrahams & Mauer, 1999; Nel, 2008; Retief, 1988).

While the transformation of the South African society greatly impacted on psychological test development in the country, the greatest impact on the assessment industry was undoubtedly made by the inauguration of post-apartheid legislation, more specifically the Employment Equity Act 55 of 1998. Where themes of misuse plagued the South African assessment industry in the past, legislation is now firmly in place to protect the public against abuse. The promulgation for the Employment Equity Act 55 of 1998, Section 8 (Government Gazette, 1998) has resulted in a greater awareness of the cultural appropriateness of psychological instruments and their application amongst culturally diverse populations (Van de Vijver & Rothmann, 2004).

The Act stipulates that:

> "Psychological testing and other similar assessments are prohibited unless the test or assessment being used – (a) has been scientifically shown to be valid and reliable; (b) can be applied fairly to all employees; and (c) is not biased against any employee or group."

The Health Professions Act 56 of 1974 furthermore requires that all psychometric tests be classified by the Professional Board of Psychology (Government Gazette, 2010). The same classification procedures apply to computerised psychological tests. The Health Professions Council of South Africa states that "computerised and Internet-delivered tests should be classified and evaluated by the Psychometrics Committee of the Professional Board for Psychology before they can be sold or used to assess persons" (HPCSA, 2012, p. 4). Through such regulatory practices, South African legislation attempts to ensure professionalism, fairness and equality in test procedures. Given the transformation of the South African society and the integration of schools, universities, the workplace and society in general since 1994 (Paterson & Uys, 2005), there is an urgent need for measuring instruments that meet EEA requirements. Van de Vijver and Rothmann (2004) state that one of the primary goals of test developers and practitioners should be to bring the psychometric practice in South Africa in line with the demands placed by current legislation.

The South African Personality Inventory

The development of the South African Personality Inventory (SAPI) can be seen as an answer to the poor item functioning and test bias often found in Western developed tests when adapted and applied in the South African context (Nel et al., 2012). The SAPI is developed from a foundational level within the South African context and is developed specifically for the diverse South African population. According to Nel (2008), the purpose of the SAPI project is to develop an instrument that is uniquely South African and that provides valid and reliable measurements of personality in the South African context and among the diverse South African population. Due to its origin within the South African context, the instrument will be applicable across all cultural, language, educational and social-economic groups.

The SAPI consists of nine primary personality factors. These factors include Extroversion, Soft-Heartedness, Conscientiousness, Emotional Stability, Intellect, Openness to Experience, Integrity, Relationship Harmony, and Facilitating (Nel, 2008). The current study will focus specifically on the "Soft-Heartedness" personality construct and the impact of test mode effect on the statistical functioning of this construct. Nel (2008, p. 124) defines soft-heartedness as "a feeling of concern for the welfare of someone else (especially someone defenseless), low concern for own interests and welfare, being thankful for others and overall life-being, an actively expressed feeling of dislike of aggressive behaviour, a

compassionate type of person who is understanding and sensitive towards others' feelings, and a concept of community from sub-Saharan Africa, often summarised as *humanity towards others.*" Nel (2008) further relates "Soft-Heartedness" to attributes of goodness, kindness and tenderness – as opposed to ruthlessness, rudeness and unpleasantness.

Although still in its development stage, the Soft-Heartedness personality construct currently consists of six sub-clusters including Amiability, Egoism, Gratefulness, Hostility, Empathy, and Active Support. When compared to internationally developed and widely studied models for personality measurement, the Soft-Heartedness construct shows high correlations with the "Agreeableness" factor from the Five-Factor Model and with Agreeableness versus Anger from the popular HEXACO Personality Model (Nel, 2008). An important step towards the development of a new psychological instrument is determining the optimal mode in which it will be administered.

Computerised vs paper-and-pencil testing

In a paper-and-pencil assessment test takers are required to "make a verbal or written response to a stimulus presented on paper or a verbal stimulus given by the test administrator" (Suris et al., 2007, p.98). Paper-and-pencil tests can be administered in individual or group settings (Murphy & Davidshofer, 2005) and are usually administered under standardised conditions in the presence of a proctor or assessment practitioner (Foxcroft et al., 2005). Paper-and-pencil tests were originally developed in response to a growing demand for large-scale group testing instruments (Foxcroft, Roodt, & Abrahams, 2005). The need to assess large numbers of military recruits during World War I and II in ways that did not require verbal responses, led to the development of paper-based tests in the 1930's and 1940's. Although paper-and-pencil testing remains the most popular and familiar mode of administration in a variety of test settings around the world (Foxcroft & Davies, 2006), the increased affordability and computational ability of modern computers (Leeson, 2006) is causing a definite shift towards computerisation.

Computer-based testing refers to selection instruments that are administered and scored via a computer (Davies et al., 2005; Tippens et al., 2006). Computer-based tests can be administered via computer in an offline setting, in network configurations, or on the Internet (Wang et al., 2008). The use of computers in testing has been reported in the literature since 1963 (King & Miles, 1995). The early 1960's saw the development of the first computer-based interpretation system, followed by the first fully automated assessment process in the 1970's (Foxcroft & Roodt, 2005). The simplest and most common variant of the computer-based test may be referred to as a linear computerised test (Mazzeo et al., 1991). This refers to an already existing, standardised test (usually in paper-based format) which is adapted for computerised administration. In such test adaptations the computerised test version is made to look and feel as similar

to the original paper-based version as possible in order to reduce the possible influence of mode effect on test performance. Test items in paper-based and computer-based assessments are generally identical in terms of sequence and content. The only major difference between the two formats can be found in the test delivery mode (Wang & Shin, 2009).

Until the 1980's, the role of the computer in testing was restricted mainly to recording answers and computing test scores (Davies et al., 2005). However, the emergence of more advanced computer technology in the 1980s and the possibilities this presented for the design, administration and scoring of tests, led to computers becoming a fundamental part of the testing process (Davies et al., 2005). Improvements in the speed and power of modern computers along with increased affordability and access have made large-scale computer testing possible (Leeson, 2006; Mills et al., 2002). As a result of such developments, computers have gained a more prominent role in both public and private sector assessment (Leeson, 2006). Despite the belief that computer-based testing will become the primary mode of assessment in the future (Davis, 1999; Vispoel et al., 2001; Wang et al., 2008), the original paper-and-pencil test remains a familiar and popular choice. It is therefore useful to consider the advantages of both paper-based and computer-based test formats when selecting the optimal mode of administration for a new instrument.

Both paper-and-pencil and computerised assessments have unique advantages and disadvantages. The advantages of paper-and-pencil assessments include, amongst others, accessibility where there are no computer facilities, standardised testing conditions, and direct supervision (Davies et al., 2005). A proctor or assessment practitioner is present during the testing session to provide instructions, motivate test takers and manage any irregularities which may arise (Griessel, 2005). In individual or small group testing, the assessment practitioner may keep a close record of the test takers' behavior during the assessment session. The information gained from the practitioner's observations is then used to support and better understand the assessment findings (Griessel, 2005).

Researchers, however, suggest that a paper-and-pencil format may not be ideal for all testing situations. The use of standard paper-and-pencil testing to report sensitive information such as drug use or sexual behaviour has been criticised for various reasons (Bates & Cox, 2008; Bressani & Downs, 2002). Bressani and Downs (2002) is of the opinion that face-to-face or written assessments are more intimidating and that test takers may be fearful of negative reactions. According to Bates and Cox (2008) paper-and-pencil assessments are notorious for eliciting exaggerated and conflicting responses. Bates and Cox (2008) further identify three important pragmatic issues arising from the paper-and-pencil administration mode: (1) paper-and-pencil tests are rigid in terms of question selection; (2) paper-and-pencil tests are time inefficient; and (3) paper-and-pencil tests are costly with regard to the printing,

transporting and processing costs involved. One solution to these pragmatic issues is to make use of computerised data collection (Bates & Cox, 2008).

Computer-based testing has enhanced the consistency and efficiency of testing in various ways (Tippins et al., 2006). In terms of consistency, computers allow for objective testing by eliminating the potential biasing effect of the assessment practitioner (Foxcroft & Roodt, 2005). In terms of efficiency, computer-based testing significantly simplifies and speeds up the testing process by largely replacing the role of the test administrator (Foxcroft & Roodt, 2005; Tippens et al., 2006). A computer presents items and collects responses resulting in an automated, simple and more efficient testing process (Lancaster & Mellard, 2005). Computer-based testing also reduces the time and expertise needed to administer and score tests (Lancaster & Mellard, 2005), resulting in significant reductions in time between test administration and score reporting (Bugbee & Bernt, 1990; Mills et al., 2002). In addition, data is automated, reducing the risk of transcription and coding errors, and there are no missing data or out of range responses (Cronk & West, 2002; Mead & Drasgow, 1993).

Although the initial setup of computerised test centers are expensive, computer-based testing requires less material, fewer administrators and less training for administrators, making it cost-effective and less labour intensive than paper-based testing (Goldberg & Pedulla, 2002; Mills et al., 2002). Computer-based testing also allows for flexible (even individualised) test scheduling by giving test takers the opportunity to take tests at times and places that suit them (Goldberg & Pedulla, 2002; Tippens et al., 2006). In addition, Internet-delivered testing creates the potential for wider distribution and increased accessibility, thus allowing for large-scale testing across geographically dispersed groups (Davies et al., 2005; Tippens et al., 2006).

Computer-based testing makes it possible to obtain information about test takers that is not available in paper-based testing (eg. time spent per item) (Wise & Plake, 1989; Mills et al., 2002). Examining such previously unavailable information can assist in test interpretation, allowing test practitioners to reach a more in-depth understanding of test performance (Mills et al., 2002). Computer-based testing also provides various opportunities for the enhancement of test batteries through the inclusion of multi-media such as audio and visual elements (Schoech, 2001). In terms of test mode preference, researchers have found that most test takers, despite their level of computer familiarity, respond favourably to computer-based tests (Arce-Ferrer & Guzman, 2009; Barak & Cohen, 2002; Bugbee & Bernt, 1990; Butler, 2003; Foxcroft et al., 2004b; Vispoel et al., 2001). Studies dating back to the 1980's report evidence of test taker preference towards computer-based assessments. Major pulling factors were the instant scoring of tests upon completion as well as the opportunity to schedule your own exam times (Bugby & Bernt, 1990). In a study by Vispoel et al. (2001), participants reported that the computerised test version was more

enjoyable and comfortable than the paper-and-pencil version and that it was easier and less fatiguing to use.

Research from the past two decades suggests that the perceived benefits of computer-based testing outweigh that of paper-based testing. However, despite the various advantages of computer-based testing, this mode of administration also holds a number of unique challenges. As the use of computers in testing rapidly spreads through both the public in private sectors of society, many legal and ethical concerns are raised (Davies et al., 2005; Leeson, 2006). As a result, various guidelines regarding the appropriate and ethical practice of computer-based testing have been published. The International Test Commission's Guidelines for Computer-Based and Internet-Delivered Testing (2005) represent the most recent attempt to provide test developers, administrators, publishers and users with such guiding principles (Foxcroft & Davies, 2006). According to the ITC's Guidelines (ITC, 2005), the major issues related to computer-based testing include computer hardware and software technology, test materials and testing procedure quality, control of the test delivery, test taker authentication, prior practice, security issues of testing materials, privacy, data protection, and confidentiality. The unsupervised setting of the computer-based test can also cause additional problems such as cheating, unstandardised test conditions, and poor response rates. When computerised tests involve the Internet, disrupted connections to the Internet may result in testing being interrupted (Davies et al., 2005; Tippins et al., 2006). As a result of such concerns, the HPCSA specifies that computerised psychological tests may not be administered unsupervised and must be administered by a person registered with the HPCSA (HPCSA, 2012).

Questions also continue to be raised regarding the fairness with which computer-based tests can be applied with technological unsophisticated test takers (Davies, 2002; Davies et al., 2005; Foxcroft & Davies, 2006; Foxcroft et al., 2002; Goldberg & Pedulla, 2002). Studies have shown that computer-based testing may lead to increased anxiety levels, especially amongst older test takers where higher levels of computer illiteracy can be found and, in turn, have a negative impact on test performance (Davies et al., 2005; Foxcroft, Watson & Seymore, 2004). Foxcroft et al. (2004) thus concluded that, given the relationship between computer familiarity and anxiety, it was important to offer test takers with low levels of computer familiarity the alternative of doing a paper-based test, or, if they preferred to do the computer-based test, they should be given a more extensive introductory tutorial and be debriefed afterwards. Where different modes of administration are used interchangeably, mode equivalence needs to be established (APA, 1986; ITC, 2005; Mead & Drasgow, 1993; Rosenfeld et al., 1996; Vispoel et al., 2001).

Equivalence

For decades, researchers and practitioners have drawn comparisons between computerised and paper-and-pencil tasks, not so much to determine which is better but rather to establish whether equivalence exists between the two different test modes (Noyes & Garland, 2008). Numerous studies have already investigated whether differences in administration mode have an effect on test performance. While the majority of recent equivalence studies suggest that computer-based tests and paper-based tests are in fact comparable (Arce-Ferrer & Guzman, 2009; Bartram & Brown, 2004; Holtzhausen, 2004; Joubert & Kriek, 2009; Pouwer, Snoek, Ploeg, Heine & Brand, 1998; Salgado & Moscoso, 2003; Simola & Holden, 1992; Vispoel et al., 2001; Wang et al., 2008), results have not been unanimous and several researchers have reported opposing results. Differences in terms of score distributions as well as psychometric properties between the two test modes have been found in various studies (Buchanan, Ali, Heffernan, Ling, Parrott, Rodgers, & Scholey, 2005; Buchanan, Johnson, & Goldberg, 2005). Some researchers have reported higher test scores for computer-based tests (Bugbee & Bernt, 1990; Clariana & Wallace, 2002; Pomplum et al., 2002), while others have found lower test scores for computer-based tests (Mazzeo et al., 1991; Russell, 1999). Existing literature is therefore said to be inconsistent and inconclusive. Testing is central to the disciplines of applied psychology and education (Garland & Noyes, 2008) and with the increased use of electronic media in testing practices, it has become of the utmost importance to determine whether the two test modes are equivalent. Establishing equivalence is particularly important in instances where different test mediums are used interchangeably.

One of the earliest reviews on equivalence testing between paper-based and computer-based testing was presented by Mazzeo and Harvey (Mazzeo & Harvey, 1988). Mazzeo and Harvey performed approximately 30 comparison studies between 1982 and 1988, focusing on intelligence, aptitude, personality and achievement and reported mixed results. While they found no test mode effect for power tests, they found significant differences for speeded tests. A meta-analytic study by Mead and Drasgow, comparing computerised and paper-and-pencil versions of 123 timed power tests and 36 speeded tests, showed similar results (Mead & Drasgow, 1993). Bodmann and Robinson (2004) found that students completing tests on familiar course content in either a paper-and-pencil or computerised format completed the computer-based test faster than the paper-based test. Despite completing the computer-based test faster, no significant differences in test scores were found.

Buchanan, Johnson et al. (2005), working with an online version of a 5-factor personality inventory, found that the latent structure of the inventory appeared to have changed slightly: a small number of items loaded on factors other than those they had loaded on in the offline development sample. Buchanan, Ali et al. (2005) compared the equivalence of online and paper-and-pencil versions of the prospective memory

questionnaire. The PMQ has four factor analytically derived subscales. In a large sample tested via the Internet, only two factors could be recovered; the other two subscales were essentially meaningless. This demonstration of non-equivalence underlines the importance of computerised test validation. Without examining the psychometric properties of a test, one cannot be sure that a test administered in computerised format actually measures the intended constructs.

Researchers have formulated various hypotheses to account for such test mode effects. Leeson (2006) divides the potential factors leading to test mode effects into two perspectives: participant-related factors and technology-related factors. Participant-related factors refer to the demographical characteristics of the test taker such as gender, ethnicity, cognitive ability, computer familiarity and computer anxiety. Technology-related factors, on the other hand, typically include computer interface legibility (such as screen size and resolution, font characteristics, line length) and user interface (such as scrolling to locate reading, permission for item review and item presentation – one at a time or several at a time) (Leeson, 2006).

Studies on subgroup differences relating to demographical characteristics have presented inconsistent results (Wang & Shin, 2009). While recent studies generally report no mode effect for gender and ethnicity subgroups (Bennett et al., 2008; Clariana & Wallace, 2002), earlier studies have reported performance differences on paper-based and computer-based tests in terms of gender, ethnicity, race and age (Parshall & Kromrey, 1993; Gallagher, Bridgeman & Cahalan, 2000). Although the differences found are small, they pose a threat in terms of test mode equivalence and bring the issue of test fairness into question. Noyes and Garland (2008) further suggest that competency-related factors such as reading speed, reading accuracy and comprehension may potentially affect test performance and therefore influence the equivalence between paper-based and computer-based tests.

Barak (2003) points out that a lack of computer skills could become a handicap when a computer unfamiliar participant is required to take a computerised or online test. Studies have shown that computer-based testing may lead to increased anxiety levels, especially amongst older test takers where higher levels of computer illiteracy can be found and, in turn, have a negative impact on test performance (Davies et al., 2005; Foxcroft et al., 2004). Again findings from past research show mixed results. While some studies indicate that computer familiarity and computer anxiety are significant predictors of test taker performance (Goldberg & Pedulla, 2002; Bennett et al., 2008), others do not replicate these results (Mazzeo et al., 1991; Clariana & Wallace, 2002; McVay, 2002; Taylor, Kirsch, & Eignor, 1999). McVay (2002) attributes his findings to the changing characteristics of modern society. The increased availability of computers coupled with the high rate of daily interaction with technology means the concept of computer-based testing is no longer foreign or distressful. Research further suggests that the effect of

computer unfamiliarity on test performance may be diminished by administering a computer-based test tutorial prior to the testing process (Foxcroft et al., 2004; Taylor et al., 1999).

In terms of technology-related factors, researchers such as McKee and Levinson (1990) and Leeson (2006) suggest that computer characteristics may affect the nature of a task to such an extent that the items from paper-based and computer-based tests no longer measure the same construct. Wise and Plake (1989) identified various features of computer-based testing that may influence scores, including the inability to (a) skip items and answer them later in the test, (b) review items already answered and, (c) change answers to items. While such features come standard in paper-based testing, they are often unavailable in computer-based testing, which could potentially impact on test performance. More recent studies by Lunz (1995) and Vispoel (2000) have focused on similar issues of flexibility in terms of item review, revision and the skipping of items. Characteristics of the computer-based testing medium, such as anonymity and the use of computer-mediated communication, have also been related to phenomena such as disinhibition effects or reduced socially desirable responding (Buchanan, Johnson et al., 2005; Joinson, 1999). Findings by Joinson (1999) suggest that individuals are likely to disclose more about themselves in online questionnaires, and thus may respond more truthfully to personality questionnaires. However, contrasting findings were reported by Rosenfeld et al. (1996) and Cronk and West (2002) who reported higher levels of impression management in the computer-linked condition.

In conclusion, it is clear that extensive research has been performed to establish whether mode equivalence exists between paper-based and computer-based test modes. However, despite the extensiveness of past research, results have at large been inconclusive. Due to the potential impact of mode effects and ongoing concerns about the potential negative effect of computer unfamiliarity on test performance, equivalence needs to be proven where different test modes are used interchangeably. In addition, the shortage of comparability studies between paper-based and computer-based tests in the multicultural South African context further necessitates the testing of equivalence. The general objective of this study is therefore to determine whether traditional paper-and-pencil and computerised assessments lead to equivalent results when testing facets from the soft-heartedness personality cluster on a dichotomous rating scale.

METHOD

The research method consisted of a literature review and an empirical study (quantitative research).

Literature review

The literature review was conducted by making use of databases such as *Academic Search Premier, EBSCO Host, SAe Publications, Science Direct, and Emerald Online*. The most recently published relevant articles were identified. Relevant journals such as *South African Journal of Industrial Psychology, South African Journal of Psychology, Journal of Occupational and Organizational Psychology, Personnel Psychology, Behavior Research Methods, Journal of Personality Assessment, International Journal of Selection and Assessment,* and *International Journal of Testing* were consulted in the search. The aim of the literature review was to explore and understand current issues relating to personality testing in the South African context. The review was also aimed specifically at investigating existing research regarding the use of computerised assessment measures and the equivalence between paper-and-pencil and computer-based modes of administration. Furthermore, the literature study motivated the need for establishing test mode equivalence between paper-and-pencil and computerised modes of administration.

Empirical study

The empirical study consisted of the research design, the research participants and the measuring instruments.

Research design

The research study was quantitative in nature. A quantitative design was used to "predict, describe and explain quantities, degrees and relationships" (Du Plooy, 2002, p.82). Findings from quantitative studies were generalised from a sample to the wider population by collecting numerical data. For the purpose of this study, a cross-sectional design was used, meaning that the sample was drawn from a population at a single point in time (Du Plooy, 2002). Information collected was used to describe the population at a specific point in time. The primary goal of such research was to assess interrelationships among variables within a population and to describe cause-and-effect relationships between observable and tangible variables (Kerlinger & Lee, 2000; Du Plooy, 2002).

Research participants

A combination of quota and convenience sampling was used. A sample was drawn from university students in South Africa (N = 724). The sample included undergraduate students from two higher education institutions in South Africa. The sample included both male and female participants from different race and language groups. Although specific inclusion and exclusion criteria were not followed, the aim was to include a diverse group of participants which were representative of the South African population demographics. Table 1 includes the characteristics of the participants.

TABLE 1

Characteristics of Participants (N=724)

		Paper-based: N = 344		Computerised: N = 380		Total: N =	724
Item	Category	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Age	17-19 years	155	45.06	63	16.60	218	30.10
	20-21 years	118	34.30	200	52.70	318	43.90
	22-23 years	46	13.37	85	22.40	131	18.10
	24-25 years	9	2.62	32	8.40	41	5.70
	Other	8	2.33	0	0	8	1.10
	Missing Values	8	2.30	0	0	8	1.10
Level of Education	Grade 12	272	79.10	315	82.90	587	81.10
	Certificate	16	4.70	12	3.20	28	3.90
	Diploma	11	3.20	12	3.20	23	3.20
	Bachelors	28	8.10	37	9.70	65	9
	Honours	10	2.90	1	0.30	11	1.50
	Masters	0	0	3	0.80	3	0.40
	Other	2	0.60	0	0	2	0.30
	Missing Values	5	1.50	0	0	5	0.70
Gender	Male	93	27.00	88	23.20	181	25
	Female	247	71.80	292	76.80	539	74.40
	Missing values	4	1.20	0	0	4	0.60
English reading ability	Very poor	2	0.60	2	0.50	4	0.60
	Poor	4	1.20	3	0.80	7	1
	Good	129	37.50	131	34.50	260	36
	Very Good	200	58.10	244	64.20	444	61.30

	Missing values	9	2.60	0	0	9	1.20
First Language	Afrikaans	69	20.10	39	10.30	108	14.90
	English	101	29.40	184	48.40	285	39.40
	isiNdebele	2	0.60	4	1.10	6	0.80
	isiXhosa	14	4.10	16	4.20	30	4.10
	isiZulu	53	15.40	50	13.20	103	14.20
	Siswati	8	2.30	19	5.00	27	3.70
	Sepedi	14	4.10	18	4.70	32	4.40
	Sesotho	17	4.90	30	7.90	47	6.50
	Setswana	30	8.70	5	1.30	35	4.80
	Tshivenda	10	2.90	6	1.60	16	2.20
	Xitsonga	11	3.20	4	1.10	15	2.10
	Other	9	2.60	5	1.30	14	1.90
	Missing Values	5	1.50	0	0	5	0.70
Race	White	103	29.90	140	36.60	243	33.60
	Black	175	50.90	171	45.00	346	47.80
	Indian	27	7.80	15	13.40	42	5.80
	Coloured	27	7.80	18	4.70	45	6.20
	Missing values	12	3.50	0	0	12	1.70

Table 1 reflects the characteristics of research participants in terms of race, gender, level of education, age, language and perceived English reading proficiency. According to Table 1, the total sample (N = 724) consisted mainly of black (47.80%) and white (33.60%) female (74.40%) participants between the age of 17 and 21 years (74%). The majority of the participants were English- (39.40%) or Afrikaans-speaking (14.90%) and had completed a minimum education level of Grade 12 (81.10%). The majority of participants also perceived their English reading ability to be Very Good (61.30%) or Good (35.90%).

A total of 344 participants completed the paper-and-pencil version of the questionnaire, while 380 participants completed the computerised version. The sample group completing the paper-and-pencil version consisted mainly of black (50.90%) females (71.80%) between the ages of 17 and 19 years (45.06%). The majority of participants were English-speaking (29.40%), had a minimum education level of grade 12 (79.10%) and perceived their English reading ability to be very good (58.10%). Similarly, the sample group completing the computerised version of the questionnaire consisted mainly of black (45%) females (76.80%) between the ages of 20 and 21 years (52.70%). The majority of participants were English-speaking (48.40%), had a minimum education level of grade 12 (82.90%) and perceived their English reading ability to be very good (64.20%).

Measuring instruments

The study made use of two parallel measuring instruments, consisting of items which measured different facets from the soft-heartedness personality construct. The soft-heartedness questionnaire was presented in both paper-and-pencil and computerised formats. Various facets from the SAPI have been measured for reliability in two independent studies with university students from various industries (Flattery, 2011; Oosthuizen, 2012; Van der Linde, 2012). The facets which generated the highest levels of reliability were then extracted and used to construct a behaviour questionnaire consisting of 24 items. Because the focus of the study was on determining the item functioning between two different modes of administration and not on measuring the overlap between clusters, sub-clusters or facets of personality, it was satisfactory to include only certain facets. The facets to be included were "generous", "compassionate" and "appreciative". The following alpha coefficients were attained in the first pilot study (Flattery, 2011): generous ($\alpha = 0.83$), compassionate ($\alpha = 0.87$) and appreciative ($\alpha = 0.83$). In the second study (Van der Linde, 2012), the following alpha coefficients were attained: generous ($\alpha = 0.77$), compassionate ($\alpha = 0.88$) and appreciative ($\alpha = 0.86$).

The paper-and-pencil and computerised formats were compared in terms of factor-loadings with the various facets, reliability and validity to determine whether equivalence exists between the two modes of administration. Besides the difference in administration mode the questionnaires were made to look as similar as possible in all other aspects such as number of questions per page, colour, numbering of questions etc, to minimise the possibility of scoring differences due to other factors. Both questionnaires were on a dichotomous (two-choice) rating scale – consisting only of "agree" or "disagree" response options. Statements such as *"I share what I have with others"* (generous), *"I am sensitive to other people's feelings"* (compassionate), and *"I value life as it is"* (appreciative) were included.

Research procedure

Both versions of the soft-heartedness questionnaire were administered to undergraduate students. The paper-and-pencil version was administered to students at the various universities during their normal lecture hours. The times and dates of the assessment sessions were pre-arranged with the various subject lecturers. The assessments took place in a controlled classroom setting. Researchers were present to provide instructions and to supervise the assessment process. The computerised assessments also took place under supervised test conditions in computer laboratories at the two universities. The computerised version of the questionnaire was loaded onto a server and students were granted access to the questionnaire through the use of passwords. Students were then requested to schedule a time to complete the questionnaire, using extra marks towards their subject grade as incentive to complete the

questionnaire. The data obtained from the two different modes of administration was then factor-analysed and compared in terms of factor loadings, variance explained, validity, and reliability to whether equivalence exists between the different test modes.

Statistical analysis

Statistical analysis was carried out with the help of the SPSS program (SPSS, 2008). The data from both questionnaires were firstly inspected for missing and unexpected values. Then descriptive statistics were analysed, followed by exploratory factor analysis. After the item loadings were inspected, reliability analyses were carried out.

Descriptive Statistics

The functionality of the data was investigated in terms of mean scores, standard deviation, skewness and kurtosis. The investigation of descriptive statistics made it possible to identify poorly functioning items and items that were arbitrary to the purpose of the study. Such items were consequently eliminated from the study. It was decided that items with skewness > 2 or kurtosis > 4 would be excluded from further analysis (see DeCarlo, 1997).

Exploratory factor analysis

Factor analysis refers to an analytical method used to determine the number of underlying factors within a construct as well as the statistical characteristics of such factors (Murphy & Davidshofer, 2005). Murphy and Davidshofer (2005) further state that factor analysis makes use of statistical means to "identify the basic underlying variables that account for the correlations between actual test scores" (p. 87). Factor analysis in this study was exploratory in nature. Exploratory factor analysis is used to uncover the underlying factor structure of relatively large sets of variables. For two formats of a test to be considered equivalent, there should be the same number of underlying factors and these factors should account for similar proportions of variance between test results (King & Miles, 1995; Salgado & Moscoso, 2003).

Principal component analysis was performed to investigate factor loadings and communalities. Communalities in variables refer to the total amount of variance that a variable shares with all other variables included in the analysis (Hair, Black, Babin, Anderson, & Tathum, 2006). Communalities are used to assess the degree to which an item is a good, reliable measure of the factor under investigation. Larger communalities are generally indicative of higher quality measures and the cut-off was set on 0.20. Coefficients below 0.20 are generally indicative of poor factoral congruence. Item loadings were also investigated by the use of principal component analysis. First the variance was inspected. Variance was measured in terms of Eigenvalues. Eigenvalues measure the amount of variance in relation to total

variance. Eigenvalues were investigated to determine the extent to which each test format account for variance in the items of the Soft-Heartedness construct. A popular way of performing factor analysis is by inserting unities (1.0) in the diagonal cells, extracting all factors with eigenvalues of 1.0 or more and rotating the factors by varimax. This is commonly referred to as the Eigenvalue 1 method (Comrey, 1988).

Reliability analysis

The reliability of a measure refers to the consistency with which it measures whatever it measures (Foxcroft & Roodt, 2005). Certain psychological measures, such as personality and attitude scales, generally consist of multi-scored items that have no right or wrong answers. Cronbach developed his Coefficient Alpha as an appropriate reliability estimator for measures containing multiple components (Osburn, 2000). Cronbach's alpha coefficient ranges from 0 to 1. According to Murphy and Davidshofer (2005) reliability estimates of 0.95 are likely to be regarded as high levels of reliability in testing applications. Estimates of 0.80 or more are generally regarded as moderate to high, with estimates lower than 0.60 indicating unacceptably low levels of reliability. Researchers such as Murphy and Davidshofer (2005) and DeVellis (2003), however, state that reliability estimates greater than 0.70 might be satisfactory for research and testing purposes. This is especially true where testing is used for initial screening processes or the making of preliminary decisions. For the purpose of this study, alpha values higher than 0.70 thus indicated acceptable levels of reliability.

Ethical considerations

Fair and ethical research procedures were vital to the success of this study. The following ethical considerations were met:

Permission to perform assessments was obtained from both universities. Ethical aspects regarding the research were discussed with all participants prior to assessment. Participation in the research was voluntary and participants were required to sign informed consent forms giving researchers permission to use their results for academic purposes. Electronic versions of the informed consent form were completed prior to the computerised assessment. Participants were informed verbally about the purpose of the study. The data obtained from the study was used for academic research purposes only, to address the item functioning and psychometric properties of the questionnaire. Participants therefore remained confidential, were not analysed on their individual personality constructs, and did not receive feedback on their performance. This was communicated to participants in advance.

RESULTS

Tables 2 and 3 summarise descriptive statistics of the data in terms of mean scores, standard deviation, skewness and kurtosis for both modes of administration.

Table 2

Descriptive Statistics for the Paper-and-Pencil mode

		Paper-and-Pencil (*N* = 344)			
Item		**Mean**	**SD**	**Skewness**	**Kurtosis**
i001	I share what I have with others	**0.95**	**0.22**	**-4.04**	**14.39**
i002	I give things to people without expecting anything in return	0.89	0.31	-2.54	4.50
i003	I share my knowledge with others	**0.94**	**0.23**	**-3.91**	**13.38**
i004	I buy things for others	0.78	0.42	-1.33	-0.23
i005	I give money to the poor	0.62	0.49	-0.51	-1.75
i006	I treat others with gifts	0.62	0.49	-0.51	-1.75
i007	I give food to people who do not have any	0.81	0.39	-1.62	0.63
i008	I provide for those in need	0.73	0.44	-1.06	-0.89
i009	I am sensitive to others' needs	0.88	0.32	-2.36	3.60
i010	I feel sympathy for people who have problems	**0.96**	**0.120**	**-4.67**	**19.92**
i011	I can share in someone's emotions	0.88	0.33	-2.32	3.40
i012	When someone cries, I also feel like crying	0.56	0.50	-0.22	-1.96
i013	I get sad when I see someone suffering	**0.96**	**0.21**	**-4.49**	**18.26**
i014	I get sad when someone I care about is sad	**0.95**	**0.22**	**-4.04**	**14.39**
i015	I feel other people's problems as my problems	0.39	0.50	-0.44	-1.82
i016	I am sensitive to other people's feelings	0.90	0.31	-2.59	4.76
i017	I like pleasant things	**0.96**	**0.19**	**-4.87**	**21.83**
i018	I value the company of people close to me	**0.98**	**0.14**	**-6.82**	**44.83**
i019	I enjoy delicious food	**0.99**	**0.12**	**-8.15**	**64.77**
i020	I value life as it is	0.90	0.30	-2.70	5.32
i021	I am happy when I see good things happening in other people's lives	**0.96**	**0.20**	**-4.67**	**19.92**
i022	I value the little things in life	0.90	0.30	-2.76	5.63
i023	I value pleasant experiences	**0.97**	**0.18**	**-5.09**	**24.07**
i024	I value other people's work	0.88	0.32	-2.36	3.60

TABLE 3

Descriptive Statistics for the Computerised mode

		Computerised (*N* = 380)			
Item		**Mean**	**SD**	**Skewness**	**Kurtosis**
i001	I share what I have with others	**0.94**	**0.23**	**-3.91**	**13.34**
i002	I give things to people without expecting anything in return	0.89	0.32	-2.45	4.03
i003	I share my knowledge with others	**0.98**	**0.15**	**-6.29**	**37.76**
i004	I buy things for others	0.86	0.35	-2.09	2.38
i005	I give money to the poor	0.71	0.45	-0.95	-1.11
i006	I treat others with gifts	0.69	0.46	-0.82	-1.33
i007	I give food to people who do not have any	0.80	0.40	-1.51	0.27
i008	I provide for those in need	0.80	0.40	-1.53	0.33
i009	I am sensitive to others' needs	**0.95**	**0.22**	**-4.02**	**14.26**
i010	I feel sympathy for people who have problems	**0.96**	**0.19**	**-4.94**	**22.49**
i011	I can share in someone's emotions	0.91	0.29	-2.89	6.37
i012	When someone cries, I also feel like crying	0.63	0.48	-0.56	-1.70
i013	I get sad when I see someone suffering	**0.95**	**0.22**	**-4.02**	**14.21**
i014	I get sad when someone I care about is sad	**0.97**	**0.17**	**-5.64**	**29.98**
i015	I feel other people's problems as my problems	0.48	0.50	0.07	-2.01
i016	I am sensitive to other people's feelings	**0.94**	**0.23**	**-3.91**	**13.34**
i017	I like pleasant things	**0.99**	**0.07**	**-13.73**	**187.48**
i018	I value the company of people close to me	**1.00**	**0.05**	**-19.494**	**380.00**
i019	I enjoy delicious food	**0.99**	**0.11**	**-8.58**	**71.97**
i020	I value life as it is	0.91	0.29	-2.88	6.35
i021	I am happy when I see good things happening in other people's lives	**0.98**	**0.14**	**-6.70**	**43.10**
i022	I value the little things in life	**0.95**	**0.22**	**-4.15**	**15.27**
i023	I value pleasant experiences	**1.00**	**0.05**	**-19.47**	**379.00**
i024	I value other people's work	**0.94**	**0.23**	**-3.91**	**13.34**

The descriptive statistics of the data are indicative of the quality of the data. Poorly functioning items were indicated by mean scores greater than 0.94, skewness values greater than 2 and kurtosis values greater than 4 (see DeCarlo, 1997). Such items were consequently excluded from further analysis.

Table 2 shows that ten items proved to be problematic in terms of mean scores in the paper-based format. Mean scores greater than 0.94 were considered to be indicative of test items that are generally too easy to

agree with. Such items can potentially skew the data. As a result, the following ten items were removed from further analysis: i001 ("I share what I have with others"), i003 ("I share my knowledge with others"), i010 (I feel sympathy for people who have problems"), i013 ("I get sad when I see someone suffering"), i014 ("I get sad when someone I care about is sad"), i017 ("I like pleasant things"), i018 ("I value the company of people close to me"), i019 ("I enjoy delicious food"), i021 ("I am happy when I see good things happening in other people's lives"), and i023 ("I value pleasant experiences").

Table 3 shows that 14 items proved to be problematic in terms of mean scores in the computer-based format. The same items that were problematic in the paper-based version also performed poorly in the computer-based version and were consequently removed from further analysis. In addition, the following four items were also removed: i009 ("I am sensitive to others' needs"), i016 ("I am sensitive to other people's feelings"), i022 ("I value the little things in life"), and i024 ("I value other people's work").

Tables 2 and 3 further show that the problematic items also performed poorly in terms of skewness and kurtosis across both modes of administration. These results are indicative of random response patterns from the participants. Particularly poor item performance can be seen in items i018 and i023 in the computer-based format. A mean score of 1 can be found on both these items, indicating that all participants taking the computer-based test answered "yes" to the presented question. Such results may suggest the possible presence of a social desirability element in the item. On the other hand, item i012 on the paper-based format showed particularly good item performance with a mean score of 0.56. These results indicate a relatively equal distribution between "yes" and "no" answers on the presented item. Ideally, mean scores should fall between 0.4 and 0.6 for items to effectively measure the presence or absence of soft-heartedness. Following the investigation of the descriptive statistics of the data, ten items were retained for further analysis. All retained items fell between the desired cut-off points for good item performance.

TABLE 4

Eigenvalues of Sample Correlation Matrix

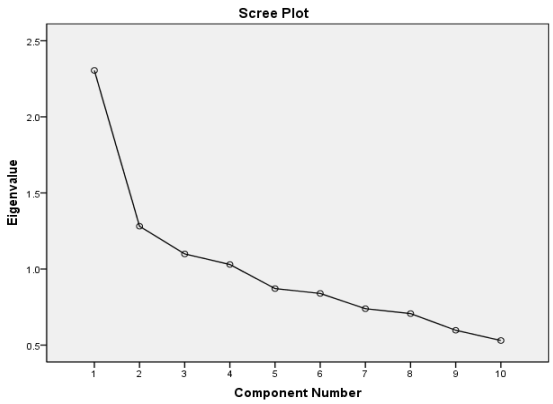
	Paper-and-Pencil (N = 344)			Computerised (N = 380)		
Component	**Eigenvalue**	**% of variance**	**Cumulative % variance explained**	**Eigenvalue**	**% of variance**	**Cumulative % variance explained**
01	**2.31**	23.05	23.05	**2.27**	22.69	22.69
02	**1.28**	12.81	35.86	**1.42**	14.23	36.93
03	**1.10**	10.99	46.85	**1.08**	10.94	47.87
04	**1.03**	10.30	57.15	**1.01**	10.07	57.94
05	0.87	8.71	65.87	0.92	9.22	67.16
06	0.84	8.40	74.26	0.82	8.15	75.31
07	0.74	7.39	81.65	0.76	7.59	82.90

08	0.71	7.07	88.72	0.61	6.08	88.97
09	0.60	5.98	94.70	0.59	5.91	94.88
10	0.53	5.30	100.00	0.51	5.12	100.00

Table 4 demonstrates the Eigenvalues, percentage of variance and cumulative percentage of variance for the two modes of administration. Exploratory factor analysis was applied on the ten retained items to determine the degree to which each successive instance of measurement accounts for more or less of the total variance of the construct under investigation (Pallant, 2007). In theory the soft-heartedness personality cluster consists of three factors, namely "generous", "compassionate", and "appreciative" (Nakani, 2011; Nel, 2008). However, when the Eigenvalue greater than 1 rule is applied, both the paper-based and computerised formats retained four factors each. With four factors retained the cumulative percentage of variance explained by the paper-based format is approximately 57.15%, with the computerised format accounting for approximately 57.94% of variance. This indicates equivalence between the two modes of administration with both formats accounting for similar percentages of variance in the Soft-Heartedness construct. Table 4 further indicates that the highest item loadings were on the first factor for both modes of administration. Factor 1 accounts for 23.05% of variance in the paper-and-pencil format, and 22.69% of variance in the computerised format. This indicates that both modes of administration account for a similar percentage of variance when a single factor is retained.

The corresponding scree-plot from the paper-based format demonstrates the existence of two distinct factors. On the other hand, the scree-plot from the computerised format demonstrates the existence of three factors, which corresponds with the theoretic view of soft-heartedness (see Nakani, 2011; Nel, 2008).

SCREE PLOT 2-POINT PAPER-AND-PENCIL



SCREE PLOT 2-POINT COMPUTERISED

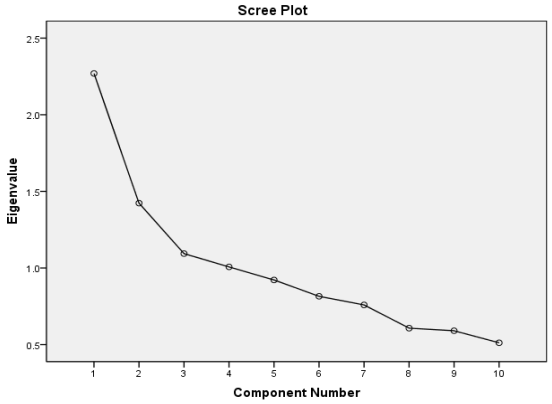


TABLE 5

Factor loadings with 4 factors extracted – 2-Point Paper-and-Pencil

	Paper-and-Pencil (N = 344)			
	1	2	3	4
i002	0.40		-0.42	-0.46
i004	0.64		-0.37	
i005	0.43	-0.39		
i006	0.53		-0.35	
i007	0.54	-0.34	0.37	
i008	0.64	-0.37		
i011	0.53	0.46		
i012	0.37	0.62	0.31	
i015	0.35	0.42	0.50	
i020		0.30		0.77

TABLE 6

Factor loadings with four factors extracted – 2-Point Computerised

	Computerised (N = 380)			
	1	2	3	4
i002	0.28	-0.06	-0.21	0.65
i004	0.60	-0.11	-0.43	-0.24
i005	0.60	-0.24	0.04	0.12
i006	0.55	-0.14	-0.60	-0.14
i007	0.52	-0.34	0.55	0.11
i008	0.68	-0.22	0.32	0.08
i011	0.42	0.59	0.15	0.08
i012	0.32	0.72	0.17	-0.08
i015	0.31	0.55	-0.14	0.10
i020	0.28	-0.04	0.18	-0.67

Tables 5 and 6 indicate the item loadings on the four extracted factors for both test modes. It is clear that the retained items loaded on more than one factor indicate a lack of congruence in the response patterns of test takers. Random response patterns could especially be identified in the computer-based test mode, where all test items loaded on more than one factor. An investigation of the strength of item loadings indicates that the strongest loadings were on the first factor. These findings correspond with the findings from Table 4. It was therefore decided to extract a single factor.

Table 7 shows the factor loadings and communalities found in both the paper-based and computer-based test modes when a single factor is extracted.

TABLE 7

Factor loadings and communalities – 2-Point

	Paper-based (*N* = 344)		Computerised (*N* = 380)	
	Factor loading	**Communalities**	**Factor loading**	**Communalities**
i002	0.40	0.16		
i004	0.64	0.41	0.60	0.36
i005	0.43	0.19	0.60	0.36
i006	0.53	0.28	0.55	0.30
i007	0.54	0.29	0.52	0.27
i008	0.64	0.41	0.68	0.46
i011	0.53	0.28	0.42	0.17
i012	0.37	0.14	0.32	**0.10**
i015	0.35	0.12	0.31	**0.10**
i020				

With one factor retained, nine out of the ten items from the paper-based format loaded onto Factor 1. In the computer-based format, eight out of the ten items loaded onto Factor 1. Factor loadings for the paper-based administration mode ranged from 0.35 to 0.64, while factor loadings for the computerised mode ranged from 0.31 to 0.68. The average factor loading for the paper-based format was 0.49, with the computerised format scoring a similar average of 0.5. According to Kerlinger & Lee (2000), factor loadings smaller than 0.3 indicate that an item is a poor measure of the factor under investigation. It can thus be concluded that the retained items are sufficiently associated with the factor structure and are adequate measures of the soft-heartedness construct.

In term of communalities, inter-item communalities ranged from 0.12 to 0.41 in the paper-based format, and from 0.10 to 0.46 in the computer-based format. The average communality for the paper-based mode was 0.25, with the computerised again scoring similarly with an average of 0.25. Inter-item correlations should be greater than 0.2 for the item to sufficiently correspond with the rest of the items measuring the construct under investigation. With several items measuring below 0.2, results indicate low communalities between the test items and the construct under investigation. The combination of test items thus poorly measures the soft-heartedness construct. In the computer-based format 4 out of the 10 retained items did not measure soft-heartedness at all. Factors i002, i012, i015 and i020 had communalities of 0.1 and lower, and were therefore not measures of the soft-heartedness construct. It is possible that these items measure another unrelated construct. This indicates that the paper-based test format had superiority item functioning when compared to the computer-based format. The poor item loadings found on both modes of administration, however, are indicative of the poor overall functioning of the two-point rating scale on both test modes.

Reliability analysis using the Cronbach Alpha coefficient of internal consistency was used to determine to what degree the paper-based or computer-based test modes were more or less reliable.

TABLE 8

Cronbach Alpha Coefficients of both test modes

Test Mode	**α (Phi)**	**Number of items**
Paper-and-Pencil	0.61	9
Computerised	0.59	8

TABLE 9

Reliability analysis – 2 point Paper-and-Pencil (9 items retained)

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
i002	5.39	3.32	0.22	0.60
i004	5.51	2.92	0.40	0.55
i005	5.66	3.02	0.24	0.59
i006	5.66	2.92	0.30	0.58
i007	5.47	3.06	0.33	0.57
i008	5.55	2.85	0.41	0.55
i011	5.41	3.17	0.33	0.57
i012	5.73	3.03	0.22	0.60
i015	5.89	3.04	0.22	0.60

With nine items retained, the paper-based questionnaire has an alpha coefficient of 0.61, which is lower than the recommended minimum cut-off of 0.7. However, a lenient cut-off of 0.6 is often considered to be acceptable for exploratory studies (Black & Porter, 2005), which means that the paper-based test mode showed acceptable reliability.

TABLE 10

Reliability analysis – 2 Point Computerised (8 items retained)

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
i004	5.03	2.42	0.37	0.54
i005	5.18	2.27	0.33	0.54
i006	5.20	2.30	0.30	0.55
i007	5.09	2.45	0.25	0.57
i008	5.09	2.27	0.42	0.52
i011	4.98	2.58	0.30	0.56
i012	5.26	2.38	0.22	0.58
i015	5.41	2.34	0.22	0.58

With 8 items retained, the computer-based questionnaire shows a reliability coefficient of 0.59 (α = 0.59), which is again lower than the recommended minimum cut-off of 0.7. The computer-based measure thus shows a lower level of internal consistency compared to the paper-based measure. In addition, fewer items could be retained in the computer-based measure, suggesting that the paper-based measure may be superior to the computer-based version. An investigation of alpha coefficients with items removed indicates an increase in reliability with the removal of items i012 and i015. Reliability analysis was consequently performed with these two items removed to investigate the effect on the internal consistency of the computer-based measure. The results are indicated in table 11.

TABLE 11

Reliability analysis – 2-Point Computerised (6 items retained)

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
i004	3.92	1.46	0.41	0.55
i005	4.06	1.31	0.39	0.55
i006	4.09	1.36	0.31	0.58
i007	3.98	1.45	0.32	0.58
i008	3.97	1.32	0.48	0.51
i011	3.87	1.72	0.16	0.62

With six items retained the computer-based measure shows a similar level of reliability to the paper-based measure (α = 0.61). However, since a larger number of items could be retained in the paper-based questionnaire, the traditional paper-and-pencil format still appears to be superior to the computer-based format.

DISCUSSION

The primary objective of this study was to determine whether traditional paper-and-pencil and computerised assessment measures will lead to equivalent results when testing facets from the soft-heartedness personality cluster on a dichotomous rating scale. Establishing equivalence forms an important part of the test development process as it would allow assessment practitioners to use the two test modes interchangeably. Another main objective of the study is to determine how the SAPI will be administered in the future. Where equivalence can not be proven, researchers need to determine the optimal mode of administration that will lead to the most valid and reliable results. The aims of the current study are also in line with the primary objective of the SAPI as a whole. According to Nel (2008), this objective is to develop an instrument that is uniquely South African and that provides valid and

reliable measurements of personality in the South African context and among the diverse South African population.

An evaluation of the descriptive statistics of the data showed that items from the paper-based format had superior item functioning when compared to items from the computer-based format. Poorly functioning items were indicated by mean scores greater than 0.94, skewness values greater than 2 and kurtosis values greater than 4 (see DeCarlo, 1997). Although comparable average mean scores were found between the two test modes (0.85 for the paper-based and 0.88 for the computer-based), the paper-based mode outperformed the computer-based mode in terms of skewness and kurtosis. For both modes of administration, however, average skewness and kurtosis values were higher than the recommended cut-off points, indicating that results were not normally distributed. Skewness values were all negative, which indicates that the mass of the distribution is concentrated on the right-hand side of the graph (Pallant, 2007). For the purpose of this study, it demonstrates that the majority of test-takers answered "yes" to the presented test items. Test items were formulated in such a way that a "yes" answer signifies the presence of soft-heartedness while a "no" answer signifies the absence of soft-heartedness. The computer-based mode in particular showed very high average kurtosis values (52.05) compared to the paper-based mode (11.61).

The high skewness and kurtosis values demonstrated on both modes of administration may indicate that test items were generally too easy to agree with. This may result in an over-measurement of the presence of soft-heartedness. Further investigation into the effective formulation and phrasing of test items is therefore required. Particularly poor item performance was seen in items i018 ("I value the company of people close to me") and i023 ("I value pleasant experiences") in the computer-based format. A mean score of 1 can be found on both these items, indicating that all participants taking the computer-based test answered "yes" to the presented question. Ideally, mean scores should fall between 0.4 and 0.6 for items to effectively measure the presence or absence of soft-heartedness. Findings such as these may suggest the presence of a social desirability element in these test items. Results further suggest that higher levels of socially desirable responding were present in the computer-based test version when compared to response patterns on the paper-based test.

These results correspond with finding from Rosenfeld et al. (1996) and Cronk and West (2002), who found higher levels of impression management in computer-based testing compared to paper-based testing. According to Cronk and West (2002) issues of confidentiality and anonymity may affect the way in which test takers respond to computer-based tests. Participants may feel uncomfortable providing information over a computer-based medium, because they believe that others may use the results. It was therefore concluded that perceiving that one's responses are linked to a larger database may lead to more

impression management on computer surveys. The use of a two-point rating scale may further have contributed to the high levels of skewness and kurtosis demonstrated in the study. Unlike a polytomous rating scale, a dichotomous scale does not allow one to measure the presence of varying levels of soft-heartedness on a test item. The fact that only "yes" or "no" responses were allowed could therefore have led test-takers to respond more favourably than they would have, had there been more response options.

According to Jacoby and Mattel (1971) too few response categories result in a loss of much of the raters' discriminating powers. This explains the poor ability of the two-point rating scale to effectively measure the presence or absence of the construct under investigation (soft-heartedness). Findings from the current study therefore suggest that it may be preferable to include more response categories when testing facets of personality in the South African context. The study consequently supports the use of polytomous rating scales over dichotomous rating scales within the applied context. Finally, further investigation is required into the causation of impression management in the computer-based test mode as well as the potential presence of a social desirability element in certain test items.

Items falling outside of the recommended cut-off points for mean scores, skewness and kurtosis may also be indicative of random response patterns amongst test takers. Results suggest that higher levels of random responding were present in the computer-based test mode. High levels of random responding could potentially have been caused by a lack of motivation from test takers. According to Wise and Kong (2005) and Osborne and Blanchard (2011) random responding is common in low stakes assessments. The degree to which test takers give their best efforts is often unclear in low stakes assessments. Unmotivated test takers may answer too quickly instead of taking the time to read and fully consider each test item. This lack of motivation to perform may affect the outcomes of the measure, resulting in test bias and decreased validity. It may be assumed that test takers will consider items more carefully and answer more accurately if they have a personal interest in the outcome of the questionnaire. Had the SAPI been administered in its entirety in a context where stakes are involved (e.g. personality to job fit), lower levels of random responding would be expected.

After investigating the descriptive statistics of the data only ten items could be retained. Items falling outside of the recommended cut-off points can potentially skew the data and were consequently removed from further analysis. In the paper-based test mode, ten items proved to be problematic in terms of item functioning. In the computer-based test mode an additional four items were problematic. The remaining ten items all fell between the desired cut-off points for good item performance. The small number of retained items can be attributed to the fact that only a single facet of personality was included in the study. Limiting the study to a single facet thus limits the pool of items to be included in the analyses.

Furthermore, poor item construction may have resulted in high levels of impression management and skewness in the response style of test takers.

Exploratory factor analysis was performed on the ten retained items. According to soft-heartedness theory, the construct of soft-heartedness consists of three factors, namely "generous", "compassionate", and "appreciative". Initial factor analysis, using the Eigenvalue greater that 1 rule, supports the extraction of four factors for both modes of administration. However, further investigation demonstrates that the greatest percentage of variance is explained by the first factor. An investigation of the strength of item loadings also indicates that the strongest loadings were on the first factor. In addition, the majority of test items loaded onto more than one factor, which further supports the extraction of a single factor. It can therefore be concluded that the retained test items measured soft-heartedness as a whole instead of measuring the three theoretical factors of soft-heartedness. The extraction of a single factor can potentially be attributed to the small number of items used for factor analysis. In addition, random response patterns from test takers caused the retained test items to load onto more than one factor, which made it impossible to clearly differentiate between different factors of the soft-heartedness personality construct.

An investigation of inter-item communalities showed that the paper-based format had superior item functioning when compared to the computer-based format. In the computer-based format, 4 out of the 10 retained items did not correlate with the construct under investigation, and also did not correlate to any of the other factors. This proves that these items do not measure the construct of soft-heartedness at all. It is possible that these items measure another unrelated construct. In addition, several items from both modes of administration measured below the recommended cut-off point for inter-item communalities. As a result, the combination of test items included in the study was found to be a poor measure of the soft-heartedness construct. Such findings imply that more time should be spent on the compilation of test items that effectively measure the presence or absence of soft-heartedness.

Reliability analysis again showed that the paper-based test format outperformed its computer-based counterpart. The paper-and-pencil test had a higher reliability coefficient and retained more test items, making it more reliable that the computer-based version. However, reliability coefficients for both modes of administration fell below the recommended cut-off point of 0.7, again indicating the poor overall functioning of the two-point soft-heartedness scale. Although initial studies supported the use of a two-point rating scale for the SAPI (see Vorster, 2011), results from the current study seem to suggest that a polytomous rating scale may be preferable for future applications of the SAPI.

Polytomous rating scales are generally considered to better account for the variance in a variable and are often associated with greater levels of reliability and validity (Alwin, 1992; Comrey, 1988; Netemeyer, Bearder, & Sharma., 2003). Bandalos and Enders (1996) did a simulation study and found that Cronbach's alpha increased when the number of response categories increased. Maximum gains were reached with five or seven categories after which reliability values levelled off. Similarly, Comrey (1988) reported the best results in his own studies when using rating scales containing seven response categories. Past research has shown that response scales with fewer than five response categories do not allow for sufficient variance in numerical scores and are more prone to distortions in the correlation matrix (Comrey, 1988; Reise, Waller, & Comrey, 2000). Taking the above mentioned findings into consideration, polytomous rating scales may therefore be able to define the structure of the soft-heartedness personality construct more clearly. Considerations such as these, together with the poor overall functioning of the two-point rating scale in the current study, support the use of polytomous rating scales for future applications of the SAPI.

Despite the fact that the paper-based test outperformed the computer-based test in terms of descriptive statistics, reliability and number of retained items, a large degree of overlap existed between the results of the two test modes. The large degree of similarity between the two modes of administration may be attributed to the fact that steps were taken to make the two test modes as similar as possible in all aspects besides presentation mode. Proctors were present during the administration of both the paper-based and the computer-based tests, which meant that questionnaires were administered in a supervised and controlled environment. Generally high levels of similarity also existed between the research participants in terms of gender, race, age, language, level of education and English language proficiency. Mode effects could therefore have been minimised by similarities in the demographical characteristics of participants.

In conclusion, the objectives of the study could be met to an extent. Although a large degree of overlap existed between the results generated from the PPT and CBT, equivalence between the two modes of administration could not be proven. The PPT slightly outperformed the CBT in terms of descriptive statistics, reliability and retained items. As a result, the PPT was proven to be more valid and reliable than the CBT. However, results may have been compromised by the poor overall functioning of the two-point rating scale, which necessitates more research into the topic under investigation. Based on the limitations found in the current study, recommendations for future research could be made.

Limitations

Based on this discussion, the following limitations of the study could be identified:

The unequal representation of all age, race and cultural groups and levels of education means that the results from this study can not be generalised to the broader South African population. It is therefore recommended that a more diverse and representative sample group be included in future studies. Poor item functioning meant that only a small number of items could be retained for further analysis. Poor item functioning could potentially have been caused by test items that were generally too easy to agree with as well random responding due to a lack of motivation from test takers. Poor item performance in terms of mean scores also indicates that a social desirability element may have been present in a number of test items. Socially desirable responding may therefore have affected test performance especially on the computer-based format.

Recommendations

Based on this discussion, it is recommended that the SAPI be developed in its entirety before the optimal mode of administration may be determined. Not only will the completion of the instrument increase the number of test items to be used in the study, but it may also lead to a decrease in random responding. Furthermore, the poor general functioning of the two-point rating scale supports the use of polytomous rating scales for future applications of the SAPI. In order to attain results that are more representative of the South African demographic it is also recommended that a more equal distribution of gender, age, race and language groups be included in future studies.

REFERENCES

Abrahams, F. (1996). *The cross-cultural comparability of the 16PF*. Unpublished doctoral thesis, Department of Industrial Psychology, University of South Africa.

Abrahams, F., & Mauer, K. F (1999). The comparability of the constructs of the 16PF in the South African context. *South African Journal of Industrial Psychology, 25*(1), 53-59.

Alwin, D. F. (1992). Transmission in the survey interview: number of response categories and the reliability of attitude measurement. *Sociological Methodology*, *22*, 83-118.

American Psychological Association. (1986). *Guidelines for computer-based tests and interpretations*. Washington, DC: Author.

Arce-Ferrer, A. J., & Guzman, E. M. (2009). Studying the equivalence of computer-delivered and paper-based administrations of the Raven Standard Progressive Matrices Test. *Educational and psychological measurement*, *69*(5), 855-867. doi:10.1177/0013164409332219

Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education, 9*(2), 151-160.

Barak, A. (2003). Ethical and professional issues in career assessment on the internet. *Journal of career assessment, 11*(1), 3-21.

Barak, A. & Cohen, L. (2002). Empirical examination of an online version of the self-directed search. *Journal of career assessment, 10*(4), 387-400. doi:10.1177/1069072702238402

Bartram, D. (2004). Assessment in organisations. *Applied Psychology: An International Review, 53*(2), 237-259.

Bartram, D., & Brown, A. (2004). Online testing: Mode of administration and the stability of the OPQ32i scores. *International Journal of Selection and Assessment, 12*(3), 278-284. doi:101111/j.0965-075X.2004.282_1.x

Bates, S. C., & Cox, J. M. (2008). The impact of computer versus paper-pencil survey, and individual versus group administration, on self-reports of sensitive behaviors. *Computers in Human Behavior, 24*, 903-916.

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on a computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*(9). Retrieved from http://www.jtla.org.

Bodmann, S. M. & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of educational computing research, 31*(1), 51-60.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York, NY: Springer.

Bressani, R. V., & Downs, A. C. (2002). Youth independent living assessment: Testing the equivalence of web and paper/pencil versions of the Ansell-Casey Life Skills Assessment. *Computers in Human Behavior, 18*, 453-464.

Buchanan, T., Ali, T., Heffernan, T. M., Ling, J., Parrott, A. C., Rodgers, J., & Scholey, A. B. (2005). Non-equivalence of online and paper-and-pencil psychological tests: The case for the prospective memory questionnaire. *Behaviour Research Methods, 37*(1), 148-154.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the internet. *European Journal of Psychological Assessment, 21*(2), 115-127. doi:10.1027/1015_5759.18.1.116

Bugbee, A. C. & Bernt, F. M. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education, 23*(1), 87-100.

Busch, M. (1993). Likert scales in L2 research: a researcher comments. *Tesol Quarterly*, *27*(4), 733-736.

Butler, D. L. (2003). *The impact of computer-based testing on student attitudes and behaviour. The Technology Source*, Jan/Feb 2003. Retrieved from http://ts.mivu.org/default.asp?show=article&id=1013

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, *6*(3), 170-175.

Claassen, N. C. W. (1995, October 24). *Cross-cultural assessment in the human sciences.* Paper presented at a work session on the meaningful use of psychological and educational tests, Human Sciences Research Council.

Claassen, N. C. W. (1997). Cultural differences, politics and test bias in South Africa. *European Review of Applied Psychology, 47*(4), 297-307.

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology, 33(*5), 593-602. doi:10.1111/1467-8535.00294

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309-319.

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56*, 754-761.

Comrey, A. L., & Montag, I. (1982). Comparison of factor analytic results with two-choice and seven-choice personality item formats. *Applied Psychological Measurement, 6,* 285-289.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, *17*(4), 407-422.

Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers, 34*, 177-180.

Davies, C., Foxcroft, C., Griessel, L., & Tredoux, N. (2005). Computer-based and internet delivered assessment. In C. Foxcroft & G. Roodt (Eds.), *An introduction to psychological assessment in the South African context* (2nd ed.), (pp 153-166) Cape Town, South Africa: Oxford University Press.

Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. *Behavior Research Methods, Instruments & Computers, 31,* 572-577.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*(3), 292-307. doi:10.1037/1082-989X.2.3.292

DeVellis, R. F. (2003*). Scale development: Theory and applications* (2nd ed.). London, UK: Sage.

Du Plooy, G. M. (2002). *Communication research: techniques, methods and applications*. Lansdowne, UK: Juta

Flattery, A. (2011). *Developing and validating a hostility, gratefulness and active support measuring instrument.* Unpublished master's dissertation. North-West University, Potchefstroom, South Africa.

Foxcroft, C. D. (1997). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal of Psychological Assessment, 13*(3), 229-235. doi:10.1027/1015-5759.13.3.229

Foxcroft, C. D. (2004). Planning a psychological test in the multicultural South African context. *SA Journal of Industrial Psychology, 30*(4), 8-15.

Foxcroft, C., Roodt, G., & Abrahams, F. (2005). Psychological assessment: A brief retrospective overview. In C. Foxcroft & G. Roodt (Eds.), *An introduction to psychological assessment in the South African context* (2nd ed.), (pp. 8-23). Cape Town, South Africa: Oxford University Press.

Foxcroft, C. D., & Davies, C. (2006). Taking ownership of the ITC's guidelines for computer-based and internet-delivered testing: A South African application. *International Journal of Testing, 6*(2), 173-180.

Foxcroft, C. D., Paterson, H., Le Roux, N., & Herbst, D. (2004). *Psychological assessment in South Africa: a needs analysis.* Retrieved from www.hsrc.ac.za/.../1716_Foxcroft_Psychologicalassessmentin%20SA.pdf

Foxcroft, C. D., Seymore, B. B., Watson, A. S. R., & Davies, C. (2002, September 24 - 27). *Towards building best practice guidelines for computer-based and Internet testing.* Paper presented at the 8th National Congress of the Psychological Society of South Africa, University of the Western Cape, Cape Town, South Africa.

Foxcroft, C. D., Watson, A. S. R., & Seymore, B. B. (2004, Augustus 8 - 13*). Personal and situational factors impacting on CBT practices in developing countries.* Paper presented at the 28th International Congress of Psychology, Beijing, China.

Freedman, R. D., & Stumpf, S. A. (1978). What can one learn from the Learning Style Inventory? *Academy of Management Journal, 21*(2), 275-282. doi:10.2307/255760

Gallagher, A., Bridgeman, B., & Cahalan, C., 2002. The effect of computer-based tests on racial- ethnic and gender groups. *Journal of Educational Measurement, 39*, 133–147.

Goldberg, A. L., & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement, 62*(6), 1053-1067.

Government Gazette (1998) (Employment Equity Act No. 55 of 1998), Republic of South Africa, Vol. 400, No. 19370. Cape Town, 19 October 1998.

Goodstein, L. D., & Lanyon, R. I. (1999). Applications of personality assessment to the workplace: A review. *Journal of Business Psychology, 13*(3), 291-322. doi:10.1023/A:1022941331649

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tathum, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Holtzhausen, G. (2004). *Mode of administration and the stability of the OPQ32n: Comparing internet (controlled) and paper-and-pencil (supervised) administration.* Unpublished master's thesis. University of Pretoria, Pretoria, South Africa.

Health Professions Council of South Africa (2012). *South African Guidelines on Computerised Testing.* Retrieved from http://www.hpcsa.co.za/downloads/psychology/Form_257.pdf

Huysamen, G.K. (1996). *Psychological measurement. An introduction with South African examples.* Pretoria, South Africa: J.L. van Schaik.

International Test Commission (2005). *International guidelines on computer-based and Internet delivered testing.* Retrieved from http://www.intestcom.org

Jacoby, J., & Mattel, M .S. (1989). Three-point Likert scales are good enough. *Journal of Marketing Research*, *8*(3), 205-230.

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers, 31*, 433-438.

Joubert, T., & Kriek, H. J. (2009). Psychometric comparison of paper-and-pencil and online personality assessments in a selection setting. *South African Journal of Industrial Psychology, 35(*1), 1-11. doi:10.4012/sajip.v35i1.727

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioural research (4th ed.).* London, UK: Wadsworth.

King, W. C., & Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology, 80*, 643-651. doi:10.1037/0021-9010.80.6.643

La Grange, L., & Roodt, G. (2001). Personality and cognitive ability as predictors of the job performance of insurance sales people. *Journal of Industrial Psychology, 27*, 35-43.

Lancaster, S. & Mellard, D. (2005). Adult learning disabilities screening using an internet-administered instrument. *Learning Disabilities: A Contemporary Journal, 3*(2), 62-73.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing, 6*(1), 1-24. doi:10.1207/s15327574ijt0601_1

Lunz, M. A., & Bergstrom, B. A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement*, *31*, 251-263.

Mazzeo, J., Druesne, B., Raffield, P. C., Checketts, K. T., & Muelstein, A. (1991). Comparability of computer and paper-and-pencil scores for two CLEP general examinations. *College Board Report No. 91-5.* New York. (ERIC Document Reproduction Service No. ED344902).

McKee, L. M., & Levinson, E. M. (1990). A review of the computerized version of the Self-Directed Search. *Career Development Quarterly, 38*(4), 325-333.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458. doi:10.1037/0033-2909.114.3.449

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

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (Eds.). (2002). *Computer-based testing: Building the foundations for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Upper Saddle River, NJ: Pearson Education Inc.

Nel, J. A. (2008). *Uncovering personality dimensions in eleven different language groups in South Africa: An exploratory study*. Unpublished doctoral dissertation. North-West University, Potchefstroom, South Africa.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications.* London, UK: Sage Publications, Inc.

Noyes, J. M. & Garland, K. J. (2008). Computer- vs. Paper-based tasks: Are they equivalent? *Ergonomics, 51*(9), 1352-1375.

Oosthuizen, T. H. (2012). *Developing and validating a measuring instrument for the relationship harmony personality cluster.* Unpublished master's dissertation. North-West University, Potchefstroom, South Africa.

Osborne, J. W. & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology, 1*, 1-7. doi:10.3389/fpsyg.2010.00220

Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5*(3), 23-57.

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

Pallant, J. (2007). SPSS Survival Manual (3rd ed.). New York, USA: Open University Press.

Parshall, C. G., & Kromrey, J. D. (1993, April). *Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics associated with mode effect*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Paterson, H., & Uys, K. (2005). Critical issues in psychological test use in the South African workplace. *South African Journal of Industrial Psychology, 31*(1), 12-22.

Pomplun, M., Frey, S., & Douglas, F. B. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337-354.

Pouwer, F., Snoek, F. J., Van Der Ploeg, H. M., Heine, R. J., & Brand, A. N. (1998). A comparison of the standard and computerised versions of the Well-Being Questionnaire (WBQ) and the Diabetes Treatment Satisfaction Questionnaire (DTSQ). *Quality of Life Research, 7*, 33-38. doi:10.1023/A:1008832821181

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*(3), 287-297.

Retief, A. I. (1988). *Method and theory in cross-cultural psychological assessment*. Pretoria, South Africa: Human Sciences Research Council.

Rosenfeld, P., Booth-Kewley, S., & Edwards, J. E. (1996). Responses to computer surveys: Impression management, social desirability, and the Big Brother syndrome. *Computers in Human Behavior, 12*(2), 263-274.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, *7*(20). Retrieved from http://epaa.asu.edu/ojs/article/view/555/678

Salgado, J. F., & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assessees' perceptions and reactions. *International Journal of Selection and Assessment, 11*(3/4), 194-205. doi:10.1111/1468-2389.00243

Schoech, D. (2001). Using video clips as test questions: The development and use of multimedia exam. *Journal of Technology in Human Services, 18*(3-4), 117–131.

Siloma, S. K., & Holden, R. R. (1992). Equivalence of computerized and standard administration of Piers-Harris Children's Self-Concept Scale. *Journal of Personality Assessment, 58*(2), 287-294. doi:10.1207/s15327752jpa5802_8

SPSS Inc. (2008). *SPSS 16.0 for Windows*. Chicago, IL: SPSS Inc.

Surís, A., Borman, P. D., Lind, L., & Kashner, T. M. (2007). Aggression, impulsivity, and health functioning in a veteran population: Equivalency and test-retest reliability of computerized and paper-and-pencil administrations. *Computers in Human Behavior, 23*, 97-110.

Tippins, N. T., Beary, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., et al. (2006). Unproctored internet testing. *Personnel Psychology, 59*(1), 189-225. doi:10.1111/j.1744-6570.2006.00909.x

Van der Linde, P. (2012). *South African Personality Inventory: Developing amiability, egoism and empathy scales for a soft-heartedness measuring instrument.* Unpublished master's dissertation. North-West University, Potchefstroom, South Africa.

Van der Merwe, R. P. (2002). Psychometric testing and human resource management. *South African Journal of Industrial Psychology, 28*(2), 77-86.

Van der Walt, H. S., Meiring, D., Rothmann, S., & Barrick, M. R. (2002, June). *Metaanalysis of the relationship between personality measurements and job performance in South Africa.* Paper presented at the Annual Conference of the Society for Industrial and Organisational Psychology of South Africa, Pretoria.

Van de Vijver, F. J. R., & Rothmann, S. (2004). Assessment in multicultural groups: The South African case. *South African Journal of Industrial Psychology, 20*(4), 1-7.

Van Dijk, J. (2006). The network society. Social aspects of new media (2nd ed). London, UK: Sage.

Vispoel, W. P. (2000). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement, 60*, 371-384.

Vispoel, W. P., Boo, J., & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg Self-Esteem Scale: A comparison of psychometric features and respondent preferences. *Educational and Psychological Measurement, 61*(3), 461-474. doi:10.1177/00131640121971329

Vorster, P. (2011). *A factor analytic comparison of dichotomous and polytomous response categories in personality inventories.* Unpublished master's thesis. University of Johannesburg, Johannesburg, South Africa.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5-24.

Wang, H., & Shin, C. D. (2009). Computer-based and paper-pencil comparability studies. *Test, Measurement and Research Services Bulletin, 9*, 1-6.

Watkins, M. W., Greenawalt, C. G. , & Marcell, C. M. (2002). Factor structure of the Wechsler Intelligence Scale for Children-Third Edition among gifted students. *Educational and Psychological Measurement, 62*(1), 164-172.

Welkenhuysen-Gybels, J. G. J., & Van de Vijver, F. J. R. (2001). A comparison of methods for the evaluation of construct equivalence in a multigroup setting. *2001 Proceedings. American Statistical Association.*

Wise**,** S. L. & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice, 8*(3), 5-10.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183.

CHAPTER 3

CONCLUSIONS, LIMITATIONS AND RECOMMENDATIONS

This chapter presents conclusions regarding the literature review and the empirical study according to the objectives of the current study. The limitations of the research are discussed, followed by recommendations for future research.

3.1 CONCLUSIONS

Taking into consideration the important role that personality testing plays in organisational processes such as intervention, promotion, educational placement and job selection (Foxcroft, 1997; Goodstein & Lanyon, 1999; Holtzhausen, 2005; Van der Merwe, 2002), as well as the rapidly changing demographics of the South African educational and job sector (Paterson & Uys, 2005), it is of the utmost concern that cross-culturally appropriate instruments be developed and validated for the diverse South African context (Nel, 2008). Joseph and Van Lill (2008) state that large inequalities still exist in South Africa's social and economic structure, and that variables such as language, race and social and educational background are therefore likely to influence individuals' test performance. Specific reference should therefore be made to the cross-cultural applicability as well as reliability and validity of psychological tests (Holzhausen, 2005). In response to the bias and poor item functioning of personality assessment measures used in South Africa (Nel, 2008), researchers are currently in the process of developing the South African Personality Inventory (SAPI) with the aim of fairly and effectively measuring personality across the eleven official language groups.

According to Zieky (2002) the concept of fairness is too complex to be proven by a single statistical method. The best way to ensure fairness in assessment is therefore to build fairness into the development, administration and scoring processes. An essential step towards the development of a new psychometric instrument is determining the optimal mode of administration that will lead to the most valid and reliable measure of the construct under investigation. The most popular forms of test administration currently being used in the fields of education and psychometrics are paper-and-pencil and computerised assessments. In a paper-and-pencil assessment test takers are required to "make a verbal or written response to a stimulus presented on paper or a verbal stimulus given by the test administrator" (Suris, Borman, Lind & Kaplan, 2007, p98). Computer-based testing, on the other hand, refers to selection instruments that are administered and scored via a computer (Davies, Foxcroft, Griessel & Tredoux, 2005; Tippins, Beaty, Drasgow, Gibson, Pearlman, Segall & Shepherd, 2006). Research from the past two decades suggests that the perceived benefits of computer-based testing outweigh that of paper-based

testing. In our fast-paced modern society these advantages are becoming all the more applicable. Based on the growing popularity of computer-based testing, it is reasonable to conclude that the concept of testing by computer has the endorsement of test developers, users and takers. However, the move to computer-based testing, although successful, has not been without some concerns. Questions continue to be raised on the reliability of the test delivery system, the psychometric quality of tests and the adequacy of theoretical models that support them (Mills, Potenza, Fremer, & Ward, 2002). In addition, many have raised concerns about the effects of differential access to technology and varying levels of computer familiarity on test performance (Barak, 2003; Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008; Foxcroft & Davies, 2006; Foxcroft, Watson, & Seymore, 2004; Goldberg & Pedulla, 2002).

Due to the clearly visible digital divide in developing South Africa (see Fuchs & Horak, 2008), concerns about the use of computer-based testing amongst technology unsophisticated test takers need to be addressed when deciding on the optimal mode of test administration. As a result of such ongoing concerns, research findings suggest that it may be preferable to make the SAPI available in both paper-and-pencil and computerised formats. Doing so will provide test-takers with the alternative of completing either a paper-based or computer-based test according to their preference. The primary objective of this study was to determine whether traditional paper-and-pencil and computerised assessment measures will lead equivalent results when testing facets from the soft-heartedness personality cluster on a dichotomous rating scale. Establishing equivalence is an important part of the test development process as it will allow future test administrators to use to use the different test modes interchangeably. Furthermore, the ability to use paper-based and computer-based test modes interchangeably within the South African context will ensure the alignment of the test mode with the socio-economic status and educational levels of the testee.

The first objective of the current study was to determine how the equivalence between paper-and-pencil and computerised assessment measures is conceptualised according to literature. A review of past literature shows that findings on the topic of equivalence between the two test modes have at large been inconclusive. While the majority of recent equivalence studies suggest that computer-based tests and paper-based tests are in fact comparable (Alexander & Davies, 2004; Arce-Ferrer & Guzman, 2009; Bartram & Brown, 2004; Holtzhausen, 2005; Joubert & Kriek, 2009; Pouwer, Snoek, Ploeg, Heine & Brand, 1998; Salgado & Moscoso, 2003; Simola & Holden, 1992; Vispoel, Boo & Bleiler, 2001; Wang, Jiao, Young, Brooks & Olsen, 2008), results have not been unanimous and several researchers have reported opposing results. Differences in terms of score distributions as well as psychometric properties between the two test modes have been found in various studies (Buchanan et al., 2005). Some researchers have reported higher test scores for computer-based tests (Bugbee & Bernt, 1990; Clariana & Wallace, 2002; Pomplun, Frey, & Becker, 2002), while others have found lower test scores for computer-based tests (Mazzeo, Druesne, Raffield, Checketts, & Muelstein, 1991; Russell, 1999). As a result of such

findings, researchers caution that the translation of paper-based questionnaires to a computer-based format represents a significant change in measurement which could affect reliability and result in inequivalent scores (McDonald, 2002; Webster & Compeua, 1996). According to Suris, Borman, Lind and Kasher (2007), these differences can be attributed to the differences in presentation mode and response requirements between the two test modes. Equivalence between different modes of administration should therefore be proven and not assumed.

Besides the psychometric and score equivalence between paper-based and computer-based tests, another main area of interest amongst researchers has been test taker attitudes towards the different test modes. Despite concerns about the potential negative effects of computer unfamiliarity and anxiety on test performance (Barak, 2003; Bennett et al., 2008; Davies et al., 2005; Foxcroft, Watson & Seymore, 2004; Goldberg & Pedulla, 2002; Mckee & Levinson, 1990), test takers have generally demonstrated a clear preference for CBT (Arce-Ferrer & Guzman, 2009; Barak & Cohen, 2002; Bugbee & Bernt, 1990; Butler, 2003; Vispoel et al., 2001). This is another important consideration to take into account when selecting the optimal test mode for psychometric instruments.

The second objective of the study was to determine equivalence using paper-and-pencil and computerised assessment measures. An evaluation of the descriptive statistics of the data showed that items from the paper-based format had superior item functioning when compared to items from the computerised format. Good item performance was indicated by mean scores smaller than 0.95, skewness values smaller than 2 and kurtosis values smaller than 4 (see DeCarlo, 1997). In the paper-based test mode ten items proved to be problematic in terms of mean scores, skewness and kurtosis. Such items were considered to be arbitrary to the purpose of the study and were consequently removed from further analysis. Items falling outside of the recommended cut-off points can potentially skew the data, thus justifying the exclusion of these items from further analysis. In the computerised test mode, fourteen items proved to be problematic in terms of item functioning. As a result, an additional four items were removed from further analysis. Items falling outside of the recommended cut-off points for mean scores, skewness and kurtosis are indicative of random response patterns amongst test takers. The computer-based test mode thus showed higher levels of random responding. The remaining ten items all fell between the desired cut-off points for good item performance. These items were therefore retained for factor analysis.

Particularly poor item performance was seen in items i018 ("I value the company of people close to me") and i023 ("I value pleasant experiences") in the computer-based format. A mean score of 1 can be found on both these items, indicating that all participants taking the computer-based test answered "yes" to the presented question. Findings such as these may suggest the presence of a social desirability element in these test items. Results further suggest that higher levels of socially desirable responding were present in

the computer-based test version when compared to response patterns on the paper-based test. These results correspond with findings from Rosenfeld et al. (1996) and Cronk and West (2002), who found higher levels of impression management in computer-based testing compared to paper-based testing. It was concluded that perceiving that one's responses are linked to a larger database may lead to more impression management on computer surveys. Cronk and West (2002) support this finding by stating that the problem of confidentiality needs to be considered in computer-based and especially Internet testing. Participants may feel uncomfortable providing information over a computer-based medium, because they believe that others may use the results. They may therefore respond differently than they would if they were certain their responses would be anonymous. On the other hand, item i012 ("When someone cries, I also feel like crying") on the paper-based format showed particularly good item performance with a mean score of 0.56. These results indicate a relatively equal distribution between "yes" and "no" answers on the presented item. Ideally, mean scores should fall between 0.4 and 0.6 for items to effectively measure the presence or absence of soft-heartedness. Due to the good item performance of item i012, developers of the SAPI should consider further investigation into the properties of this test item. Poorly performing test items could potentially be rephrased in a way which is similar to the above-mentioned item in order to achieve more optimal item performance.

Exploratory factor analysis was performed on the ten retained items. When the Eigenvalue greater than 1 rule is applied, both test modes retained four factors each. With four factors retained both the paper-based and computer-based test modes account for similar proportions of variance. According to soft-heartedness theory, the construct of soft-heartedness consists of three factors, namely "generous", "compassionate" and "appreciative" (Nakani, 2011; Nel, 2008). When the corresponding scree plots are investigated, the paper-based measure shows the existence of two distinct factors, while the computer-based scree plots show the existence of three distinct factors which corresponds with soft-heartedness theory. An investigation of factor loadings, with four factors retained, shows that all items loaded on more than one factor – indicating a lack of congruence in item loadings. Due to the fact that the highest item loadings were on the first factor, a single factor was retained. This factor was then further factor analysed. Another investigation of factor loadings with a single factor retained show that in the paper-based test mode 9 out of the 10 items loaded onto the first factor. In the computerised test mode, 8 out of the retained 10 items loaded onto the first factor. These results further support the existence of a single factor. The fact that a greater number of items from the paper-based test format loaded onto the retained factor again support the superiority of the paper-based test over its computer-based counterpart.

An investigation of factor loadings and communalities again showed that the paper-based format outperformed the computer-based format. Kerlinger and Lee (2000) suggest that factor loadings should be greater than 0.3 in order to adequately measure the underlying factor construct. In the current study, factor

loadings for the paper-based test mode ranged from 0.35 to 0.64, while factor loadings from the computer-based format ranged from 0.31 to 0.68. This demonstrates that the retained items were sufficiently associated with the factor structure under investigation and are therefore adequate measures of the soft-heartedness construct. Inter-item communalities ranged from 0.12 to 0.41 in the paper-based format, and from 0.10 to 0.46 in the computer-based format. The average communality for the paper-based mode was 0.253, with the computerised testing again scoring similarly with an average of 0.252. Theoretically, inter-item communalities should be greater than 0.2 to adequately measure the construct under investigation. With several items measuring below 0.2, results indicate low communalities between the test items and the soft-heartedness construct. Results such as these suggest that the combination of test items poorly accounts for variance in the soft-heartedness construct.

In the computer-based format, 4 out of the 10 items did not correlate with the construct under investigation, and also did not correlate to any of the other factors. Factors i002, i012, i015 and i020 had communalities of 0.1 and lower, and were therefore not measures of the soft-heartedness construct. It is possible that these items measure another unrelated construct. This again indicates the superiority of the paper-based format. The poor item loadings found on both modes of administration are indicative of the poor overall functioning of the two-point scale on both test modes. The current study forms part of the larger SAPI project. As a result of the poor functioning of the two-point rating scale, it has been decided to only make use of five point personality scales for the SAPI in the future.

The third objective of the study was to compare the reliability and validity of paper-and-pencil and computerised assessment measures. Reliability analysis again showed that the paper-based test format outperformed its computer-based counterpart. With 9 factors retained, the paper-and-pencil test had a reliability coefficient of 0.61. With 8 factors retained, the computerised test had a reliability coefficient of 0.59. An investigation of reliability coefficients with certain items removed, demonstrated an increase in reliability in the computer-based test with the removal of items i012 and i015. These items were consequently removed before testing again for the alpha coefficient. With 6 items retained, the reliability of the computer-based format increases to 0.61 – equal to that of the paper-based test mode. However, since the paper-based version retained more items, it is still more reliable than the computer-based version. Reliability on both modes of administration, however, fell below the recommended cut-off point of 0.7. This again indicates poor overall functioning on the 2-point rating scale.

Researchers provide several explanations for the poor performance of dichotomous rating scales. Jacoby and Mattel (1971) suggest that too few response categories result in a loss of much of the raters' discriminating powers. Polytomous rating scales are generally considered to better account for the variance in a variable and are often associated with greater levels of reliability and validity (Alwin, 1992;

Comrey, 1988; Netemeyer, Bearder, & Sharma., 2003). This is based on the assumption that test takers are better able to select their particular trait and make more accurate ranking decisions than on a dichotomous scale. Increasing the number of rating scale categories is therefore believed to increase the reliability, validity and discriminating ability of a scale (Alwin, 1992; Lui, Wu, & Zumbo, 2010; Netemeyer et al., 2003). Polytomous rating scales may therefore be able to define the structure of the soft-heartedness personality construct more clearly. Considerations such as these support the use of polytomous rating scales for future applications of the SAPI.

The final objective of the study was to make recommendations for future research. Recommendations for future research, as well as recommendations for the SAPI, will follow after a discussion of the limitations of the study.

3.2 LIMITATIONS

The current study was not without limitations. The following limitations could be identified:

Firstly, the use of convenience sampling meant that the population group was limited to university students from the North-West and Gauteng provinces. There were generally high levels of similarity between research participants in terms of age, level of education and socio-economic status. The study can therefore not be generalised to the broader South African population.

The focus of the current study was on establishing structural equivalence between the two modes of administration in terms of descriptive statistics, reliability, validity, communalities and factor structure. Due to the focus on psychometric functioning in a broader sense, the current study did not investigate the effect of factors such as cultural differences, race, language, socio-economic status and level of education on test functioning and mode equivalence. The current study also only focused on a single personality cluster, thus limiting the scope of the study. Furthermore, poor item functioning meant that only a small number of items could be retained for further analysis. Poor item functioning could potentially have been caused by poor item formulation as well random responding resulting from a lack of motivation from test takers. Poor item performance in terms of mean scores further indicates that a social desirability element may have been present in a number of test items. Socially desirable responding may therefore have affected test performance especially on the computer-based format.

Research suggests that self-efficacy, perceived ease of use and computer anxiety are potential moderating factors in equivalence studies involving computer-based testing (Davies et al., 2005; Wang, 2004; Watson & Seymore, 2004). However, limiting research participants to university students, raised in the so-called

"era of technology", is likely to have eliminated the effect of computer anxiety to a large extent. More significant differences between the two modes of administration could potentially have been found had a more representative sample group been tested.

3.3 RECOMMENDATIONS

In the face of these limitations, the current study has important implications for the organisation and for future research.

3.3.1 Recommendations for the organisation

The purpose of this study was to determine whether equivalence exist between PPT and CBT when administered amongst demographically dissimilar test takers in South Africa. It is clear from the results that small, yet potentially significant, mode effects exist between PPT and CBT in the South African context. Due to the popularity of psychological testing in the organisational setting, a greater awareness should be created concerning the need for equivalence testing. PPT and CBT should be proven to be comparable within its applied context before the two modes may be used interchangeably.

In addition, test takers with lower levels of computer-familiarity should, if possible, be offered the choice of completing a paper-based questionnaire. Where test takers do not have the option to choose between PPT and CBT a CBT tutorial should be presented prior to testing in order to familiarise the test-taker with the computerised test format and to minimise potential mode effects.

3.3.2 Recommendations for future research

Findings from the current study may have important implications for future research within the field of I/O psychology and for the development of the SAPI. Firstly, the use of convenience sampling meant that results from the current study could not be generalised for the broader population. Future studies should therefore aim to include a more diverse sample group, representative of the broader South African population. Researchers should specifically strive to include participants from more widely dispersed age groups, socio-economic backgrounds, and levels of education. More in-depth investigations of the possible causes of non-equivalence between paper-based and computer-based test modes are also recommended for the future. Future studies could study the effects of factors such as cultural differences, race, language, socio-economic status and level of education on test functioning and mode equivalence.

Researchers suggest that, more specifically, factors such as age, socio-economic status and level of education are potential moderating factors in equivalence studies comparing paper-based and computerised modes of assessment. The effect of such factors can be attributed to the possible influence of computer familiarity, computer anxiety and perceived ease of use on test results in computerised testing. Older participants, and those from less privileged backgrounds and lower levels of education, would be assumed to have lower levels of computer familiarity and perceived ease of use, and potentially higher levels of anxiety concerning the use of computers in assessment. Factors such as these could have potentially negative effects on results attained via computerised testing. Future studies could therefore investigate the effects of computer familiarity and perceived ease of use on computer-based test performance in the South African context.

The focus of the current study was specifically on establishing structural equivalence between the paper-based and the computerised modes of administration. It was therefore sufficient to only include items from the soft-heartedness personality construct. A large number of items, however, had to be excluded from further analysis due to poor item functioning, thus limiting the scope of the study. Items may need to be revised and formulated differently in order to better measure the construct of soft-heartedness. Poor item functioning in terms of mean scores also suggested the possible influence of a social desirability element on test takers' responses. Further investigation into the social desirability of test items is therefore recommended. Future studies could also consider including a more widespread array of personality constructs to measure the equivalence between two comprehensive personality questionnaires in the South African context.

Another important area of concern in test mode selection is preference. Paterson and Uys (2005) state that although a test or method may be scientifically proven as fair, valid and reliable, the perception of the method of assessment can still be perceived as negative by the testee. Perceiving a test more favourably may have a potential positive impact on test performance. As research with regard to mode preference is currently lacking in South Africa, this is an important consideration for future research.

Due to the poor overall performance of the two-point rating scale, it is recommended that a polytomous rating scale be used for future applications of the SAPI. Poor item performance from the two-point scale has not only been demonstrated in the current study, but also in a number of related studies. As a result, it was decided to move away from the two-point scale and only make use of polytomous rating scales in the further development of the SAPI.

Lastly, future studies can be performed comparing the equivalence of a greater number of test formats, for example internet-based administration as opposed to supervised computer-based testing. Since the main

focus of the SAPI will be for selection purposes, different formats can also be considered in order to broaden the scope of the mode of administration and making it more applicable to a diverse target group, including the disabled and the poorly literate. The possibility of developing a personality inventory that can be used fairly and reliably for participants from all levels of ability will be a beneficial advancement for the field of South African psychometrics.

REFERENCES

Alwin, D. F. (1992). Transmission in the survey interview: number of response categories and the reliability of attitude measurement. *Sociological Methodology*, *22*, 83-118.

Arce-Ferrer, A. J., & Guzman, E. M. (2009). Studying the equivalence of computer-delivered and paper-based administrations of the Raven Standard Progressive Matrices Test. *Educational and psychological measurement*, *69*(5), 855-867. doi:10.1177/0013164409332219

Barak, A. (2003). Ethical and professional issues in career assessment on the internet. *Journal of career assessment, 11*(1), 3-21.

Barak, A. & Cohen, L. (2002). Empirical examination of an online version of the self-directed search. *Journal of career assessment, 10*(4), 387-400. doi:10.1177/1069072702238402

Bartram, D., & Brown, A. (2004). Online testing: Mode of administration and the stability of the OPQ32i scores. *International Journal of Selection and Assessment, 12*(3), 278-284. doi:10.1111/j.0965-075X.2004.282_1.x

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on a computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*(9). Retrieved from http://www.jtla.org.

Buchanan, T., Ali, T., Heffernan, T. M., Ling, J., Parrott, A. C., Rodgers, J., & Scholey, A. B. (2005). Non-equivalence of online and paper-and-pencil psychological tests: The case for the prospective memory questionnaire. *Behaviour Research Methods, 37*(1), 148-154.

Bugbee, A. C. & Bernt, F. M. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education, 23*(1), 87-100.

Busch, M. (1993). Likert scales in L2 research: a researcher comments. *Tesol Quarterly*, *27*(4), 733-736.

Butler, D. L. (2003). *The impact of computer-based testing on student attitudes and behaviour. The Technology Source*, *Jan/Feb 2003*. Retrieved from http://ts.mivu.org/default.asp?show=article&id=1013

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, *6*(3), 170-175.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology, 33(*5), 593-602. doi:10.1111/1467-8535.00294

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, *17*(4), 407-422.

Cronk, B. C., & West, J. L. (2002). Personality research on the internet: A comparison of web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers*, *34*(2), 177–180.

Davies, C., Foxcroft, C., Griessel, L., & Tredoux, N. (2005). Computer-based and internet delivered assessment. In C. Foxcroft & G. Roodt (Eds.), *An introduction to psychological assessment in the South African context* (2nd ed.), (pp 153-166) Cape Town: Oxford University Press.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*(3), 292-307. doi:10.1037/1082-989X.2.3.292

Foxcroft, C. D. (1997). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal of Psychological Assessment, 13*(3), 229–235. doi:10.1027/1015-5759.13.3.229

Foxcroft, C. D., & Davies, C. (2006). Taking ownership of the ITC's guidelines for computer-based and internet-delivered testing: A South African application. *International Journal of Testing, 6*(2), 173-180.

Foxcroft, C. D., Watson, A. S. R., & Seymore, B. B. (2004). *Personal and situational factors impacting on CBT practices in developing countries.* Paper presented at the 28th International Congress of Psychology, Beijing, China, 8-13 Augustus 2004.

Fuchs, C., & Horak, E. (2008). Africa and the digital divide. *Telematics and Informatics, 25,* 99-116.

Goldberg, A. L., & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement, 62*(6), 1053-1067.

Goodstein, L. D., & Lanyon, R. I. (1999). Applications of personality assessment to the workplace: A review. *Journal of Business Psychology, 13*(3), 291-322. doi:10.1023/A:1022941331649

Holtzhausen, G. (2005). *Mode of administration and the stability of the OPQ32n: Comparing internet (controlled) and paper-and-pencil (supervised) administration.* Unpublished master's thesis. University of Pretoria, Pretoria, South Africa.

Jacoby, J., & Mattel, M .S. (1989). Three-point Likert scales are good enough. *Journal of Marketing Research*, *8*(3), 205-230.

Joseph, L., & Van Lill, B. (2008). Investigating subscale differences among race and language groups on the Occupational Personality Profile. *South African Journal of Psychology, 38*(3), 501-514.

Joubert, T., & Kriek, H. J. (2009). Psychometric comparison of paper-and-pencil and online personality assessments in a selection setting. *South African Journal of Industrial Psychology, 35(*1), 1-11. doi:10.4012/sajip.v35i1.727

Liu, Y., Wu, A. D., & Zumbo B. D. (2010). The impact of outliers on Cronbach's coefficient alpha estimate of reliability: ordinal/rating scale item responses. *Educational Psychological Measurement*, *70*(1), 5-21.

Mazzeo, J., Druesne, B., Raffield, P. C., Checketts, K. T., & Muelstein, A. (1991). Comparability of computer and paper-and-pencil scores for two CLEP general examinations. *College Board Report No. 91-5.* New York. (ERIC Document Reproduction Service No. ED344902).

McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computer & Education, 39*(3), 299-312.

McKee, L. M., & Levinson, E. M. (1990). A review of the computerized version of the Self-Directed Search. *Career Development Quarterly, 38*(4), 325-333.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (Eds.). (2002). *Computer-based testing: Building the foundations for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Nakani, L. (2011). *The factor analytic equivalence of paper-based and computerised personality scales*. Unpublished master's thesis. University of Johannesburg, Johannesburg, South Africa.

Nel, J. A. (2008). *Uncovering personality dimensions in eleven different language groups in South Africa: An exploratory study*. Unpublished doctoral dissertation, North-West University, Potchefstroom, South Africa.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications.* London, UK: Sage Publications, Inc.

Paterson, H., & Uys, K. (2005). Critical issues in psychological test use in the South African workplace. *South African Journal of Industrial Psychology, 31*(1), 12-22.

Pomplun, M., Frey, S., & Douglas, F. B. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*(2), 337-354.

Pouwer, F., Snoek, F. J., Van Der Ploeg, H. M., Heine, R. J., & Brand, A. N. (1998). A comparison of the standard and computerised versions of the Well-Being Questionnaire (WBQ) and the Diabetes Treatment Satisfaction Questionnaire (DTSQ). *Quality of Life Research, 7*, 33-38. doi:10.1023/A:1008832821181

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, *7*(20). Retrieved from http://epaa.asu.edu/ojs/article/view/555/678

Salgado, J. F. & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assesses' perceptions and reactions. *International Journal of Selection and Assessment, 11*(3/4), 194-205. doi:10.1111/1468-2389.00243

Siloma, S. K., & Holden, R. R. (1992). Equivalence of computerized and standard administration of Piers-Harris Children's Self-Concept Scale. *Journal of Personality Assessment, 58*(2), 287-294. doi:10.1207/s15327752jpa5802_8

Surís, A., Borman, P. D., Lind, L., & Kashner, T. M. (2007). Aggression, impulsivity, and health functioning in a veteran population: Equivalency and test-retest reliability of computerized and paper-and-pencil administrations. *Computers in Human Behavior, 23*, 97-110.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing. *Personnel Psychology, 59*(1), 189-225. doi:10.1111/j.1744-6570.2006.00909.x

Truell, A. D., Alexander, M. W., & Davis, R. E. (2004). Comparing Post-secondary Marketing Students' Performance on Computer-Based and Handwritten Essay Tests. *Journal of Careers and Technical Education*, *29*(2), 69-78.

Van der Merwe, R. P. (2002). Psychometric testing and human resource management. *South African Journal of Industrial Psychology, 28*(2), 77-86.

Vispoel, W. P., Boo, J., & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg Self-Esteem Scale: A comparison of psychometric features and respondent preferences. *Educational and Psychological Measurement, 61*(3), 461-474. doi:10.1177/00131640121971329

Vorster, P. (2011). *A factor analytic comparison of dichotomous and polytomous response categories in personality inventories.* Unpublished master's thesis. University of Johannesburg, Johannesburg, South Africa.

Wang, H., & Shin, C. D. (2009). Computer-based and paper-pencil comparability studies. *Test, Measurement and Research Services Bulletin, 9*, 1-6. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/93727FC9-96D3-4EA5-B807-5153EF17C431/0/Bulletin_9.pdf

Wang, S. (2004). *Online or paper: Does delivery affect results? Administration mode comparability study for Stanford Diagnostic Reading and Mathematics Tests.* Pearson Assessment Report. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/D381C2EA-18A6-4B52-A5DC-DD0CEC3B0D40/0/OnlineorPaper.pdf

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5-24.

Webster, J., & Compeau, D. (1996). Computer-assisted versus paper-and-pencil administration of questionnaires. *Behavior Research Methods, Instruments & Computers, 28*, 567-576.

Zieky, M. (2002). Ensuring the fairness of licensing tests. *CLEAR Exam Review*, *xii*(1), 20.