Psychometric comparison of paper-and-pencil and online personality assessments in a selection setting

Authors:
Tina Joubert[1]
Hendrik J. Kriek[2]

Affiliations:
[1]SHL, South Africa

[2]Department of Industrial and Organisational Psychology, University of South Africa, South Africa

Correspondence to:
Tina Joubert

e-mail:
tina@shl.co.za

Postal address:
PO Box 1305, Groenkloof, 0027, South Africa

Keywords:
online assessment; personality questionnaire; equivalence; Occupational Personality Questionnaire; paper-and-pencil assessment

Dates:
Received: 10 July 2008
Accepted: 03 Mar. 2009
Published: 09 June 2009

How to cite this article:
Joubert, T., & Kriek, H.J. (2009). Psychometric comparison of paper-and-pencil and online personality assessments in a selection setting. *SA Journal of Industrial Psychology/SA Tydsrkif vir Bedryfsielkunde 35*(1), Art. #727, 11 pages. DOI: 10.4102/sajip.v35i1.727

This article is available at:
http://www.sajip.co.za

ABSTRACT

The goal of the study was to determine whether the Occupational Personality Questionnaire (OPQ32i) yielded comparable results when two different modes of administration, namely paper-and-pencil and Internet-based administration, were used in real-life, high-stakes selection settings. Two studies were conducted in which scores obtained online in unproctored settings were compared with scores obtained during proctored paper-and-pencil settings. The psychometric properties of the paper-and-pencil and Internet-based applications were strikingly similar. Structural equation modelling with EQS indicated substantial support for the hypothesis that covariance matrices of the paper-and-pencil and online applications in both studies were identical. It was concluded that relationships between the OPQ32i scales were not affected by mode of administration or supervision.

INTRODUCTION

Since the development of Kraepelin's free association test in 1892, an early forerunner of personality testing, psychological assessment has advanced dramatically (Anastasi & Urbina, 1997). Not only have a variety of assessments specifically developed for the world of work emerged, but psychological assessment has evolved into the use of information technology and technological innovations. The development of computers in the twentieth century moved the focus from traditional standardised paper-and-pencil tests to computer-based testing. Computerisation has strongly influenced psychological research and practice (Mead & Drasgow, 1993). The latest trend has been made possible by the growth of the Internet as a generic communication medium (Bartram, 2001a). Salgado and Moscoso (2003) predicted that the next 10 years would be the decade of the Internet in personnel selection.

In a review of online instruments Brown, Bartram, Holtzhausen, Mylonas and Carstairs (2005) found that most of the major traditional paper-and-pencil personality questionnaires are available on the Internet. Cronbach (1990) points out the necessity of determining the equivalence of scale scores on a measuring instrument when it is used in more than one mode. It is therefore important to investigate the psychometric properties of traditional paper-and-pencil tests when they are adapted for use on the Internet. A literature review of previous research appears to indicate that measurement equivalence between web-based and paper-and-pencil tests is generally established (Bartram & Brown, 2002; Bartram & Brown, 2004; Brown *et al*., 2005; Buchanan & Smith, 1999; Mylonas & Carstairs, 2003; Templer, 2005; Trippe, 2005). However, most of the earlier research was performed under laboratory conditions. Consequently, it is important to investigate the equivalence of Internet-based and paper-and-pencil measures in real applicant contexts (Ployhart, Weekley, Holtz & Kemp, 2003). To address this need, the current study focuses on how people behave under real rather than laboratory conditions. It also investigates the effect of supervised versus unsupervised settings.

Computerised and Internet-based testing

The use of computers has increased dramatically since its introduction in educational and psychological assessment (Davis, 1999; Skinner & Pakula, 1986; Van de Vijver & Harsveld, 1994; Vispoel, Boo & Bleiler, 2001; Webster & Compeau, 1996). The popularity of computer-based testing can be attributed to the automation of test administration and the almost immediate scoring and interpretation of results (Buchanan & Smith, 1999; Mead & Drasgow, 1993). Computer-based testing can also be applied such that there are no missing data, that data entry is automated and that there are no out-of-range responses (Cronk & West, 2002; Rosenfeld, Booth-Kewley & Edwards, 1993).

In order to take advantage of the benefits of computer-based testing, test publishers converted traditional paper-and-pencil tests for administration via the computer. According to Bartram (2000), the available computer-based tests are mostly computer adaptations of existing paper-and-pencil tests. However, Webster and Compeau (1996) caution that when paper-and-pencil tests are converted to any other mode of administration, one should be aware that the change in procedure may lead to inequivalence of scores. One should therefore assess the equivalence of scores collected by the paper-and-pencil method and the new method.

The use of computer-based testing has been reported in the literature since 1963 (King & Miles, 1995). Concerns that medium effect size differences between means on the various applications might exist were already discussed in 1968 at the Computer-Assisted Testing Conference in the United States (USA) (Mead & Drasgow, 1993). Various studies on the equivalence of paper-and-pencil and computerised administrations of psychological assessments have been published since then (Donovan, Drasgow & Probst, 2000; King & Miles, 1995; Mead & Drasgow, 1993; Rosenfeld *et al*., 1993; Skinner & Pakula, 1986; Van de Vijver & Harsveld, 1994; Vispoel *et al*., 2001; Webster & Compeau, 1996). Most of the research, with the exception of some individual studies, supported the measurement equivalence of computer and paper-and-pencil modes of administration. Conflicting results pertained mostly to speeded tests, whereas equivalence was demonstrated for non-speeded non-cognitive measures (Bartram, 2001b; Bartram & Bayliss, 1984; Evans, 2002; Stanton, 1998).

Technological innovations in the computer industry have led to the development of better interfaces and dramatic increases in the volume of and accessibility to hardware. The Internet is currently being used widely and access from home and office is readily available. This has had a major impact on the way assessment and testing is carried out (Bartram, 2000). 'The interconnectivity reached through the Internet is leading to a revolution in testing and assessment' (Evans, 2002, p. 3). The Internet offers many of the same benefits as computer-based testing and can provide added benefits, but it also leads to concerns. Evans (2002, p. 7) asks the question: 'Can equivalence on personal computer be generalized to the Internet?'

The growth of the Internet has, in recent years, been phenomenal (Bartram, 2000; Buchanan & Smith, 1999; Drasgow, 2005). The Internet was originally developed in the 1950s mainly for use by academics and the military as a medium for communicating and transferring files. It was the development of the World Wide Web in 1992 that led to the widespread use of the Internet (Bartram, 2001a). It has been estimated that the number of people worldwide with access to the Internet (Internet usage statistics – the big picture, 2008) reached 100 million (1 114 274 426) in 2007. This is a 208.7% increase from 2000 to 2007. In Africa the Internet usage growth increased by 638.4% from 2000 to 2007, and in South Africa by 112.5%. A consequence of these developments was that psychologists realised the potential that the Internet presented for a range of activities. Their interest was originally focused on research as a result of increased access to information (Buchanan & Smith, 1999; Davis, 1999), but the possibilities that it offered for various human resource functions were quickly noticed. In recent years many test publishers have developed personality questionnaires, surveys, inventories and cognitive ability tests to be administered on the Internet (Evans, 2002). Schrage concludes that:

> *the Internet is the greatest medium ever invented for conducting standardized tests. Any company, or any executive, believing in the value of testing for intellectual acuity or emotional stability or managerial potential is going to treat the Net as an irresistible opportunity to poke into people's psyches.*
>
> (Schrage 1999, p. 170)

Over the past five years there has been a marked increase in employment tests available on the Internet for recruitment, selection and development. Unfortunately, these also include tests that are not scientifically validated. The Internet made it possible for people to bring their tests to the world, but it appears that the number of people with programming skills exceeds the number of those with the psychological expertise required to design tests (Buchanan & Smith, 1999). Drasgow (2005) notes that The Association of Test Publishers had grown from less than 150 in 2000 to 700 in 2005, but that some of these businesses were dominated by sales and information technology specialists and not by psychometricians and industrial psychologists. Anecdotal evidence of this increase of instruments available on the Internet is the search performed by Kriek (2005) on the World Wide Web by using the key words 'personality tests'. He obtained 5 880 000 hits on Google and 3 560 000 hits on Yahoo. He repeated the search in 2006 and obtained 26 300 000 hits on Google and 8 990 000 hits on Yahoo. Some of these hits included for example The Sense of Humor Test, The Neighbor Test, The Paranoid Test and The Drama Queen Test. It is therefore evident that there is an enormous amount of personality tests on the Internet, but it is not known how many of these are scientifically sound.

In spite of its potential disadvantages/misuses, the Internet offers many advantages for psychological assessment. These include enhanced convenience for the applicants with its 24 hour a day access, cost-effectiveness and a shortened hiring cycle (Cappelli, 2001; Chapman & Webster, 2003; Cronk & West, 2002; Potosky & Bobko, 2004; Salgado & Moscoso, 2003; Tippins *et al.*, 2006). Internet-based platforms have a broader reach and therefore allow testing in rural settings where it is difficult to travel to a testing centre. Tests are also easier to maintain, because norms can be changed readily and new translations or items can be made available around the world almost instantly (Jones & Dages, 2003; Kriek, 2005; Tippins *et al.*, 2006). The Internet also leads to integrated personnel assessment platforms with the potential to vastly improve the communication between employers and candidates (Holtzhausen, 2004).

Another advantage of the Internet is the positive reaction of candidates towards Internet assessments. In a study conducted by Evans (2002), investigating candidates' perceptions toward Internet assessments, it was found that the candidates expressed a strong preference for the Internet versions of the assessment. They perceived that online testing caused less anxiety and found it easier to use the mouse compared to a pencil. Similar studies performed by Potosky and Bobko (2004) and Salgado and Moscoso (2003) also found that Internet-based tests are more positively perceived than the paper-and-pencil tests. Dowdeswell (2006) conducted a study in South Africa in which an online survey questionnaire was administered to an applicant pool following the applicants' completion of online assessments as part of a selection process. The objective of the study was to determine how the applicants perceived the online assessments in general and whether they considered them to be fair. Overall, online assessments were perceived as fair by South African graduate applicants.

There are also some drawbacks and potential problems when using Internet-based assessments. The first major issue involves the quality of assessments available on the Internet. As discussed previously, a range of tests is made available by unqualified persons and reports are generated and feedback given on the results (Kriek, 2005). With Internet assessments, the problem of confidentiality must also be considered. Participants may feel uncomfortable providing information over the Internet, because they believe that others may see the results (Cronk & West, 2002). There are also some technical issues relating to the speed and bandwidth of the Internet, as well as network integrity and reliability (Jones & Dages, 2003; Tippins *et al*., 2006).

Another concern expressed by Tippins *et al*. (2006) is that there are no studies that have examined measurement equivalence for paper-and-pencil versus unproctored Internet test administration for previously disadvantaged groups. Such groups may experience the Internet medium as difficult. Also, if disadvantaged groups have less access to the Internet, recruitment and selection via the Internet raise ethical and legal concerns (Tippins *et al*., 2006). Kriek (2006) uses the selection process of a South African financial company in which open-mode online assessments were used as an initial screening tool for operational positions as an example. There were 66 706 applicants that entered the website over a five-year period to complete the assessments, of which 64.81% (n = 43 231) were African, 11.67% (n = 7 782) were Coloured, 8.58% (n = 5 726) were Indian and 14.94% (n = 9 967) were White. From this it appears that previously disadvantaged groups in South Africa have sufficient access to the Internet. Foxcroft (2009) reports on two studies performed to assess the differences in test performance between students with different levels of computer familiarity. The results of the first study indicated that there was 0.5 to 1 standard deviation difference between the two groups. In a follow-up study a few years later, however, it was found that there was no longer any significant difference between the two groups. She ascribed this phenomenon to the increasing access to technology such as cell phones by everyone in South Africa, even in rural areas and recommended that the impact of technology be monitored over an extended period and that the results of one study not be blindly followed.

'The very same factors that lead to the ease and convenience of data collection via the Internet are also responsible for the varied testing conditions possible under web-based measurement' (Trippe, 2005, p. 8). Unproctored testing, in which the applicant completes a test battery without direct supervision, holds some of the disadvantages associated with Internet testing. Internet testing almost ensures a lack of standardised testing conditions. Candidates are subject to any number of physical and psychological variables and are more distracted than those taking tests under proctored conditions (Tippins *et al*., 2006; Trippe, 2005). Bartram (2001a) identifies issues of test security, authentication of users and the protection of test results as potential problems. To a large extent, these issues can be addressed by controlled access to the assessments via passwords and honesty contracts. Most of these drawbacks can be overcome or reduced via testing standards and end-user training (Jones & Dages, 2003).

> *Psychological assessment is guided by published standards of professional principles for test construction. These standards are revised periodically to reflect the latest professional developments in the field of psychological assessment.*
>
> (Muchinsky, 2004, p. 175)

Professional associations, such as the International Test Commission (ITC), the American Psychological Association (APA) and the Association for Test Publishers (ATP), started to develop guidelines and principles to address the quality and control of assessments on the Internet. Changing the medium of assessment does not change the basic requirements of testing, such as reliability and validity (Drasgow, 2005). SHL South Africa (2006) conducted a global survey on the acceptance of and current legislation on Internet testing. A questionnaire containing items on the general acceptance and legislation with regard to Internet testing was emailed to the managing directors of SHL offices in 27 countries, including Australia, Belgium, Botswana, France, Italy, Israel, the Netherlands, Russia, Sweden, the United Kingdom (UK) and the USA. SHL offices from 21 countries responded to the questionnaire, including those listed above. Only 10% of the SHL offices indicated that they have a statutory body or laws on regulating the use of psychological tests in their countries. Internet testing also appears to be generally well accepted by 57.89% of the countries (e.g. Australia, Portugal, Switzerland, the UK and the USA), with only medium acceptance by 5.26% of countries, but these countries indicate definite growth potential. The Internet is not widely used by 36.84% of countries (e.g. Greece, Italy and Israel).

The ITC responded to the rapid growth in computer- and Internet-based testing by developing the International Guidelines on Computer-based and Internet Delivered Testing, Version 1.0, January 2005. The first draft of this document was completed in March 2003, after which it went through a consultation period and a number of revisions. The ITC Council formally accepted the latest version of the document in July 2005. There are four general issues that form the basis of the development of the guidelines:

- Technology – ensuring that the technical aspects of CBT/Internet testing are considered, especially in relation to the hardware and software required to run the testing
- Quality – ensuring and assuring the quality of testing and test materials and ensuring good practice throughout the testing process
- Control – controlling the delivery of tests, test taker authentication and prior practice
- Security – security of testing materials, privacy, data protection and confidentiality (Bartram & Coyne, 2005, pp. 3–4).

The ITC Guidelines (Bartram & Coyne, 2005) identify four modes of test administration:

- Open mode – anyone can access and complete the test without supervision. No user identification is necessary.
- Controlled mode – no direct supervision of the session, but access to the tests can only be obtained via logon name and password.
- Supervised mode – certain degree of supervision. The supervisor handles the login process and can verify the test taker’s identification.
- Managed mode – high level of supervision and a controlled test-taking environment.

According to Drasgow (2005), the APA entered the fray by establishing a Taskforce on Internet Testing in 2001 that made three recommendations for the use of Internet-based tests:

- Traditional psychometric standards (reliability and validity) should apply to Internet tests
- The validity of inferences from the many diverse types of Internet tests must be demonstrated
- Internet site authors should be accountable so that users receive the same type of protection as in traditional testing, e.g. through the use of test manuals and norming studies.

Most of the issues regarding Internet testing relate to the level of supervision (Kriek, 2005). Common issues underlying unproctored/unsupervised Internet-based testing include the identity of the test taker, security of the test content, cheating and the monitoring of the test administration (Tippins *et al*., 2006; Weiner, Hayes, Reynolds & Doverspike, 2005). The online controlled mode of administration will become more secure as new technologies become embedded, such as remote identification and authentication through thumbprint, retinal eye pattern and voice recognition. Strategies to increase test security include test delivery in kiosk mode (controlled navigation, disabling of system commands during the test), encrypting data transmission and password logins (Kriek, 2005).

With regard to cheating, Schrage (1999) states that wherever there is testing, there will be cheating. Tippins *et al*. (2006) declare:

> *Applicants have successfully cheated on non-Internet tests and will no doubt try to cheat on Internet tests and succeed. Cheating is not a death-knell for non-Internet tests, nor should it be for Internet tests. The relevant question is how to minimize its effects.*
>
> (Tippins *et al.* 2006, p. 207)

To control cheating, unsupervised tests can be used as a pre-screen, followed by a supervised retest for the final shortlist. Access to the tests can be controlled giving single-test access to a candidate with limited time available. The use of honesty contracts with a warning of follow-up assessments and the consequences of cheating or breeching security can be used (Kriek, 2005; Tippins *et al*., 2006).

In spite of the above-mentioned disadvantages of Internet and specifically unproctored Internet testing, there is a growing body of evidence supporting the use of Internet-based testing for psychological assessments. Lievens and Harris (2003) reviewed available studies and concluded that there was initial evidence of measurement equivalence between Internet and paper-and-pencil testing. Davis (1999) examined responses to the Ruminative Responses Scale (RRS) administered via Internet and paper-and-pencil formats and found that individuals in the Internet sample reported significantly higher levels of rumination. He interpreted this finding to mean that Internet-based measures may encourage more candid responding due to perceived anonymity. The internal consistency of the RRS was similar across administration modes. Buchanan and Smith (1999) studied the Internet-based test and paper-and-pencil versions of the Self-Monitoring Scale measure (SMS-R). They found that the psychometric properties of the Internet-based version compared favourably with its paper-and-pencil equivalent. Potosky and Bobko (2004) studied the cross-mode equivalence of an untimed situational judgement test administered on the Internet (proctored) as well as by paper-and-pencil (proctored) and found promising cross-mode equivalence. Trippe (2005) evaluated the measurement invariance of unsupervised Internet and supervised paper-and-pencil forms of the International Personality Item Pool (IPIP) designed to parallel the Five Factor Model scales of the Revised NEO Personality Inventory (NEO-PI-R). His findings suggest that it is safe to assume equivalence and comparability between Internet and paper-and-pencil versions of this personality inventory. Pasveer and Ellard (1998) investigated the psychometric properties of a Self-Trust

Questionnaire (STQ) administered via the Internet (unproctored) and paper-and-pencil modes. They concluded that measures of internal consistency as well as the factor structure for the STQ were similar for the respective administration modes.

There are only a limited number of studies in which personality questionnaires have been researched under real-life conditions, and not in a laboratory setting. Of these studies even fewer studies were done in a high-stakes selection environment. Mylonas and Carstairs (2003) compared unsupervised Internet and supervised computer administrations of the Motivation Questionnaire (MQ). They attempted to create a high-stakes environment by requesting participants to complete the MQ as they would if they were applying for a position of employment. Their conclusion was that the data supported the equivalence of the two administration modes. Ployhart *et al*. (2003) researched the equivalence of proctored Internet-based and paper-and-pencil tests of assessment instruments in a real-life selection setting. The Internet-based measures showed better distributional properties, lower means, higher variance, higher internal consistency reliabilities and stronger intercorrelations.

In another study that simulated a high-stakes selection environment, Salgado and Moscoso (2003) examined whether the paper-and-pencil version of a Big Five personality questionnaire could be adapted for the Internet without loss of psychometric properties. They tried to create a real-life situation by informing the participants (students) that the results would be used for selecting individuals for an assessment centre training course. The data showed that the two versions were equivalent in terms of distributions, reliability and factor structure. Bartram and Brown (2004) explored the equivalence of Internet administration with no supervision and supervised paper-and-pencil versions of the Occupational Personality Questionnaire 32i (OPQ32i) (ipsative format) involving samples tested for development and selection. They found that scale reliabilities and covariances appeared to be unaffected by the differences between the administration conditions. Holtzhausen (2004) obtained similar results on the Occupational Personality Questionnaire 32n (the normative version). Templer (2005) examined the equivalence of proctored and unproctored Internet testing in a combined laboratory-field condition. The results did not provide evidence that the test conditions affected the test results.

In the case of some of the abovementioned studies, it was found that the samples that had completed the assessments on the Internet obtained larger standard deviations than their paper-and-pencil counterparts. Pasveer and Ellard (1998) and Salgado and Moscoso (2003) obtained similar mean scores for the different modes of administration and larger standard deviations, whereas Buchanan and Smith (1999), Ployhart *et al*. (2003) and Potosky and Bobko (2004) found that Internet testing resulted in lower mean scores, but larger standard deviations.

The current study was carried out to provide further evidence regarding the equivalence of a personality questionnaire administered unproctored on the Internet and proctored with paper-and-pencil in a selection setting, because it is clear that Internet testing is gaining ground. Chapman and Webster (2003) studied the use of technology in recruitment and selection by conducting a survey and interviewing human resource managers who were members of the Society for Human Resource Management. Every company they interviewed had plans for increasing their reliance on technology-based approaches to recruitment and selection, as it would reduce hiring cycles and costs. If companies are intent on using technology for recruitment and selection, they should have access to instruments that are scientifically proven to be reliable and valid.

It is essential to conduct comparison studies if paper-based tests converted to Internet-based tests are to be considered valid and fair measures (Jones & Dages, 2003).

Cronbach (1990) also strongly advises that the equivalence of different versions of the same instrument be researched when different modes are to be used. Traditional paper-and-pencil tests and their Internet derivatives should be highly and significantly correlated before Internet-based versions may be used with confidence. 'Comparability studies are essential if paper-based tests converted into technology-friendly tests are to be considered valid and fair measures' (Jones & Dages, 2003, p. 249). The ITC Guidelines on Computer-based and Internet Delivered Testing (2005) make it clear that where the Internet-based test has been developed from a paper-and-pencil version of the same test, equivalence testing becomes important. Section 2 (c) of the ITC Guidelines (2005) states that 'where the CBT/Internet test has been developed from a paper-and-pencil version, ensure that there is evidence of equivalence' (p. 11). It 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 (p. 11).

The present study addressed two of the four requirements – comparable reliabilities and intercorrelation between subtests. Correlation with external criteria and comparable means and standard deviations were not addressed – because independent groups were used the latter was not possible given the present information.

In seven of the equivalence studies discussed previously, independent samples were asked to complete the paper-and-pencil and Internet-based tests (Bartram & Brown, 2004; Buchanan & Smith, 2003; Davis, 1999; Holtzhausen, 2004; Pasveer & Ellard, 1998; Ployhart *et al*., 2003; Trippe, 2005). Although these authors noted that this increased the possibility of alternative explanations for their results, circumstances in most instances made it impossible to use the same group for both modes of administration. Salgado and Moscosso (2003) express their concern about the fact that most of the equivalence studies done to date have been on independent groups of examinees. It is however notable that all the studies irrespective of the sample found cross-mode equivalence.

In this study the equivalence of a personality test was evaluated in two different studies in which the assessments were done as part of a real-life, high-stakes personnel selection and decision-making process. Based on the literature review, the hypothesis can be stated that there will be no differences between the psychometric properties of the unsupervised online personality assessment and the supervised paper-and-pencil personality assessment as measured by the Occupational Personality Questionnaire (OPQ).

RESEARCH DESIGN

Research approach

This research can be categorised as quasi-experimental quantitative research. The aim of this research is to investigate the construct equivalence of the Occupational Personality Questionnaire 32 (OPQ32i) when it is administered in an online and paper-and-pencil mode of administration. The results were analysed using the statistical methods of mean differences, reliabilities and structural equation modelling by means of the SPSS and EQS 6.1 programs. The online samples used were samples of convenience, whereas the paper-and-pencil samples were drawn randomly from an existing database to reflect the biographical data of the online samples.

Research method

Design

Two separate studies were conducted in which samples had to complete an Internet version of the OPQ32i in an unproctored, controlled selection context in South Africa. In the first study, the sample consisted of students in their final year of study or students who had already completed a degree and who had applied for various positions at a financial institution. In the second study, the Internet sample consisted of managers at a transporting company who had applied for leadership training. For both studies, the results of the Internet testing were compared with scores obtained by samples randomly drawn from an existing database of proctored paper-and-pencil OPQ32i testing results of applicants tested for various positions in different industry sectors.

As mentioned previously, Salgado and Moscoso (2003) express their concern with previous research that compared the similarity of the responses on the paper-and-pencil and Internet-based tests using independent groups of examinees. They state that:

> *this means that the actual equivalence was not directly examined, because the failure to detect differences between groups does not mean that the two versions are really equivalent, as such a failure could be due to other causes (e.g. a third variable or the situation).* (Salgado & Moscoso, 2003, p. 200)

In the researched literature it is evident that most of the previous studies did indeed use independent groups for the different modes of administration. However, in an attempt to address Salgado and Moscoso's concerns in the current research, the Internet-based samples used were drawn from high-stakes selection settings and the paper-and-pencil samples were randomly drawn to reflect the biographical data of the online samples with respect to age, gender, ethnicity and education. These biographical variables were identified as a result of the research done by Bartram, Brown, Fleck, Inceoglu and Ward (2006), which indicated that small to medium differences were found on certain personality scales between the gender and ethnic groups. Males scored higher on Persuasive, Controlling, Data Rational, Innovative, Relaxed, Tough Minded and Competitive, while the females scored higher on Outgoing, Affiliative, Caring, Behavioural, Detail Conscious, Conscientious, Worrying and Vigorous. When gender differences were investigated using a Five-Factor Model, the men scored higher on Emotional Stability and the women on Agreeableness and Conscientiousness. This is similar to the results reported by Costa, Terracciano and McCrae (2001) in that across countries, females scored higher on Neuroticism and Agreeableness.

Small score differences were also found for the different ethnic groups in a South African sample. Black candidates described themselves to be more Outspoken and less Controlling, Affiliative and Modest than the White candidates. Age was also controlled for as it was shown that younger people scored slightly higher, with a very small effect size, on certain of the OPQ32 scales (Outgoing, Affiliative and Achieving).

Although Bartram *et al*. (2006) explain that none of the differences found were large and that they reflect mostly 'real but minor differences in the typical style of members of different groups in a work setting', it was thought important to control for these differences (p. 209). Following this process, no statistically significant differences between these biographical variables of the two groups were found. The samples to be compared were therefore assumed to have similar biographical information. It should however be noted that there are other third variables, such as job type and job level, that cannot be controlled for as a result of unavailability of data and that this is a limitation of the current study.

Participants

Study 1

For the first study the Internet sample of students (n = 1091) consisted of 512 (46.93%) males and 579 (53.07%) females. The mean age of the candidates was 23.14 (SD = 2.36) with a range from 19 to 35. In terms of ethnic distribution, the sample consisted of 586 (53.71%) Africans, 239 (20.91%) Indians, 43 (3.94%) coloured people and 223 (20.44%) white people. Their highest qualifications included Grade 12 (n = 116; 10.63%), certificate (n = 32; 2.93%), degree (n = 612; 56.10%) and postgraduate degree (n = 331; 30.34%).

The randomly drawn paper-and-pencil sample (n = 1 136) proportionately reflects the graduate Internet sample and included 495 (43.57%) males and 641 (56.43%) females. The mean age of the candidates was 23.70 (SD = 2.62) with a range from 18 to 35. In terms of ethnic distribution, the sample consisted of 650 (57.22%) Africans, 198 (17.43%) Indians, 47 (4.14%) coloured people and 241 (21.21%) white people. Their highest qualifications included Grade 12 (n = 150; 13.20%), certificate (n = 29; 2.55%), degree (n = 615; 54.14%) and postgraduate degree (n = 342; 30.11%).

The Internet sample and the paper-and-pencil sample were compared on age (p = 0.275), gender (p = 0.112), education (p = 0.287) and ethnicity (p = 0.68) and no statistical significant difference between the two groups on any of the biographical variables was found. It can be accepted that these variables will not have any effect on our dependant variable, namely the possible differences in the psychometric properties of the OPQ32i based on the results of the two groups' performance on the OPQ.

Study 2

For the second study the Internet sample of managers (n = 1 159) consisted of 852 (73.51%) males and 307 (26.49%) females. The mean age of the candidates was 42.89 (SD = 8.99) with a range from 24 to 65. In terms of ethnic distribution, the sample consisted of 383 (33.05%) Africans, 94 (8.11%) Indians, 94 (8.11%) coloured people, 581 (50.13%) white people and seven (0.60%) that indicated another ethnicity. Their highest qualifications included Grade 12 (n = 208; 17.95%), certificate (n = 145; 12.51%), degree (n = 495; 42.71%) and postgraduate degree (n = 311; 26.83%).

The randomly drawn paper-and-pencil sample to compare with the Internet sample of managers (n = 950) included 662 (69.68%) males and 288 (30.32%) females. The mean age of the candidates was 41.67 (SD = 9.52) with a range from 24 to 64. In terms of ethnic distribution, the sample consisted of 325 (34.21%) Africans, 81 (8.53%) Indians, 80 (8.42%) coloured people, 462 (48.63%) Whites and two (0.21%) that indicated another ethnicity. Their highest qualifications included Grade 12 (n = 202; 21.26%), certificate (n = 130; 13.68%), degree (n = 404; 42.53%) and postgraduate degree (n = 214; 22.53%).

Regarding the biographical variables age (p = 0.052), gender (p = 0.244), ethnicity (p = 0.643) and education (p = 0.06), no statistically significant differences were found between the Internet and paper-and-pencil samples. It can be accepted that these variables will not have any effect on our dependant variable, namely the possible differences in the psychometric properties of the OPQ32i based on the results of the two groups' performance on the OPQ.

Measuring instrument

The ipsative version of SHL's OPQ32 was used in this study. The OPQ series of personality questionnaires, containing normative and ipsative versions, were designed to provide information on individual styles or preferences at work. The structure of the OPQ includes three broad domains, namely Relationships with People, Thinking Style, and Feelings and Emotions, which can be subdivided into 32 dimensions (Bartram *et al*., 2006). The OPQ32i consists of 416 items arranged in 104 blocks of four statements each of which the test taker has to choose one item as 'Most like me' and one item as 'Least like me'.

Evidence supporting the job-related validity of the OPQ instruments has been reported in a number of studies (e.g. Robertson & Kinder, 1993; Saville, Sik, Nyfield, Hackston & MacIver, 1996). The British Psychological Society (BPS) reviewed the OPQ32 and stated that it is at the top of personality tests (Marshall & Lindley, 2007). The review resulted in the OPQ32 obtaining the highest possible ratings for quality, plus a high rating for reliability (Marshall & Lindley, 2007).

The ipsative version of the OPQ32 was used in the selection process, because it was specifically designed to counter the effects of response distortion. The items of the normative version of the OPQ32 are more transparent and open to socially desirable responses in which the applicants may attempt to present themselves in a more favourable light. Bank and Ramsey (2001) found in a study conducted to examine the effect that unproctored Internet administration may have on responses to work styles instruments that the applicants tend to present themselves in a socially desirable way. They concluded that alternative measures such as dichotomised or ipsative response formats that may not be as prone to impression management as normative scales should be used. In another study, Richman, Kiesler, Weisband and Drasgow (1999) report that the mode of administration had little effect on socially desirable responses, but they conclude that the context (e.g. selection or development) within which the testing takes place had an effect.

With regard to the statistical analysis of ipsative data, Baron (1996) concludes that the artificial dependence (when raw scale scores sum to a constant for an individual) between ipsative scores does affect their psychometric properties, especially when the instrument has few scales. However, with the OPQ32i's 32 scales and 416 items, 'ipsative measurement does provide some interpretable psychometric parameters' (Baron, 1996, p. 55). Bartram (1996) and Saville and Wilson (1991) report that calculating reliabilities on ipsative measures will result in lower internal consistency reliabilities than their respective normative counterparts. For factor analysis, it is apparent that a much larger number of scales will be needed before the results resemble those provided by normative data (Baron, 1996). For this reason, factor structures of the OPQ32i were not compared in this study, but scale covariances were compared where one scale was removed from the correlation matrix in order to reduce the degrees of freedom to equal the number of scales (Bartam & Brown, 2004). This is done to free some variance, because the variance of the OPQ32i as an ipsative instrument is constrained.

Procedure

Participants in the student Internet sample were drawn from a graduate recruitment programme hosted by a South African financial institution. A hurdle approach was adopted to screen and select the graduates from this programme. Approximately 3 400 candidates applied online through the company's website and completed a structured application form. As part of the application process, candidates completed a behavioural questionnaire (i.e. an online interview) designed to screen for styles of behaviour that were identified as important for the role. Once the deadline for applications had been reached, the 20% most suitable applicants were invited for further assessments. These candidates were then invited to do online verbal and numerical competency screening tests and to complete the OPQ32i online. This was done in a controlled assessment mode, because it would not require any direct supervision of the session and allow candidates to complete where and when it would be comfortable for them. Access to the tests was controlled via logon name and password.

TABLE 1
Descriptive statistics and estimated effect sizes of differences between means of graduate supervised paper-and-pencil and controlled Internet-based OPQ32i results (study 1)

OPQ SCALES	SUPERVISED PAPER-AND-PENCIL (N=1136)				CONTROLLED INTERNET-BASED (N=1091)				
	MEAN	SD	SKEWNESS	KURTOSIS	MEAN	SD	SKEWNESS	KURTOSIS	d
Persuasive	12.67	4.92	0.04	-0.33	13.55	4.59	-0.01	-0.40	0.18
Controlling	12.68	5.07	0.13	-0.6	13.7	4.68	-0.06	-0.44	0.21
Outspoken	14.04	4.50	-0.09	-0.34	14.22	4.04	-0.01	-0.3	0.04
Independent minded	11.49	3.54	0.04	-0.01	9.78	3.47	0.2	-0.14	0.47
Outgoing	10.85	4.55	0.34	-0.2	10.55	4.33	0.51	0.08	0.07
Affiliative	12.11	4.09	0.19	-0.03	11.09	3.94	0.38	0.36	0.25
Socially confident	12.73	4.07	0.04	-0.2	13.6	3.81	0.04	-0.2	0.22
Modest	12.07	4.29	0.27	-0.08	11.08	4.45	0.39	-0.07	0.23
Democratic	14.59	3.72	0	-0.3	15.33	3.72	-0.04	-0.28	0.20
Caring	14.80	3.89	-0.09	-0.25	14.46	3.72	-0.08	-0.11	0.09
Data rational	15.48	5.14	-0.28	-0.3	15.44	6.41	-0.31	-0.79	0.01
Evaluative	14.12	3.57	0.25	-0.21	15.45	3.5	0.08	-0.25	0.37
Behavioural	11.97	4.50	0.31	-0.25	11.81	4.87	0.46	-0.17	0.03
Conventional	11.96	3.49	0.22	0.22	10.6	3.61	-0.08	-0.23	0.38
Conceptual	12.77	3.63	0.25	0.37	13.34	3.86	0.2	-0.03	0.15
Innovative	12.31	4.63	0.22	-0.09	14.63	4.91	-0.15	-0.27	0.47
Variety seeking	12.22	4.04	0.3	-0.06	12.31	3.95	0.36	-0.13	0.02
Adaptable	10.92	4.71	0.38	-0.36	9.79	4.84	0.48	-0.26	0.24
Forward thinking	14.96	4.13	-0.14	-0.41	16.16	4.18	-0.39	-0.1	0.29
Detail conscious	14.76	3.79	-0.17	0.02	15.04	3.81	-0.26	-0.08	0.07
Conscientious	17.83	3.80	-0.43	0.03	19.9	3.15	-0.49	0.23	0.57
Rule following	15.31	5.10	-0.35	-0.38	14.56	4.97	-0.17	-0.26	0.15
Relaxed	11.73	3.99	0.16	0.1	10.87	3.6	0.21	0.66	0.23
Worrying	9.04	4.77	0.41	-0.35	6.62	4.21	0.72	0	0.52
Tough minded	12.31	4.04	0	-0.13	11.9	3.86	0.18	0.14	0.10
Optimistic	15.62	3.86	-0.27	0	16.12	3.84	-0.21	-0.14	0.13
Trusting	8.58	4.37	0.45	0.21	8.66	4.04	0.2	-0.24	0.02
Emotional control	10.32	4.05	0.31	0.12	8.16	3.65	0.28	0.17	0.54
Vigorous	14.10	3.87	-0.05	-0.38	15.24	3.82	-0.09	-0.2	0.29
Competitive	13.12	5.15	-0.04	-0.44	13.01	4.83	0.07	-0.36	0.02
Achieving	17.90	3.55	-0.47	0.26	19.45	3.04	-0.59	0.2	0.46
Decisive	10.66	4.58	0.44	-0.25	9.6	4.73	0.56	0	0.23
Consistency	22.78	7.44	-0.11	-0.4	26.52	6.26	-0.4	0.33	0.54

TABLE 2
Descriptive statistics and estimated effect sizes of differences between means of managerial supervised paper-and-pencil and controlled Internet-based OPQ32i results (study 2)

OPQ SCALES	SUPERVISED PAPER-AND-PENCIL (N=950)				CONTROLLED INTERNET-BASED (N=1159)				
	MEAN	SD	SKEWNESS	KURTOSIS	MEAN	SD	SKEWNESS	KURTOSIS	d
Persuasive	13.46	5.13	-0.12	-0.66	11.25	5.45	0.31	-0.57	0.41
Controlling	14.72	5.26	-0.18	-0.52	14.37	5.31	-0.18	-0.6	0.07
Outspoken	13.95	4.56	-0.15	-0.39	13.83	4.5	-0.12	-0.43	0.03
Independent minded	11.56	3.83	0.17	-0.07	12.18	3.83	0.18	-0.14	0.16
Outgoing	9.36	4.71	0.54	0.1	8.36	4.2	0.57	0.01	0.22
Affiliative	11.12	4.39	0.48	0.12	9.83	4.21	0.5	0.16	0.30
Socially confident	12.19	4.41	0.04	-0.29	11.11	4.05	0.21	-0.17	0.25
Modest	12.57	4.56	0.22	-0.38	13.15	4.72	0.12	-0.46	0.12
Democratic	14.81	3.98	0.01	-0.31	14.97	3.85	0.02	-0.46	0.04
Caring	15.28	4.07	-0.09	-0.21	14.14	4.01	0.01	-0.31	0.28
Data rational	14.11	5.46	-0.02	-0.58	14.82	6.11	-0.14	-0.85	0.12
Evaluative	14.49	3.65	0.05	-0.32	15.81	3.52	0.03	-0.34	0.36
Behavioural	13.05	4.29	0.16	-0.25	11.63	4.38	0.36	-0.18	0.32
Conventional	11.91	4.11	0.08	-0.38	12.04	4.38	0.24	-0.24	0.03
Conceptual	11.91	4.13	0.33	0.2	13.14	4.29	0.28	0	0.29
Innovative	13.42	4.90	0.03	-0.37	14.7	4.95	0.03	-0.46	0.26
Variety seeking	12.69	3.92	0.29	-0.17	14.29	4.3	0.12	-0.39	0.38
Adaptable	11.32	4.69	0.24	-0.4	10.43	4.52	0.38	-0.15	0.19
Forward thinking	14.56	4.55	-0.08	-0.48	16.25	4.25	-0.34	-0.13	0.38
Detail conscious	14.21	4.59	-0.1	-0.35	14.13	4.31	-0.18	-0.17	0.02
Conscientious	17.56	3.92	-0.49	-0.01	18.52	3.59	-0.52	0.23	0.26
Rule following	13.36	5.49	-0.11	-0.66	13.24	5.91	0.06	-0.83	0.02
Relaxed	11.11	4.22	0.07	-0.22	10.17	3.93	0.18	0.23	0.23
Worrying	8.67	4.89	0.57	-0.19	8.73	4.81	0.47	-0.34	0.01
Tough minded	11.36	3.76	0.15	0.16	10.73	3.55	0.21	0.05	0.17
Optimistic	15.69	4.01	-0.15	-0.39	16.14	4.08	-0.26	-0.3	0.11
Trusting	10.19	4.58	0.28	-0.06	10.12	4.71	0.22	-0.29	0.01
Emotional control	10.82	4.22	0.54	0.3	10.42	4.23	0.44	0.11	0.10
Vigorous	14.07	3.76	-0.04	-0.17	14.51	3.9	-0.08	-0.29	0.12
Competitive	13.22	5.28	-0.04	-0.58	13.14	4.89	-0.08	-0.57	0.01
Achieving	16.44	3.80	-0.32	-0.18	16.94	3.6	-0.5	0.03	0.13
Decisive	12.82	5.15	0.14	-0.61	12.89	5.02	0.11	-0.64	0.01
Consistency	23.89	6.84	-0.14	-0.04	25.79	6.25	-0.25	-0.07	0.29

Participants in the Internet sample of managers were drawn from a leadership development programme hosted by a South African transport company. The aim was to identify employees that would be able to lead the organisation through a structural change and that were best fitted for development as future leaders of the company. Employees from junior, middle, senior and executive management levels were invited to complete the online assessment. The candidates were assessed by means of the online OPQ32i version and were asked to complete online verbal and numerical ability tests.

RESULTS

The main focus of the study was to determine the degree of equivalence of the OPQ32i administered unproctored on the Internet and proctored by means of the paper-and-pencil version in a real-life, high-stakes selection setting. This was achieved by comparing the samples on their mean scores, reliabilities and analysis of covariance structures for both studies. The ITC (2005) provides guidance for the use of Internet-based tests and suggests that the equivalence between paper-and-pencil and Internet-based tests be proven where the Internet version was developed from the paper-based version. It is also proposed that the mean scores and standard deviations as well as the reliabilities be comparable between modes of administration.

The first step in the analysis of the results explored possible differences between the comparison samples by examining their distributions (skewness and kurtosis), intercorrelations between the variables, and differences between the sample means expressed in terms of estimated effect sizes. In the second step, the internal consistency reliabilities of the instruments under supervised and unsupervised conditions were evaluated, and lastly, the effect of the unproctored Internet-based administration on the pattern of scale intercorrelations was examined using EQS.

Distributions, intercorrelations and mean differences

Study 1 (Graduates)

In Table 1 the means, standard deviations, skewness and kurtosis of the paper-and-pencil and Internet-based applications for Study 1 are presented. The results indicate that the distributions do not deviate substantially from the normal distribution. The absolute values of the skewness and kurtosis statistics were less than one in all instances.

The intercorrelations between the OPQ32i scales for the traditional paper-and-pencil sample for Study 1 ranged from -0.37 to 0.42, with 86% falling between -0.20 to 0.20. For the unproctored Internet-based sample, the OPQ32i scale intercorrelations ranged from -0.35 to 0.39, with 90% falling between -0.20 and 0.20.

Estimated effect sizes (d-statistics) were used to assess the magnitudes of the differences between the scale means when the supervised paper-and-pencil administration results were compared with the Internet-based controlled administration results. This means that mean differences were assessed in practical terms by expressing them in pooled standard deviation units using the d-statistic (Cohen, 1988). The values of *d* for Study 1 are reported in the final column of Table 1.

The effect sizes for Study 1 ranged from very small (0.02) to medium (0.57). The medium effect size differences found between some of the scales (Conscientious, Worrying, Emotional Control and Consistency) may be due to actual sample differences (see Table 1). Although great care was taken to equate the samples in terms of their biographical information, no previous work experience data were available. The Internet-based sample consisted solely of young graduates with possibly no work experience applying for positions, whereas the paper-and-pencil sample more likely consisted of candidates with previous work experience.

TABLE 3
Internal consistency reliabilities and standard errors of measurement for Studies 1 and 2

	STUDY 1				STUDY 2			
	PAPER-AND-PENCIL		INTERNET-BASED		PAPER-AND-PENCIL		INTERNET-BASED	
	Alpha	SEM	Alpha	SEM	Alpha	SEM	Alpha	SEM
Persuasive	0.79	2.25	0.78	2.15	0.81	2.24	0.85	2.11
Controlling	0.82	2.15	0.8	2.09	0.83	2.17	0.84	2.12
Outspoken	0.73	2.34	0.7	2.22	0.75	2.28	0.75	2.25
Independent Minded	0.60	2.24	0.61	2.17	0.65	2.26	0.64	2.3
Outgoing	0.76	2.23	0.76	2.12	0.79	2.16	0.75	2.1
Affiliative	0.76	2.00	0.76	1.93	0.80	1.96	0.81	1.84
Socially Confident	0.70	2.23	0.71	2.05	0.76	2.16	0.74	2.07
Modest	0.75	2.14	0.8	1.99	0.78	2.14	0.78	2.21
Democratic	0.60	2.35	0.65	2.2	0.66	2.32	0.67	2.21
Caring	0.69	2.16	0.72	1.97	0.73	2.12	0.74	2.04
Data Rational	0.83	2.12	0.91	1.92	0.86	2.04	0.89	2.03
Evaluative	0.60	2.26	0.6	2.21	0.62	2.25	0.61	2.2
Behavioural	0.74	2.30	0.81	2.12	0.72	2.27	0.75	2.19
Conventional	0.63	2.12	0.69	2.01	0.72	2.17	0.76	2.15
Conceptual	0.61	2.27	0.66	2.25	0.69	2.3	0.71	2.31
Innovative	0.81	2.02	0.84	1.96	0.83	2.02	0.85	1.92
Variety Seeking	0.69	2.25	0.69	2.2	0.67	2.25	0.7	2.35
Adaptable	0.78	2.21	0.82	2.05	0.78	2.2	0.78	2.12
Forward Thinking	0.73	2.15	0.77	2	0.79	2.08	0.77	2.04
Detail Conscious	0.63	2.31	0.67	2.19	0.74	2.34	0.7	2.36
Conscientious	0.72	2.01	0.69	1.75	0.75	1.96	0.74	1.83
Rule Following	0.84	2.04	0.86	1.86	0.86	2.05	0.89	1.96
Relaxed	0.71	2.15	0.7	1.97	0.75	2.11	0.72	2.08
Worrying	0.82	2.02	0.81	1.84	0.84	1.96	0.83	1.98
Tough Minded	0.67	2.32	0.68	2.19	0.64	2.26	0.61	2.22
Optimistic	0.69	2.15	0.73	1.99	0.72	2.12	0.72	2.16
Trusting	0.81	1.90	0.8	1.81	0.82	1.94	0.84	1.88
Emotionally Controlled	0.72	2.14	0.74	1.86	0.74	2.15	0.75	2.12
Vigorous	0.67	2.22	0.7	2.09	0.68	2.13	0.72	2.07
Competitive	0.79	2.36	0.79	2.21	0.81	2.3	0.79	2.24
Achieving	0.64	2.13	0.6	1.92	0.67	2.18	0.66	2.1
Decisive	0.77	2.19	0.8	2.11	0.81	2.24	0.81	2.19
Mean	**0.72**	**2.18**	**0.74**	**2.04**	**0.75**	**2.16**	**0.76**	**2.12**

Study 2

In Table 2 the means, standard deviations, skewness and kurtosis of the paper-and-pencil and Internet-based applications for Study 2 are presented. Similar to findings reported for Study 1, the results indicate that the distributions do not deviate substantially from the normal distribution. The absolute values of the skewness and kurtosis statistics were less than one in all instances.

For Study 2 the intercorrelations between the OPQ32i scales for the traditional paper-and-pencil sample ranged from -0.38 to 0.49, with 82% falling between -0.20 to 0.20. For the unproctored Internet-based sample for Study 2, OPQ32i scale intercorrelations range from -0.38 to 0.49, with 80% falling between -0.20 and 0.20. As for Study 1, these results point to the relative independence of the OPQ32i scales.

The values of *d* for Study 2 are reported in the final column of Table 2. The effect sizes for this study were generally substantially lower than for Study 1 and ranged from very small (0.01) to below medium (0.41).

Internal consistency reliabilities

The Cronbach alpha coefficients of the OPQ32i scales for the two modes of administration are reported separately for the two studies in Table 3. Generally, alpha coefficients with magnitudes between 0.60 and 0.80 are considered reasonable for personality instruments (SHL South Africa, 2004). The alpha coefficients and standard errors of measurement (SEM) for both studies are given in Table 3. The SEM values are based on scale raw scores.

For Study 1, the alpha coefficients for the supervised paper-and-pencil sample ranged from 0.60 to 0.84, with a mean alpha of 0.72, a median alpha of 0.73, and a mean SEM of 2.18. For the Internet-based sample, alpha coefficients ranged from 0.60 to 0.91, with a mean alpha of 0.74, a median alpha of 0.74, and a mean SEM of 2.04.

Similar results were obtained for Study 2 (see Table 3). The alpha coefficients for the supervised paper-and-pencil sample ranged from 0.62 to 0.86, with a mean alpha of 0.75, a median alpha of 0.75, and a mean SEM of 2.16. For the Internet-based sample, the alpha coefficients ranged from 0.61 to 0.89, with a mean alpha of 0.76, a median alpha of 0.75, and a mean SEM of 2.12.

It is clear from Table 3 that for both studies the mean and median alpha coefficients for the scales are very similar for the paper-and-pencil and Internet applications, indicating that administration mode does not compromise scale reliability. With regard to SEM the results indicate that they were marginally smaller for the Internet-based sample than for the paper-and-pencil sample.

Similarity of scale intercorrelations

Comparison of the covariance structures of the two samples was carried out using structural equation modelling with EQS 6.1 (Bentler, 2006). The scale covariance structures of the paper-and-pencil and Internet applications for the two samples of studies 1 and 2 were directly compared with prior removal of one scale from the correlation matrix to free variance, so that the degrees of freedom became equal to the number of scales. The models tested were that the covariance matrices were identical, firstly for the two samples in Study 1, and secondly, for the two samples in Study 2.

There are a number of statistics that measure how adequately the hypothesised model describes the data. The first is the Maximum Likelihood Chi-square where a significant value is indicative of a poor fit between the two groups. The second, the Comparative Fit Index (CFI), ranges from zero to 1.00 and provides a measure of complete covariance in the data and a cut-off value of 0.95 is considered representative of a well-fitting model (Bentler, 2006). Another fit measure is the Root Mean Square Error of Approximation (RMSEA) (Bentler, 2006). This index has been recognised as one of the most informative criteria in covariance structure modelling where values of less than 0.05 indicate good fit (Byrne, 2006).

The CFI obtained for Study 1 (graduates) was 0.985 and the RMSEA was equal to 0.015. A significant Chi-square of 705.99 with df = 465 was obtained for Study 1. The CFI obtained for Study 2 (managers) was equal to 0.993 and the RMSEA 0.012. The Chi-square (594.66; df = 465) for Study 2 was also found to be significant. The CFI and RMSEA indices represent an exceptionally good fit for both studies and imply that the relationships between the OPQ32i scales may be considered equivalent for the different application formats of the instrument. The significant Chi-square values found in both studies are an indication of either the large size of the samples or a poor fit between the two modes of administration (Byrne, 2006).

The above results indicate that the paper-and-pencil and controlled Internet-based versions of the OPQ32i yielded comparable psychometric properties in terms of reliability and covariance structures. The null hypothesis that the OPQ32i scale reliabilities and covariances are unaffected by the differences between the supervised and unsupervised administration conditions cannot be rejected.

DISCUSSION

The main focus of this study was to determine psychometric equivalence of the OPQ32i administered unproctored on the Internet and proctored with paper-and-pencil in a real-life, high-stakes selection setting. The analyses of the data supported the psychometric equivalence of these two administration modes of the OPQ32i. Therefore, the unproctored Internet-based measure of the OPQ32i had similar psychometric properties to the traditional paper-and-pencil instrument. Given the overall findings of the two studies, the general hypothesis that there will be no differences in the psychometric properties of the two modes of administration could not be rejected.

Previous research indicates that measurement equivalence between Internet-based and paper-and-pencil modes of administration is generally established. However, two concerns were raised with these studies. Firstly, Ployhart *et al*. (2003) point out that most of the research has been performed under laboratory conditions and not in real applicant contexts. Secondly, Salgado and Moscoso (2003) express their concern with research that compares the similarity of the responses on the paper-and-pencil and Internet-based tests using independent groups of examinees. In an attempt to address these issues in the current research, the Internet-based and paper-and-pencil samples used in this study were drawn from high-stakes selection settings. In order to address the independence of the settings of where the data were collected, the paper-and-pencil samples were randomly drawn to reflect the biographical data of the online samples with respect to age, gender, ethnicity and education. Although equivalent in terms of the biographical variables, these samples are still independent, and this presents a limitation for this study. However, the possible impact of biographical data was controlled for. Salgado and Moscoso's (2003) concern that there are still extraneous variables that could not be controlled for could not be addressed.

The first step in the analysis investigated differences between the samples by examining their distributions, mean differences and the intercorrelations between scales. In both studies the skewness and kurtosis values did not deviate substantially from the normal distribution curve, because all the values were smaller than 1.00.

The means and standard deviations of the Internet-based and paper-and-pencil assessment scores in this research were very similar. The mean score differences, when expressed in terms of effect size, ranged from very small to medium. This is similar to what Mylonas and Carstairs (2003) and Ployhart *et al*. (2003) found. The largest difference (d = 0.57) was found in Study 1 (graduates). Although the samples were drawn to reflect similar biographical details, the graduate sample presented an interesting issue. The graduate Internet-based sample consisted solely of inexperienced candidates, whereas the work experience of the paper-and-pencil-based sample was unknown. The mean differences (Independent Minded, Innovative, Conscientious, Worrying, Emotional Control and Achieving) obtained between the samples may therefore have been due to actual sample differences rather than being the result of mode of administration.

Although previous research found that Internet-based assessments yielded larger standard deviations than their paper-and-pencil counterparts (Buchanan & Smith, 1999; Pasveer & Ellard, 1998; Potosky & Bobko, 2004; Salgado & Moscoso, 2003), this was not supported in the current study. The standard deviations for both studies were very similar. The foregoing results imply that the same norms can be used for both modes of administration.

Intercorrelations between the OPQ32i scales are a very important consideration for multi-scale instruments such as the OPQ32i. These intercorrelations indicate how closely related different constructs are to one another and they support the construct validity of an instrument (Bartram *et al*., 2006). High intercorrelations among instruments or scales reduce the unique variance explained by each scale (Ployhart *et al*., 2003). The low intercorrelations between the OPQ32i scales for both the Internet-based and paper-and-pencil applications suggest a high degree of independence for the scales.

Ployhart *et al*. (2003) found that the Internet-based measures used in their study yielded higher intercorrelations than the paper-and-pencil measures. The intercorrelations obtained in the current research between the OPQ32i scales proved to be very similar for the Internet-based and paper-and-pencil samples for both studies 1 and 2. For studies 1 and 2 respectively, 90% and 80% of the coefficients fell between -0.20 and 0.20, and for the paper-and-pencil samples 86% and 82%. The range of correlations is similar for both modes of administration and does not support the finding of Ployhart *et al*. (2003).

The internal consistency reliabilities of the instrument scales were also investigated for the different modes of administration. The alpha coefficients for both modes of administration for both studies were acceptable, ranging between 0.60 and 0.91. In both studies the mean alphas obtained were higher for the Internet-based samples than for the paper-and-pencil samples. Ployhart *et al*. (2003) also report reliabilities that are higher for the Internet-based mode of administration than for the paper-and-pencil mode of administration. It is clear that administration mode does not have a negative impact on scale reliability. The SEM results obtained for the Internet-based samples were somewhat better than for the paper-and-pencil samples. This result is similar to those found by Bartram and Brown (2004).

Lastly, the effect of the unproctored Internet-based administration on the pattern of scale intercorrelations was examined using EQS. The comparison between the covariance structures of the Internet-based and paper-and-pencil samples for both studies produced an exceptionally good fit. Although the Chi-squares obtained for both studies were significant, the CFI and RMSEA indices indicated the equivalence of the two modes of administration. An explanation for the significant Chi-square is the fact that all the samples involved had approximately 1 000 respondents each. Byrne (2006) explains that using the Chi-square for determining invariance is unrealistic, as it is too sensitive for sample size and non-normality. There is an increasing inclination among researchers to base evidence for invariance on adequate model fit. It therefore appears that the results supported the equivalence of the covariance structures for the Internet-based and paper-and-pencil modes of administration.

Overall, the data obtained from the unproctored Internet-based mode of administration yielded comparable psychometric properties to the proctored paper-and-pencil administration in terms of mean scores and intercorrelations, as well as reliability and relationships between scales. This finding supports the findings of Bartram and Brown (2004) on the OPQ32i and implies that there is little to no distortion to the instrument itself.

Conclusion

The biggest limitation of the current study is the fact that two independent samples were used for the different modes of administration. To address this, the researchers attempted to equalise the two sample groups by controlling for biographical variables up to the point that no statistical significant differences were found between the groups based on these variables. There were, however, variables that could not be controlled for that might have influenced the results. It is recommended that a follow-up study be done in which the same group of candidates is used for both modes of administration.

The study addresses certain of the equivalence issues raised by the ITC in their Internet guidelines. Evidence is provided of comparable reliabilities, means and standard deviations between the different modes of administration. However, their recommendation that scores of both modes of administration be correlated with other tests and external criteria falls beyond the scope of this research and it is recommended for future research.

It should be noted that the present study focussed solely on structural equivalence and that full-scale equivalence was not investigated. Further research is recommended to address this issue.

In summary, this research demonstrated that the traditional proctored paper-and-pencil and the Internet-based versions of the OPQ32i are equivalent in terms of their correlations and means and that relationships between scales are not affected by mode of administration and supervision when used during real-life, high-stakes assessments.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the assistance provided by Anna Brown of SHL UK in conducting the EQS6.1 statistical analyses for this study. Due to limitations in space, the intercorrelation matrix of the OPQ scales is not published, but copies can be made available by the authors.

REFERENCES

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* (7th edn.). New York: Prentice Hall.

Bank, J., & Ramsey, M.A. (2001, April). *The impact of unproctored Internet administration of normative work styles questionnaires on response data.* Paper presented at the 16th Annual SIOP conference, San Diego, CA.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organisational Psychology*, *69*, 49–56.

Bartram, D. (1996). The relationship between ipsatized and normative versions of personality. *Journal of Occupational and Organisational Psychology*, *69*, 25–39.

Bartram, D. (2000). Internet recruitment and selection: Kissing frogs to find princes. *Internet Recruitment and Selection*, *8*, 261–274.

Bartram, D. (2001a, April). *Testing through the Internet: Mapping the issues for managing the future.* Paper presented at the 16th Annual SIOP conference, San Diego, CA.

Bartram, D. (2001b, June). *The impact of the Internet on testing for recruitment, selection and development.* Paper presented at the 4th Australian I/O Psychology Conference, Sydney, Australia.

Bartram, D., & Bayliss, R. (1984). Automated testing: Past, present and future. *Journal of Occupational Psychology*, *57*, 221–237.

Bartram, D., & Brown, A. (2002, June). *Mode of administration and the stability of the OPQ32i.* Paper presented at the ITC Conference on computer-based testing and the Internet, Winchester, England.

Bartram, D., & Brown, A. (2004). Online testing: Mode of administration and the stability of OPQ32i scores. *International Journal of Selection and Assessment*, *12*, 278–284.

Bartram, D. Brown, A., Fleck, S., Inceoglu, I., & Ward, K. (2006). *OPQ32 Technical Manual.* Thames Ditton: SHL Group.

Bartram, D., & Coyne, I. (2005). *ITC computer-based and Internet delivered testing guidelines.* Granada: International Test Commission.

Bentler, P.M. (2006). *EQS 6 structural equations program manual.* Ecino: Multivariate Software.

Brown, A., Bartram, D., Holtzhausen, G., Mylonas, G., & Carstairs, J. (2005, April). *Online personality and motivation testing: Is unsupervised administration an issue?* Paper presented at the 20th annual SIOP conference, Los Angeles, CA.

Buchanan, T., & Smith, J.L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, *90*, 125–145.

Byrne, B.M. (2006). *Structural equation modelling with AMOS: Basic concepts, applications and programming.* (2nd edn.). New Jersey: Lawrence Erlbaum.

Cappelli, P. (2001). Toolkit: Making the most of online recruiting. *Harvard Business Review*, *79*, 139–146.

Chapman, D.S., & Webster, J. (2003). The use of technologies in the recruiting, screening, and selection processes for job candidates. *International Journal of Selection and Assessment*, *11*, 113–120.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (2nd edn.). New York: Academic Press.

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

Cronbach, L.J. (1990). *Essentials of psychological testing.* (5th edn.). New York: HarperCollins.

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.

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

Donovan, M.A., Drasgow, F., & Probst, T.M. (2000). Does computerizing paper-and-pencil job attitude scales make a difference? New IRT analyses offer insight. *Journal of Applied Psychology*, *85*, 305–313.

Dowdeswell, K.E. (2006). *Applicant perceptions of the fairness of online assessment in the selection process.* Unpublished master's thesis, University of Pretoria, Pretoria, South Africa.

Drasgow, F. (2005, April). *Computerized testing and assessment: Boon or boondoggle?* Presidential address at the SIOP conference, Los Angeles, CA.

Evans, T. (2002). *Equivalence between Internet-based and paper-and-pencil cognitive ability tests and factors associated with the test modes effect.* Unpublished master's thesis, Birkbeck College, London, UK.

Foxcroft, C. (2009, January). *Advantages and disadvantages of computer-based and Internet based testing in South Africa.* Paper presented at the Knowledge Resources Seminar on Assessment: Selecting and developing talent, Bryanston, South Africa.

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.

Internet usage statistics – the big picture. (2005). Retrieved May 8, 2007, from http://www.internetworldstats.com/stats.htm

Jones, J.W., & Dages, K.D. (2003). Technology trends in staffing and assessment: A practice note. *International Journal of Selection and Assessment*, *11*, 247–252.

King, W.C., Jr., & 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.

Kriek, H.J. (2005, June). *International trends and Internet-based assessment.* Paper presented at the SIOPSA conference, Pretoria, South Africa.

Kriek, H.J. (2006, June). *Unproctored Internet testing: Local reliability and validity trends.* Paper presented at the SIOPSA conference, Pretoria, South Africa.

Lievens, F., & Harris, M.M. (2003). Research on Internet recruitment and testing: Current status and future directions. In I. Robertson & C. Cooper (Eds.), *The international review of industrial and organisational psychology.* Chichester: Wiley.

Marshall, L.A., & Lindley, P.A. (Eds.). (2007). Occupational Personality Questionnaire (OPQ32). In *Review of Personality Assessment Instruments (B) For Use in Occupational Settings.* Leicester: BPS.

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

Muchinsky, P.M. (2004). When the psychometrics of test development meets organizational realities: A conceptual framework for organizational change, examples, and recommendations. *Personnel Psychology*, *57*, 175–209.

Mylonas, G., & Carstairs, J. (2003). *Comparison of a computer administered motivation questionnaire under supervised and unsupervised conditions.* Unpublished master's thesis, Macquarie University, Sydney, Australia.

Pasveer, K.A., & Ellard, J.H. (1998). The making of a personality inventory: Help from the WWW. *Behavior Research Methods, Instruments and Computers, 30*, 309–313.

Ployhart, R.E., Weekley, J.A., Holtz, B.C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata and situational judgment tests comparable? *Personnel Psychology*, *3*, 733–752.

Potosky, D., & Bobko, P. (2004). Selection testing via the Internet: Practical considerations and exploratory empirical findings. *Personnel Psychology*, *57*, 1003–1035.

Richman, W.L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires and interviews. *Journal of Applied Psychology*, *84*, 754–775.

Robertson, I.T., & Kinder, A. (1993). Personality and job competencies: An examination of the criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology*, *65*, 225–244.

Rosenfeld, P., Booth-Kewley, S., & Edwards, J.E. (1993). Computer-administered surveys in organizational settings. *American Behavioral Scientist*, *36*, 485–511.

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*, 194–205.

Saville, P., Sik, G., Nyfield, G., Hackston, J., & MacIver, R. (1996). A demonstration of the validity of the Occupational Personality Questionnaire (OPQ) in the measurement of job competencies across time and in separate organisations. *Applied Psychology: An International Review*, *45*, 243–262.

Saville, P., & Wilson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal of Occupational Psychology*, *64*, 219–238.

Schrage, M. (1999). How the Web will test your sanity and competence. *Fortune*, *139*, 170.

SHL South Africa. (2004). *The Occupational Personality Questionnaire in South Africa: Psychometric research perspectives.* Pretoria: SHL Group.

SHL South Africa. (2006). *Global trends and the regulation of psychological tests*. Pretoria: SHL Group.

Skinner, H.A., & Pakula, A. (1986). Challenge of computers in psychological assessment. *Professional Psychology: Research and Practice*, *17*, 44–50.

Stanton, J.M. (1998). An empirical assessment of data collection using the Internet. *Personnel Psychology*, *51*, 709–725.

Templer, K. (2005, April). *Internet testing: Equivalence between proctored lab and unproctored field conditions.* Paper presented at the 20th Annual SIOP Conference, Los Angeles, California.

Tippins, N.T., Beaty, J., Drasgow, F. Gibson, W.M., Pearlman, K., Segall, D.O., & Shepherd, W. (2006). Unproctored Internet testing. *Personnel Psychology*, *59*, 189–225.

Trippe, D.M. (2005, April). *Equivalence of online and traditional forms of a Five Factor Model measure.* Paper presented at the 20th Annual SIOP Conference, Los Angeles, California.

Van de Vijver, F.J.R., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized version of the General Aptitude Test Battery. *Journal of Applied Psychology, 79*, 852–859.

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*, 461–474.

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

Weiner, J.A., Hayes, T.L., Reynolds, D.H., & Doverspike, D. (2005, April). *Unproctored Internet-based testing: Emerging issues and challenges.* Paper presented at the 20th Annual SIOP Conference, Los Angeles, California.