

https://dspace.org/

Raven's advanced progressive matrices within a South African context

Israel, Nicole

2006-11-14T12:26:28Z
http://hdl.handle.net/10539/1703

Downloaded from a DSpace repository. https://dspace.org/

RAVEN'S ADVANCED PROGRESSIVE MATRICES WITHIN A SOUTH AFRICAN CONTEXT

Nicole Israel (9807986H)

A Research Report submitted to the Faculty of Humanities, University of the Witwatersrand, Johannesburg, in partial fulfillment of the requirements for the degree of Master of Arts in Psychology by Coursework and Research Report.

Johannesburg, February 2006

Declaration

I hereby declare that this research report is my own independent work, and has not been presented for any other degree at any other academic institution, or published in any form.

It is submitted in partial fulfillment of the requirements for the degree of Masters of Arts in Psychology by Coursework and Research Report at the University of the Witwatersrand, Johannesburg.

_____ _____

Nicole Israel
(9807986H)

February 2006

Acknowledgements

I wish to extend sincere thanks and grateful acknowledgement to the following:

Dr Kate Cockcroft, supervisor and true mentor, whose unconditional support, academic guidance re the conceptualization and completion of the research, personal wisdom, commitment and exceptional patience underpinned every step of this project – the phrase 'thank you' is pitifully inadequate to express my gratitude for all you have done for me over the past few years, both academically and personally.

Mr Mike Greyling, not only for conducting the computer-based statistical analysis (being all too aware of my limitations with those 'infernal' machines) but also for his invaluable contribution of statistical theory, support, encouragement and occasional (necessary) lecture.

Mrs Gillian Haiden-Mooney, for her enthusiasm and encouragement in discussing the topic of academic writing and the research process generally, and her support and mentorship.

Mr Peter Fridjhon, Ms Sumaya Laher, Mr Michael Pitman, Prof. Charles Potter and Ms Adilia Silva, for both their endless patience in discussing statistical and psychometric theory and their personal support and encouragement.

The Cognitive Research Unit based at the University of the Witwatersrand, particularly Prof. Mervyn Skuy and Joseph Seabi, for their financial and academic support of the project.

Berl, Naeema, Grant, Mbongi, Saintha and Stephen, for their voluntary assistance in transporting the questionnaires and collecting the data; and Mr Michael Pitman and Ms Adilia Silva, for their voluntary assistance in presenting the research to potential subjects.

Dr Yvonne Broom, Lisa Church, Ms Andee Deverell, Prof. Norman Duncan, Prof. Gill Eagle, Prof. Gillian Finchilescu, Prof. James Fisher, Dina Grasko, Dr Kirston Greenop, Ms Jenny Haddingham, Ms Leonie Human, Dr Vinitha Jithoo, Prof. Heila Jordaan, Dr Lesley-Anne Katz, Ms Anisa Keshav, Ms Peace Kiguwa, Dr Karen Milner, Prof. Thokozile Mayekiso, Elena Mouyis, Dr Ali Peirson, Gregg Ravenscroft, Vera Schneider, Dr Andrew Thatcher, Mr Xolani Xaba, Ms Natalie Zahringer and numerous other members of staff and fellow students in the Discipline of Psychology and at WITS, for their various contributions of advice, support and encouragement during the process.

To all my family and friends, especially Debby, Kirston and Joy, for infinite patience and wisdom: *"True friends are those who believe in you even when you've ceased to believe in yourself"*.

To my parents, for everything…

Abstract

The issue of bias, whether a psychological test measures what it claims to measure similarly across different groups, remains a fundamental concern within the field of psychometrics, particularly within South Africa, where a history riddled with oppression, discrimination and malpractice in the area has led to suspicion, mistrust and legislation banning the use of many psychological tests as invalid and unfair (Foxcroft & Roodt, 2001; Murphy & Davidshofer, 2001; Nell, 1999). There is thus clearly a need for additional and more detailed investigations of the way specific individual tests function in the South African context. This study attempted to creatively examine systematic differences in performance on one specific test, the Raven's Advanced Progressive Matrices (RAPM), on the basis of home language and gender, factors seldom investigated in the literature.

A sample of one hundred Psychology first-year students completed a demographic questionnaire, the RAPM, the Similarities sub-test of the South African Wechsler Adult Intelligence Scales (SAWAIS) and an adapted version of the Reading Comprehension subtest of the Stanford Diagnostic Reading Test (SDRT). The data gathered was then utilized to explore four main research questions.

Firstly, in order to establish construct comparability, the relationship between the RAPM and a more verbally-oriented measure of *g,* the Similarities sub-test of the SAWAIS, was assessed. Results indicated a relatively strong positive relationship between the two measures ($r = 0.66$), and no significant differences between the correlations on the basis of either gender or home language.

Secondly, in order to explore the relationship between the RAPM and English comprehension, the study assessed the relationship between overall, literal and inferential scores on an adapted version of the Reading Comprehension sub-test of the SDRT and the RAPM. Results indicated only a moderate relationship between the two tests ($r = 0.65$), no difference in the relationship between RAPM performance and literal comprehension or inferential comprehension, and no difference in the relationship between the two tests on the basis of either gender or home language.

Thirdly, in order to establish whether items were found to be difficult in a similar way across the different gender and home language groups, p-values and regression lines were calculated. These indicated that significant differences in level of item difficulty were experienced between English and African language speakers, although no differences were apparent in item difficulty on the basis of gender.

Lastly, in order to establish whether qualitative differences in performance on the basis of ability (as estimated by performance on the RAPM), gender or language existed, a discrimination analysis examining the types of errors made by each group was performed. Repeated measures ANOVAs and multiple comparison post-hoc analyses revealed significant differences in the types of errors made on the basis of ability and home language, but not gender. The post-hoc analyses suggested that those of higher ability or first language English speakers were more likely to make incomplete correlate errors, while those of lower ability or speaking African first languages were more likely to make confluence of ideas errors. In general, the findings of the study seemed to suggest that the RAPM, while not biased on the basis of gender, might contain a deep-seated language bias despite their non-verbal presentation.

Table of Contents

Acknowledgements **3**

Abstract **4**

CHAPTER ONE: INTRODUCTION **9**

CHAPTER TWO: LITERATURE REVIEW **12**

2.1. Introduction **12**
2.2. Statistical Bias: Types and Sources **12**
2.3. Assessment and Testing in South Africa: A Brief Overview **17**
2.4. Raven's Progressive Matrices: Theoretical Background **20**
2.5. Raven's Advanced Progressive Matrices: Structural Description **25**
2.6. Raven's Advanced Progressive Matrices: Discussion of Bias **28**
2.7. Raven's Advanced Progressive Matrices: Discussion of Errors **32**
2.8. Raven's Advanced Progressive Matrices: Research in South Africa **35**
2.9. Rationale for the Current Study **38**

CHAPTER THREE: METHODS **40**

3.1. Sample **40**
3.2. Research Design **41**
3.3. Instruments **42**
3.4. Procedure **44**
3.5. Threats to Validity **45**
3.6. Data Analysis **46**

CHAPTER FOUR: RESULTS **50**

CHAPTER FIVE: DISCUSSION **69**

REFERENCE LIST **77**

APPENDICES **84**

APPENDIX A: Informed Consent Sheet
APPENDIX B: Demographic Questionnaire
APPENDIX C: Examples of RAPM Items and RAPM Scaling Studies
APPENDIX D: Categorisation of RAPM Distractors
APPENDIX E: Scatterplots for the Correlations between the RAPM, SIM and ARC
APPENDIX F: Comparative p-values for RAPM Items on the basis of Gender and Home Language
APPENDIX G: Regression Analyses
APPENDIX H: Analysis of Frequency of Errors by Gender
APPENDIX I: Post-hoc Analysis of Gender by Error Type Repeated-Measures ANOVA
APPENDIX J: Ethical Clearance Certificate for the Study

List of Tables

Table 1: Breakdown of Sample According to Demographic Factors **41**
Table 2: Reliability Estimates for the Tests **50**
Table 3: Chi-Squared Tests of Association between Variables in the Study **51**
Table 4: Basic Descriptive Statistics for the Tests **53**
Table 5: Kolmogorov-Smirnov Tests of Normality for the Tests **54**
Table 6: Pearson Correlation Coefficients: RAPM and Similarities sub-test **55**
Table 7: Pearson and Spearman Correlation Coefficients: RAPM, the Similarities sub-test and the adapted Reading Comprehension sub-test **56**
Table 8: Pearson and Spearman Correlation Coefficients on the basis of Gender and Home Language: RAPM and the adapted Reading Comprehension sub-test **57**
Table 9: Regression of p-values for Gender **59**
Table 10: Regression of p-values for Gender: Backwards Elimination Step One **59**
Table 11: Regression of p-values for Home Language **60**
Table 12: Frequency of Error Types Prior to Conversion to Proportions **61**
Table 13: Repeated Measures ANOVA: Error Type x Ability **62**
Table 14: Post-hoc Analysis: Error Type x Ability **62**
Table 15: Basic Descriptive Statistics: Error Type x Ability **63**
Table 16: Repeated Measures ANOVA: Error Type x Gender **65**
Table 17: Repeated Measures ANOVA: Error Type x Home Language **65**
Table 18: Post-hoc Analysis: Error Type x Home Language **66**
Table 19: Basic Descriptive Statistics: Error Type x Home Language **66**
Table 20: Breakdown of Sample: Ability x Home Language **68**
Table 21: Categorisation of RAPM Distractors **App.D**
Table 22: Comparative p-values for RAPM Items: Gender **App.F**
Table 23: Comparative p-values for RAPM Items: Home Language **App.F**
Table 24: Post-hoc Analysis: Error Type x Gender **App.I**
Table 25: Basic Descriptive Statistics: Error Type x Gender **App.I**

List of Figures

Figure 1: Histogram: RAPM **52**
Figure 2: Histogram: Similarities **52**
Figure 3: Histogram: adapted Reading Comprehension **53**
Figure 4: Post-hocA: RP and WP **63**
Figure 5: Post-hocA: WP and CI **63**
Figure 6: Post-hocA: RP and IC **64**
Figure 7: Post-hocA: RP and CI **64**
Figure 8: Post-hocA: IC and WP **64**
Figure 9: Post-hocA: IC and CI **64**
Figure 10: Post-hocL: WP and CI **67**
Figure 11: Post-hocL: RP and IC **67**
Figure 12: Post-hocL: RP and WP **67**
Figure 13: Post-hocL: RP and CI **67**
Figure 14: Post-hocL: IC and WP **67**
Figure 15: Post-hocL: IC and CI **68**
Figure 16: Studentized Residuals: Gender **App.G**
Figure 17: Coot's D Influence: Gender **App.G**
Figure 18: Regression Lines: Gender **App.G**
Figure 19: Studentized Residuals: Home Language **App.G**
Figure 20: Coot's D Influence: Home Language **App.G**
Figure 21: Regression Lines: Home Language **App.G**

Chapter 1: Introduction

Assessment, the process of "providing information to guide individuals, groups and organizations to understand and make informed and appropriate decisions about their functioning" (Foxcroft & Roodt, 2001, p.3), forms an integral part of professional psychological practice. From occupational and educational selection, placement and evaluation to clinical diagnosis and treatment, most sub-disciplines of psychology depend on gathering and collating data from numerous sources in order to address specific problems or questions (Kaplan & Saccuzzo, 2001; Murphy & Davidshofer, 2001; Owen & Chamberlain, 1996). Within this process, psychological tests and measures provide one key method of assembling this information, so much so that the terms 'testing' and 'assessment' are often, and erroneously, treated as synonymous (Murphy & Davidshofer, 2001). It is thus hardly surprising that psychometrics, the field of psychology "dealing with measurable factors" (Reber, 1985, p.594), has established itself as a recognized and vital specialization within the discipline.

One of the primary reasons for the popularity of testing is the notion that it provides an objective and balanced method for accurately placing individuals along a continuum of performance on various abilities or characteristics, thus allowing for relative comparison in an unprejudiced manner (Lippmann, 1976). In addition, the flexibility of tests in allowing a range of assessment formats and contexts (written, oral, individual, group and so forth) provides test users with the opportunity to accurately gauge performance on a wide variety of factors in a manner that best suits their specific needs (Foxcroft & Roodt, 2001, Murphy & Davidshofer, 2001). Tests – "objective and standardized measures of samples of behaviour" (Anastasi, 1976, p.23) – are thus popularly deemed as systematic and scientific, contributing to the objective and professional practice of judgement (Foxcroft & Roodt, 2001; Owen & Chamberlain, 1996). These ideas, while extremely appealing, are not always accurate, and provide justification for the importance of constantly re-assessing the utility and suitability of individual psychological tests, particularly within new contexts (Nell, 1999).

Generally speaking, there are three primary areas of concern dealt with by psychometrists or researchers in exploring this: firstly, whether the test is actually measuring an aspect of behaviour consistently; secondly, whether the test is measuring what it claims to measure; and lastly, whether the test measures what it claims to measure in the same way across different groups and people (Murphy & Davidshofer, 2001; Rosnow & Rosenthal, 1996). Of particular concern is establishing that the test, be it

an inventory, measure, scale or questionnaire, is able to measure an aspect of behaviour consistently (termed the test's reliability) and whether the test does in fact measure the aspect of behaviour it purports to measure (termed the test's validity) (Murphy & Davidshofer, 2001). There are numerous forms of both test reliability and validity. For example, a test's reliability can be established internally, by calculating the inter-item correlations, or over time, by examining the stability of tests scores, while validity can be examined in terms of test content, the match between test and task specification; criterion validity, the correlation between the test's prediction and an independent measure of the same behaviour; or construct validity, the correlation between the test score and other measures of similar/dissimilar ideas (Murphy & Davidhofer, 2001). In addition, it is necessary to establish whether the test makes any systematic errors in predicting performance (in either direction), in other words, to establish whether the test is biased. A biased test is clearly invalid, as it does not measure what it claims to accurately for all people (Murphy & Davidshofer, 2001).

Arguably, the majority of recognized and popular psychological tests and measures utilized in assessment today have been established as both reasonably reliable and reasonably valid across populations (Foxcroft & Roodt, 2001). Of particular concern to practitioners, however, is the mounting evidence that the majority of these tests are subject to some form of bias (Foxcroft & Roodt, 2001). This concern, in turn, has fuelled a major drive towards research in the area, and investigating the ability of a test to predict performance accurately and fairly 'across the board' has become a leading psychometric concern within the last few decades, particularly within South Africa, still suffering the ongoing effects of the historical segregation and discrimination that affected every aspect of life under the colonial and Apartheid governments, including psychometric testing (Foxcroft, 1997; Foxcroft & Roodt, 2001; Nell, 1999; Owen, 1996).

Against this background, the need for closer investigations of psychometric tests in order to establish their reliability, validity and particularly lack of bias within the multi-cultural, multi-lingual, democratic context of the 'New' South Africa is unquestionable (Foxcroft, 1997; Nell, 1997; Nzimande, 1995; Shuttleworth-Jordan, 1996). To this end, this research attempts to add to the growing database of psychometric knowledge available with regard to one specific test often utilized in the South African context, the Raven's Advanced Progressive Matrices (RAPM).

It is particularly useful to investigate the RAPM in the South African context for several reasons, not least of which is the test's wide-spread use in both applied and research psychology as an easy-to-administer, group or individual assessment of intellectual functioning (Carpenter, Just & Shell, 1990; DeShon, Chan & Weissbein, 1995). In addition, the test contains a relatively large number of items, making it particularly suitable for in-depth statistical analysis; and a large database of performance profiles and norms from numerous different populations and countries is available, enabling high levels of cross-cultural comparison (Carpenter et al., 1990). Due to their simplicity and non-verbal nature, the RAPM are also often touted as more culturally fair than other common measures of intellectual performance, an idea particularly appealing within South Africa, as it implies the test could potentially be suitable for use across cultural, linguistic and racial groups (Owen, 1992).

This study thus specifically attempted to examine whether the RAPM are biased in the South African context in terms of language and gender, factors not typically examined as sources of bias in this context. Furthermore, the research attempted to utilize innovative methods of establishing whether bias existed, including construct comparison through correlation, item difficulty analysis via regression and distractor analysis on the basis of the types of errors made across different groups.

Chapter 2: Literature Review

Introduction

In order to sensibly conduct an investigation of bias with regard to the RAPM in the South African context, it is first necessary to explain what precisely is meant by bias and why the issue remains so important to explore in this country today. In addition, it is necessary to provide a detailed description of the utility, underlying theory and structure of the RAPM test itself, as well as a detailed overview of previous research carried out involving the test.

Statistical Bias: Types and Sources

It is important to emphasise that test bias refers not to an opinion of whether the test is 'fair' or not, but to a mathematically based and statistically provable concept of systematic over- or under- prediction of performance across different groups on the test (Murphy & Davidshofer, 2001; Owen, 1996), or to quote Retief (1988) - " An item [or test] is biased if persons of equal ability [from different groups] do not have an equal opportunity for answering the item correctly" (p.45). Jensen (1976; 1980) refers to bias as systematic errors in the validity of the test scores of an individual on the basis of the group membership of that individual, and further suggests that bias could be evident both psychometrically and on closer examination of the individual performances of group members. Green (n.d., as cited in Green, Griffore & Simmons, 1976) expands on this by noting that bias could exist in terms of the test's content, norms or even the testing situation or context.

There are several common types of bias that are typically searched for in relation to the test content. The most basic of these, construct comparability, relates to whether the concepts or ideas being measured or examined translate in the same way across groups, in other words, whether the ideas expressed mean the same thing to different people (Owen, 1996). This form of bias relates closely to language and translation issues, but also extends beyond the merely pragmatic or literal to incorporate similarity of metaphorical or abstract notions, also termed conceptual equivalence (Retief, 1988). It is typically investigated by establishing whether similar internal consistencies, rank order of item difficulties, item discrimination values and factor structures exist between the groups (Jensen, 1980; Owen, 1996)

Once a broad similarity of meaning has been established, it is also possible that individual items within the test may be 'biased', in other words, that individuals from different groups may not have the same opportunity to answer any particular individual item correctly or may answer the item in a specific way. This form of bias, termed item bias or score comparability, essentially provides an overview of whether specific patterns of answers on the basis of group membership exist (Jensen, 1980; Owen, 1996). It is typically explored using analysis of variance, item response theory or distractor analysis (Owen, 1996)

In addition to establishing that both the general meaning and individual items of the test are comparable, a further concern is whether the test will predict performance on an independent criterion equally between the groups. This is estimated by formulating and comparing regression lines for the different groups, to establish whether the slopes and/or intercepts differ substantially from one another (Jensen, 1980; Owen, 1996). A 'bias' in this sense has particularly serious implications, as it means that the test will systematically overestimate or discriminate in relation to the performance of a particular group, and this error in measurement will be assumed to have been caused by an innate difference (Owen, 1996). Predictive bias is typically established by comparing the standard error estimate, the validity coefficient and particularly the slopes and cut-offs of the regression lines formulated for the two groups (Jensen, 1980; Owen, 1996).

Beyond the test content itself, bias could exist in terms of the way the scores are interpreted, the norms scores are compared to, the instructions or level of understanding of test demands and even the environment in which the test is taken (Owen, 1996; Raven, Raven & Court, 1998). These types of bias, however, are typically accounted for by the standardization process and common sense, and are thus not often formally investigated (Murphy & Davidshofer, 2001).

It is evident that bias is not a term that applies to the individual taking the test *per se*, but rather to the group to which that individual belongs. Theoretically, any difference or dividing factor between those people taking the test could be a source or cause of bias. In practical terms, however, the tendency is to focus primarily on culture and experience as those factors which are most likely to lead to substantial group differences (Owen, 1996).

Veroff & Goldberger (1995) define culture as:

> *… a system of shared meanings that emerges from cultural phenomena - such as people who share a common history; live in a specific geographic region; speak the same or a closely related language; observe common rituals, beliefs, values, rules and laws; and share culturally normative practices or patterns that constitute the fabric of how a society is set-up and functions (child-rearing practices; kinship patterns; power relations; ascribed roles…) – and that provides a common lens for perceiving and organising (structuring) reality (the world and the ideas one has about who one is and who others are) for and by its members* (p.1).

This definition is particularly useful as it highlights the myriad aspects that, having been deemed to provide a sufficiently 'common perceptive lens', could potentially constitute a 'culture' or grouping, and thus a source of bias. Language, age, gender, socio-economic status, literacy, location, religion – all could act as potentially rich sources of variability (and thus possibly bias) in terms of test question interpretation, stylistic approach and performance (Nell, 1999; Owen, 1996). Furthermore, the likelihood of these aspects causing systematic differences in experience or content for those taking the test is exacerbated by the fact that tests inevitably reflect the cognitive style, method of working and testing context typical within the society in which they were developed – in most cases Western, urban, middle-class assumptions, particularly within intelligence testing (Cockcroft, 2002; Neisser et al., 1996; Pellegrino, 1986). Lastly, it must be noted that even within particular 'cultures', there tend to be more sub-cultural differences than commonalities, and thus caution is indicated in terms of assuming comparability even within specified 'groups' (Green et al., 1976; Neisser et al., 1996).

Test scores reflect not only the test-taker's actual ability or characteristics in terms of what is being measured, but also their "level of emotional and social preparation" for taking the test, an area "saturated with cultural assumptions" (Bower, 2003, p.1.). Nell (1999), for example, argues that many of the test-taking skills inherently assumed to be present in test-takers by examiners, such as recognition of the need to work both speedily and accurately, understanding of the role of the examiner and even the level of communication between test-takers, are in fact directly influenced by the level of exposure the test-taker has had to testing within schooling and psychometric settings. It is not unreasonable to assume that factors such as level and quality of education, access to resources and level of functional literacy (ability to read and write) will thus impact dramatically on test scores in this respect, with those having less formalized exposure or experience at a distinct disadvantage. The

importance of ‘test-wiseness’, as this is often termed, was highlighted in a study carried out by Crawford-Nutt (1976), who was able to show that when concerted efforts were made to ensure that all test-takers understood the testing context and requirements to a similar degree, differences in scores between the ‘privileged’ and ‘disadvantaged’ groups were non-significant. Crawford-Nutt’s (1976) research, in turn, was based on several studies illustrating that methods of test presentation ensuring universal understanding of test demands produced substantially different results to those methods simply presented in a singular fashion to all test-takers regardless of culture (Irvine, 1962/1963; Pons, 1974; Schwarz, 1961 as cited in Crawford-Nutt, 1976; Serpell, 1979, as cited in Neisser et al., 1996).

The concept of ‘test-wiseness’ also provides a partial explanation for the strong effect that socio-economic status has been shown to have on test scores. Level of access to financial resources also tends to determine level of access to numerous other aspects of life, including diet, health-care and sanitary facilities, quality of education received, level of education achieved, exposure to books and technology, level of cultural stimulation and generally level of familiarity with Western culture and expectations, which in turn determine the level of performance that person is capable of cognitively and socially (Eysenck, 1986; Nell, 1999; Owen, 1996; Van den Bergh, 1996). Furthermore, many theorists have argued that ‘culture’ itself, the experiences intrinsically associated with a person’s ethnic and social background or ‘cultural attitude’, will fundamentally affect the approach the person adopts to the test (Kottak, 1994, as cited in Cockcroft, 2002; Melck, 2003; Neisser et al., 1996; Pellegrino, 1986). This can be demonstrated in terms of the person’s motivation to perform on the test, the ‘cognitive style’ that they adopt, assumptions with regard to what the most important factors within items or test sections are and even the test-taker’s level of confidence or attitude to performance (Bower, 2003; Owen, 1996). Bower (2003) presents evidence suggesting that even so-called ‘non-verbal’ problems, items designed specifically to capture ‘universal’ as opposed to culture-specific concepts in the form of shapes or numbers, are influenced by the test-taker’s mindset and cultural experience, for example, reading left to right, acceptance of traditional behavioural guidelines and so forth (Melck, 2003; Murphy & Davidshofer, 2001).

Language itself is often presented as an obvious source of bias. Oakland (1977, as cited in Owen, 1996) emphasizes that unless every test-taker understands clearly what is being asked of them and is comfortable responding not only on a literal level, but also in idiomatic and metaphorical terms, a bias against those not fluent is likely to occur. This could be in terms of non-equivalent or ‘untranslatable’ terms or concepts being presented in the test, misunderstanding of instructions or in the test itself

becoming a measure of language ability in addition to what it is supposed to be examining (Jensen, 1986; Nell, 1999; Owen, 1996; Van den Bergh, 1996).

Language has also been shown to fundamentally determine the way that people think and organize their world: "We dissect nature along lines laid down by our native languages…the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds – and this means largely by the linguistic systems in our minds" (Whorf, 1956, p213, as cited in Sternberg, 1996). The concept of linguistic relativity, also known as the Sapir-Whorf hypothesis, suggests that people will develop and utilize different cognitive styles and interpretations on the basis of the limitations imposed on them by the language they speak, in other words, language directly moulds and defines the type of knowledge people can gain about the world around them by determining their form of thinking about that world (Sternberg, 1996). This could be problematic in that people speaking a specific language may be able to only consider test items in light of the existing models, procedures and contexts permitted by their language, and may thus be tied to certain mental sets or stereotypes (Sternberg, 1996). Hunt and Banaji (1998, as cited in Sternberg, 1996) suggest that this phenomenon may occur even in the interpretation of visual imagery, in that both perception and memory can be influenced by the verbal tags attached to the object being encoded, which, in turn, may create connotations or understandings potentially unique to the language of the test-taker.

There is one further dividing factor that, although not a cultural group *per se*, provides an important potential source of systematic difference in test performance, namely gender. The extensive body of research conducted in this area suggests that on certain tasks large and consistent performance differences exist between the sexes (Jensen, 1980; Neisser et al., 1996). This includes a distinct advantage for males in terms of visuo-spatial tasks, performance tasks and mathematical and quantitative reasoning, whilst females tend to perform better on verbal tasks, particularly those related to verbal fluency and synonym generation (Gordon & Lee, 1986; Hines, 1990; Law, Pellegrino & Hunt, 1993; Linn & Peterson, 1985; Meehan, 1984; Stanley, Benbow, Brody, Dauber & Lupkowski, 1992, as cited in Neisser et al., 1996). Numerous reasons for this difference have been proposed, including biological-structural differences, socialization, level of confidence and inherent ways of acquiring and processing information (Neisser et al., 1996; Pallier, 2003; Ryan & David, 2003). It is worth noting, however, that most tests are specifically constructed to avoid overall differences in total scores on this regard (Neisser et al., 1996).

The list of potential sources of bias provided above is by no means complete, and goes some way towards emphasizing the critically important nature of investigations of test bias for the successful practice of psychometrics. Numerous international studies have indicated potentially problematic areas leading to erroneous placement, selection and assessment procedures, occasionally with disastrous consequences. The most well-documented of these problems probably remains the 'Spearman-Jensen effects', evident across the board in intelligence testing, which refer to the fact that, on average, Black Americans and Africans tend to score approximately one standard deviation below Whites on intelligence tests despite the items appearing similarly difficult and loading equally on *g* for both groups (Jensen, 1980; Lynn & Owen, 1994; Neisser et al., 1996; Owen, 1996; Rushton & Skuy, 2000). The debate as to whether this represents a genuine difference in ability or a substantial bias within the tests themselves remains both controversial and bitter, with acrimonious accusations of prejudice and scientific ineptitude (Neisser et al., 1996), however it does serve to highlight the need for further research in terms of test and concept comparability between groups not only in terms of race, but also in terms of other potential divisions that could substantially alter the decisions taken on the basis of those scores. Future research on this topic is particularly important within the South African context, given our country's unique and disturbing history (Foxcroft & Roodt, 2001).

Assessment and Testing in South Africa: A Brief Overview

While in many ways the development of the field of testing and assessment in South Africa paralleled the pattern of development found internationally, it is crucial to acknowledge the impact of the country's unique historical and social context. As Claassen (1995) states "…*testing and measurement…reflect some aspect of the world. It can be expected that tests will reflect the nature of the society in which they are used*" (p.8, as cited in Foxcroft, 1997, p.229). It is thus impossible to ignore the fundamental impact that South Africa's policies of racial segregation and discrimination had on every aspect of life, including psychological assessment and testing (Foxcroft, 1997; Foxcroft & Roodt, 2001; Nell, 1999; Shuttleworth-Jordan, 1996).

Appel (1989) provides a lucid discussion of the conflict created within White South African ideology at the turn of the twentieth century by the rise of capitalism, specifically the threat imposed by the Victorian ideal of "civilising the Africans by turning them into Black Europeans" (pp.545-546). He illustrates how White South Africans became increasingly pessimistic about the desirability of creating 'good, clean, obedient workers' (p.546) in light of increased competition for land and employment.

Appel (1989) further goes on to highlight how this, in turn, created scope for the imposition of legal barriers separating levels of access to resources between the races, and also propelled forward the movement towards 'scientific racism'. He goes so far as to argue that eugenics became "part of both the middle-class and academic common sense of White South Africans" (Dubow, 1987, as cited in Appel, 1989, p. 547).

Foxcroft and Roodt (2001) provide a clear example of how this consciousness impacted on testing in the work of M.L. Fick, a prominent South African psychologist working for the Department of the Interior. As early as 1929, Fick presented results indicating that the mental tests employed in his study showed that 'Native' children had I.Q.'s roughly comparable to those of "mental defectives", and observed that "…'the native child may have a different type of intelligence' than [sic] White children" (Fick, 1929, as cited in Appel, 1989, p.551). By 1939, Fick was attributing differences in performance between the groups to 'differences in original ability' or innate causes (Foxcroft & Roodt, 2001).

Foxcroft and Roodt (2001) also clearly outline the consequences that notions of racial difference had on testing prior to World War II. These included the standardization of mental tests primarily for White groups, the tendency to utilize measures standardized on certain groups with other groups without regard for the suitability of such, and in particular, the tendency to interpret test results as indicative of fundamental differences between the groups without considering the impact of environmental, educational and cultural factors on performance (Foxcroft & Roodt, 2001). Although work such as Fick's was criticized quite heavily as based on culturally inappropriate and biased measures (c.f. Biesheuvel, 1943, as cited in Foxcroft & Roodt, 2001), these issues remained within the field of testing, and were exacerbated by the introduction of the Apartheid system in 1948.

After World War II, especially in light of the government's policies of job reservation and access to education, the need to identify 'occupational suitability' fuelled a giant leap forward in the development of the field of testing in South Africa (Claassen, 1997). The lack of competition between the groups encouraged the development of tests along 'similar, but separate lines', and also the application of test norms developed specifically for international/White groups to other groups 'with caution' (Foxcroft, 1997). To quote Owen (1992), "there was little specific need for common tests because the various groups did not compete with each other" (p.112). Testing was also utilized relatively cynically, very often with the purpose of maintaining the distinctions between the racial

groups in terms of access to education and employment, and very few tests were developed for groups other than Whites (Foxcroft, 1997; Foxcroft & Roodt, 2001).

During the political shifts of the late nineteen seventies and early eighties, testing increasingly came under attack as biased and a tool for continued discrimination and segregation (Foxcroft, 1997). Fundamental questions regarding the utility of testing in South Africa began to gain prominence on the national agenda, and questions as to possible future directions were raised (Foxcroft & Roodt, 2001; Nzimande, 1995). This was epitomized in a study by Sehlapelo and Terre Blanche (1996), who collected the views of both workers and psychometric practitioners on testing and its future. Interestingly both groups felt that existing tests were biased and that a lack of trust in the process was justified, although the practitioners felt that technical solutions would solve the issues, while the workers called for more transparent testing policies and in many cases, the abandonment of testing in favour of less biased assessment procedures such as interviews. Foxcroft (1997) termed this point of view, that is, the idea that testing was simply too fundamentally flawed to be of any further use, the anti-test lobby (c.f. Nell, 2000, as cited in Rushton, Skuy & Fridjhon, 2003; Sehlapelo & Terre Blanche, 1996).

Plug (1996, as cited in Foxcroft, 1997) provided a coherent response to this, emphasizing primarily the lack of suitable alternatives to testing and the notion that despite its flaws, testing still remains more reliable and valid than its alternatives. He also gave emphasis to the fact that testing plays a crucial role in the field of assessment internationally, and that within a multi-cultural, multi-lingual society such as South Africa, the need for valid and fair assessment and placement is even greater than overseas (Plug, 1996, as cited in Foxcroft, 1997). Foxcroft (1997) utilizes Plug's (1996) point of view to emphasise the inevitability of continued testing in South Africa, but points out that the anti-test lobby is correct, and that there is an urgent need for both increased transparency in testing policy and disclosure of test results, as well as immediate investigations into the validity and potential bias of currently employed tests and measures in South Africa (c.f. Foxcroft & Roodt, 2001; Nell, 1999; Nzimande, 1995; Shuttleworth-Jordan, 1996).

Foxcroft and Roodt (2001) also describe how this conclusion has been acknowledged by various players since the advent of democracy. In many provinces, the government has banned the administration of group tests and school readiness assessments, and the ANC government has committed itself to a policy of redressing past imbalances, particularly illustrated in the new

Constitution, which specifically provides legal protection against all forms of discrimination (Foxcroft & Roodt, 2001; Nell, 1999). Furthermore, the Labour Relations Act of 1996, specifically states that:

> *Psychological testing and other similar forms or 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*
> (Employment Equity Act No. 55, section 8, as cited in Foxcroft & Roodt, 2001, p.28).

Foxcroft and Roodt (2001) further point out how these measures have huge implications for those involved in assessment in South Africa, as increasingly practitioners are required to demonstrate the validity, fairness and unbiased nature of their assessments (Nell, 1999). They emphasize that the response to this challenge needs to be met by those working within the field, particularly as currently the vast majority of tests are internationally developed, lacking entirely in local norms or with norms relevant only to certain sections of the population, and largely available only in English (Foxcroft & Roodt, 2001).

The above outline clearly demonstrates the need for deeper and more innovative research into the validity and suitability of specific psychological measures in the South African context. Although in the last decade the HPCSA (Health Professions Council of South Africa) has worked tirelessly to introduce new, more reliable measures and to provide more suitable norms and translations, the lack of funding and resources available makes the potential contribution of researchers working within academia vital. Not only do academically-based researchers have the opportunity to add crucial information to the growing databases aimed at ensuring the fair and ethical application of tests in assessment, but they also have the opportunity to drive the process forward by approaching the issue in different and novel ways. To this end, this study attempts to creatively examine the validity of a specific test, the Raven's Advanced Progressive Matrices (RAPM), within the South African context.

Raven's Progressive Matrices: Theoretical Background

Initially developed as an assessment tool for evaluating the mental ability of military recruits in the United Kingdom independent of their educational background, the Raven's Progressive Matrices

(RPM) tests epitomize one of the first and most successful attempts to present inductive reasoning and analogical tasks in non-verbal format (Kaplan & Saccuzzo, 2001; Paul, 1985). The non-verbal, or rather linguistically minimized, nature of the tests is deemed particularly important, as it allows an evaluation of intellectual ability without substantial influence by linguistic, educational and cultural factors, in other words, the tests are considered significantly culture-reduced, and thus a less biased and fairer measure across different populations (Kaplan & Saccuzzo, 2001; Owen, 1992; Paul, 1985; Raven et al., 1998; Spreen & Strauss, 1998; Valencia, 1984, as cited in Grieve & Viljoen, 2000). This assumption, namely that the tests are good measures of non-verbal intelligence not unduly influenced by culturally-specific factors, largely explains the extensive use and popularity of the tests throughout the world in both psychological practice and research (Arthur & Woehr, 1993; DeShon et al., 1995; Raven, 1989; Raven et al., 1998; Rushton & Skuy, 2000; Rushton et al., 2003).

The RPM tests were originally based on Spearman's (1904, 1923, 1927) research into intelligence and cognition, specifically his use of factor analytic methods to prove the existence of a 'positive manifold', a set of significant correlations between many different subtests of intelligence, suggesting the notion of a global or 'general' intellectual ability underlying performance (as cited in Cockcroft, 2002; Coleman, 2001; Eysenck, 1986; Lubinski, 2004; Neisser et al., 1996). Spearman termed this global ability *'g'*, and furthermore proposed that it consisted of two components, namely eductive ability, the ability to reason, and "reproductive ability, the ability to recall acquired information" (Raven, 1989, p.1; Raven et al., 1998). The RPM matrices were developed specifically to provide as direct a measure of eductive ability as possible, to act as "…a test of a person's capacity to form comparisons, reason by analogy, and develop a logical method of thinking, regardless of previously acquired information" (Raven, 1938, p.12, as cited in Paul, 1985, p.95).

Eduction, from the Latin *educare* (to draw out), refers to the ability to make meaning out of confusion, to discern relationships and perceive connections, and to infer new relationships from what is already known. It also refers to the ability to evolve higher-order mechanisms for processing and integrating information holistically, allowing for combined or individuated analysis within an active, largely non-verbal process of cognition (Raven, 2000; Raven et al., 1998). The RPM tests are designed specifically as tests of induction, and require the test-taker to identify and solve problems by inferring rules and relationships and generating answers on the basis of those (Alderton & Larsen, 1990; Carpenter et al., 1990; Raven et al., 1998). Carpenter et al. (1990) argue that integral parts of this solution process include matching corresponding elements, abstracting representations based only loosely on perceptual

input and successfully generating, tracking and managing problem-solving goals. They conclude that people vary in their ability to achieve this, with those scoring lower on the test tending to fixate on the answer through scanning and eliminating options rather than focusing on educing the complete pattern (Carpenter et al., 1990; Forbes, 1964).

In addition to eductive ability, there is varied and contradictory evidence as to other constructs that the RPM tests might measure (Babcock, 2002; Carpenter et al., 1990). Evidence suggests that the tests might function as measures of nonverbal intelligence (Bathurst & Kee, 1994; Giles, 1964; Jensen, 1983, as cited in DeShon et al., 1995); inductive ability (Rogers, Fitz & Hertzog, 1994, as cited in DeShon et al., 1995); fluid ability (Carroll, 1983; Cattell, 1971, as cited in DeShon et al., 1995); pattern perception and perceptual accuracy (Dillon, Pohlmann & Lohman, 1981; Owen, 1992) and working memory (Babcock, 2002; Carpenter et al., 1990; Kyllonen & Christal, 1990, as cited in DeShon et al., 1995). The possibility that the tests might act as a measure of spatial reasoning is particularly controversial, with some proposing that test performance does depend moderately on spatial ability (c.f. Ackerman & Kanfer, 1993; Hunt, 1974, as cited in DeShon et al., 1995; Babcock, 2002), and others categorically stating that it does not (Jensen, 1998).

Spreen and Strauss (1998) suggest that reasonable correlations (0.7-0.8) exist between the RPM tests and other measures of IQ, particularly given the verbally-oriented nature of most conventional IQ tests (Bors & Stokes, 1998; McLaurin, Jenkins, Farrar & Rumore, 1973; Mills & Ablard, 1993; Raven et al., 1998). There are also established moderate relationships between the RPM tests and measures of inspection time (Kranzler & Jensen, 1989; Nettlebeck, 1987, as cited in Bors & Stokes, 1998), processing speed and reaction time (Verguts & De Boeck, 2002) and tests involving goal generation and management (Kotovsky & Simon, 1973, as cited in Carpenter et al., 1990). The RPM tests, however, show relatively low correlations with measures of academic achievement, possibly suggesting that they measure intellectual potential as opposed to actual performance (Esquivel, 1984; Llabre, 1984, as cited in Spreen & Strauss, 1998; Mills & Ablard, 1993).

It is important to note that eductive ability is considered to be conceptually distinct from the notion of '*general intelligence' per se*, and that the RPM tests were never designed with the intention of functioning as measures of *g*, as opposed to eductive ability or fluid intelligence (Lubinski, 2004; Raven et al., 1998). Nevertheless, extensive research has shown conclusively that the RPM tests remain one of the best single measures of *g* available, illustrated typically through either scaling studies or

factor analytic research (DeShon et al., 1995; Jensen, 1998; Kaplan & Saccuzzo, 2001; Lynn, 2002; Paul, 1985; Raven, 2000; Raven et al., 1998).

Although a few factor analytic studies have indicated multiple factors on the RPM tests (c.f. Rimoldi (1948, as cited in Dillon et al., 1981); Carlson & Weidl (1979, as cited in Dillon et al., 1981) and, most notably, Dillon et al. (1981)), the vast majority have indicated a single latent variable responsible for determining performance (c.f. Alderton & Larsen, 1990; Arthur & Woehr, 1993; Bors & Stokes, 1998; DeShon et al., 1995; Matarazzo, 1990, as cited in Raven et al., 1998; Paul, 1985). While Arthur and Woehr (1993) argue that differences in sampling procedure might be responsible for the few discrepant results, others are quick to point out that the notion that a single factor might underlie performance on the RPM does not necessarily contradict the notion that it may include heterogenous problem-solving requirements or processes, for example, both visual and verbal reasoning (Alderton & Larsen, 1990; DeShon et al., 1995).

Possibly even more compelling is the evidence provided by several multidimensional scaling projects undertaken to prove that the RPM tests measure processes central to the construct of analytic intelligence. Studies undertaken by Ackerman and Kanfer (1993, as cited in DeShon et al., 1995); Marshalek, Lohman and Snow (1983, as cited in Kaplan & Saccuzzo, 2001) and Snow, Kyllonen and Marshalek (1984, as cited in Carpenter et al., 1990) clearly demonstrate that the RPM tests lie at the centre of a wide range of complex reasoning tests, with simpler tests ranging outwards. Those tests lying nearest the centre have strong correlations with both the RPM tests and each other, and tend to involve abstract reasoning or induction, whilst those on the periphery are less complex tests that tend to involve unidimensional tasks (Carpenter et al., 1990).

One possible reason for the considerable overlap between eductive ability and *g*, as represented by construct explication studies on the RPM tests, is the similarity of definition between the two. Whilst *g,* like intelligence, is often considered extremely difficult to define, it is, in essence, usually regarded as:

> [a] *very general mental capacity that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings – "catching on", "making sense" of things, or "figuring out" what to do…*
> (Gottfredson, 1997, p. 13, as cited in Lubinski, 2004, p.97)

While Raven et al. (1998) specifically warn against falling into the habit of ascribing to the concept of *g* considerably more explanatory power and relevance than it deserves, recent research has highlighted its importance in relation to numerous facets of life, including achieved socioeconomic status, delinquency, creativity, mate selection, health-risk behaviour and educational-vocational choice and performance (c.f. Deary, Leaper, Murray, Staff & Whalley, 2003; Deary, Whalley, Lemmon, Crawford & Starr, 2000; Gottfredson, 1997; Lubinski, 2000; Moffitt, Caspi, Harkness & Silva, 1993; Moffitt, Caspi, Silva & Stouthamer-Loeber, 1995; Plomin, DeFries, McClearn & McGuffin, 2001; Shea, Lubinski & Benbow, 2001, as cited in Lubinski, 2004; Jensen, 1986). Gottfredson (2004, as cited in Lubinski, 2004) argues that this is only to be expected within modern society, as educational, occupational and social tasks become less defined and more flexible. Increasingly, the demand is for people able to cope with change and novelty, anticipate developments, make inferences and 'stay ahead of the curve' (Holden, 2003). In addition, many modern theorists in the field of intelligence subscribe to a hierarchical model that places *g* at the apex of an increasing number of specific abilities, and regard *g* as central or core to almost all intellectual and cognitive abilities (Carroll, 1993, as cited in Lubinski, 2004; Cockcroft, 2002; Paul, 1985). Furthermore, Carpenter et al. (1990) point out that the specific model or construct subscribed to is, in fact, irrelevant, as the RPM tests have been shown conclusively to lie at the centre of intellectual and cognitive theory, and they thus argue that the constructs underlying performance on the RPM will account for performance on almost all other tests in those fields as well (DeShon et al., 1995).

From the above discussion, it is not difficult to see the importance of the RPM tests, as they represent a nearly flawless measure of a construct deemed fundamental to modern life (Anastasi, 1988; Jensen, 1972; Spearman & Wynn-Jones, 1951; Vernon, 1984, as cited in DeShon et al., 1995). Rushton et al. (2003) emphasise that the RPM tests remain the best available single measure of *g* even within samples of restricted scoring range, such as university students. Nevertheless, despite their theoretical suitability, it is vital that practical concerns, including those of potential bias, be explored prior to embracing the RPM tests within South Africa (Foxcroft & Roodt, 2001).

Raven's Advanced Progressive Matrices: Structural Description

There are three forms of the Raven's Progressive Matrices tests – of which the Raven's Standard Progressive Matrices (RSPM), developed as a measure of intellectual functioning across the ability spectrum, remain the most widely utilised (Raven, 2000; Spreen & Strauss, 1998). Despite their continuing use, however, it was quickly recognized that the RSPM were unable to provide sufficient discrimination of performance at extreme levels in both directions (Raven, 2000; Spreen & Strauss, 1998). In order to compensate for this and allow for the 'spreading out' of scores, the Raven's Colored [sic] Progressive Matrices (RCPM) and the Raven's Advanced Progressive Matrices (RAPM) were developed (Paul, 1985; Raven, 2000; Spreen & Strauss, 1998). The RCPM distinguish performance at the lower end of the spectrum, and are primarily utilized with children, the elderly or those mentally impaired (Spreen & Strauss, 1998). The RAPM, on the other hand, are intended for those designated as intellectually superior, and were originally designed to discriminate between those performing in the top ten percent on the RSPM (Mills & Ablard, 1993; Paul, 1985; Raven, 1989; Spreen & Strauss, 1998). In recent years, as the overall increase in performance on IQ tests (the Flynn Effect) has become more apparent, this has shifted slightly, with the RAPM routinely being used to screen the top twenty-five percent of performers on the RSPM (Bors & Stokes, 1998; Raven, 2000; Raven et al., 1998).

Similarly to the other Raven's tests, the RAPM consist of a series of geometric patterns or designs represented within three-by-three matrices or grids. Eight of the nine squares contain a completed pattern, the ninth, in the lower right corner, is blank, and must be filled by selecting one of eight possible options given below (Babcock, 2002; Paul, 1985; Raven, 2000). Test-takers are instructed to select the most logical option of the eight based on the rules governing changes in the pattern across the rows, columns and diagonals (Green & Kluever, 1992; Kaplan & Saccuzzo, 2001; Verguts & De Boeck, 2002). These changes commonly include variations in size, shape, number, direction, symmetry, proportion, shading and/or texture of the elements; and the correct alternative must fit the changes made both horizontally and vertically (Alderton & Larson, 1990; Bors & Stokes, 1998; Carpenter et al., 1990; Green & Kluever, 1992; Paul, 1985; Vodegel Matzen, van der Molen & Dudink, 1994). The incorrect alternatives (distractors) are constructed such that they represent a range of similarities to the correct answer, with some indicating similar but incomplete changes and others completely incorrect or unreasonable options (Carpenter et al., 1990; Thissen, 1976). The items themselves require no verbal response, and instructions regarding how to answer the test are both as minimal and as simplistic as possible, thus, at least theoretically, reducing the impact of language skill

on test performance (Kaplan & Saccuzzo, 2001; Zaidel et al., 1981, as cited in Spreen & Strauss, 1998). In order to maintain interest throughout the test and reduce the effects of fatigue, each item is presented in bold outline and drawn and labelled clearly (Melck, 2003; Rushton et al., 2003).

As is suggested by the results of analytic studies identifying several factors determining performance on the RAPM, it is possible to solve RAPM items in a variety of ways (Spreen & Strauss, 1998). Hunt (1974, as cited in DeShon et al., 1995), for example, identifies two potential problem-solving strategies that could be applied to resolving RAPM items: visual strategies dependent on processes such as continuation and superimposition; and analytic strategies dependent on reasoning and logic. The choice of strategy, while unlikely to make a huge difference in terms of overall score, does represent an important interaction between the key features of the problem and the personal profile of the respondent, and as such, could provide a unique source of insight into RAPM performance (DeShon et al., 1995; Mislev & Verhelst, 1990, as cited in DeShon et al., 1995). Carpenter et al. (1990) argue that these different strategies and their determinants can, in fact, be illustrated by a relatively small number of rules representing fundamentally different approaches to solving RAPM items (Jacobs & Vandeventer, 1972 as cited in Carpenter et al., 1990). The more complex the problem, the more rules required to solve it correctly, thus providing the basis for the argument that working memory plays a crucial role in successful RAPM performance, as it is necessary to correctly track and remember an increasing number of complex interrelationships between elements (Carpenter et al, 1990; Green & Kluever, 1992; Raven et al., 1998; Vodegel Matzin et al., 1994). Carpenter et al. (1990) also demonstrate that the rate of errors made is related to both the complexity of the item and the time taken to solve the item.

Maistriaux (1959, as cited in Raven et al., 1998) argues that failure to perform successfully on RAPM items most likely represents a failure or unwillingness to work analytically through deconstructing and then synthesizing information into a suitable gestalt. Raven et al. (1998) further suggest that successful pattern completion requires not only a recognition of the overall or whole, but also correct identification of the parts necessary to focus on or attend to and their correspondence. As an example, they cite the frequent choice of otherwise correct answers of the wrong size, and argue that this provides a clear demonstration of the importance of attention to detail and accurate perception and reasoning (Raven et al. 1998). Verguts and De Boeck (2002) expand on this hypothesis by emphasizing the importance of previous experience on the methods chosen to solve RPM items. In particular, they discuss the possibility of a learnt mental set developing while progressing through RAPM items, based

on the success with which rules or strategies could be implemented to solve initial items (Ferrara, Brown & Campione, 1986, as cited in Verguts & De Boeck, 2002; Verguts & De Boeck, 2002). This is supported by the forced choice nature of the test, which provides partial feedback to respondents as to the correctness of their thinking (Verguts & De Boeck, 2002). Thus, although not originally designed for the purpose, the RAPM can be utilized relatively successfully to analyse the cognitive determinants of successful performance (Carpenter et al., 1990).

Used without a time limit, the RAPM act as a powerful tool for assessing a person's general capacity for rational and logical thought, coherent perception and detailed observation, however the test can also be given under timed conditions to establish the test-taker's intellectual efficiency (Forbes, 1964; Paul, 1985; Spreen & Strauss, 1998). While it has been demonstrated that success on the RAPM does not depend heavily on the time aspect, Paul (1985) notes that untimed administrations of the test drastically reduce the number of items left unattempted at the end of the test, thus enabling far more accurate assessments of item difficulty uncontaminated by speed issues (Yates, 1961, as cited in Forbes, 1964). The RAPM are divided into two sets. Set I, designed to cover the full range of item difficulty presented in the RSPM, contains twelve items, and is widely used as a method of familiarizing people with the expectations and format of the test, although it can also provide a very rough assessment of mental ability (Bors & Stokes, 1998; Paul, 1985; Raven et al., 1998). Set II, consisting of thirty-six items, follows an identical presentation format, but contains increasingly complex and involved problems, thus providing a far more detailed and finely-grained analysis of ability (Babcock, 2002; Bors & Stokes, 1998; Paul, 1985; Raven et al., 1998). The correlation between the two sets, however, generally lies around 0.5, suggesting that performance on the practice items does not provide a reasonable estimate of performance on the test overall (Bors & Stokes, 1998).

The items presented in both sets, but particularly in Set II, are arranged in an ascending order of difficulty based on the frequency with which they are likely to be solved, calculated by assessing the error rates based on both failure to attempt and failure to solve (Bors & Stokes, 1998; Carpenter et al., 1990; Forbes, 1964; Paul, 1985; Raven et al., 1998). This absolute order, initially developed on the basis of research by Foulds, Raven and Forbes (1962, as cited in Paul, 1985), has proved remarkably durable across time and context, with very few items being assessed as 'out of place' in specific item difficulty analyses (c.f. Bors & Stokes, 1998; Paul, 1985; Raven et al., 1998). As a result of this, unlike the RSPM, the validity of the overall score for the RAPM does not depend on the test-taker having attempted every item (Rushton et al., 2003; Raven et al., 1998). Furthermore, as the items become

progressively more difficult throughout the test, earlier items can serve as learning tools for later items, with the test-taker becoming increasingly adept at the analytic strategizing and reasoning required by all but the very first items in Set I, which work more on the basis of visual algorithms (Hunt, 1974, as cited in Carpenter et al., 1990; Mills & Ablard, 1993; Rushton & Skuy, 2000). Despite this supposed learning effect, however, it is worth noting that research by Andrich and Dawes (n.d., as cited by Raven et al., 1998) has clearly indicated a lack of a 'leap and plateau' effect indicating sudden and dramatic increases in ability. Instead, on the basis of item response theory, they are able to demonstrate a steady and consistent relationship between the probabilities of solving progressive items (Raven, 2000).

Perhaps the most important effect of the remarkable similarity of order of item difficulty across populations and contexts, however, is the evidence it provides for a lack of bias within the test. While critics have attempted to argue that the test disadvantages certain groups on the basis of education, socio-economic status or culture, extensive research has indicated that the test exhibits a similar rank order of item difficulty across numerous ethnic and social groups (Raven, 2000; Raven et al., 1998; Rushton, Skuy & Bons, 2004). The fact that the items scale in the same way across disparate people suggests that the test is found similarly 'difficult' across different groups, thus refuting the notion that the demands of the test are in some way unfamiliar to particular types of test-taker (Raven et al., 1998; Rushton et al., 2003; Rushton et al., 2004). Furthermore, evidence has clearly indicated that the RAPM produces similar estimates of predictive validity within a variety of groups in relation to the same criteria (Raven, 2000; Raven et al., 1998). Raven (2000) argues that these findings, taken together, provide significant evidence that the test is not biased, as the items 'behave' in the same way across different groups (Rushton et al., 2003; Rushton et al., 2004).

Raven's Advanced Progressive Matrices: Discussion of Bias

Despite the compelling evidence provided by the 'absolute order' phenomenon, the debate as to whether the RAPM are, in some way, biased remains both complex and unresolved (Babcock, 2002; Owen, 1992; Raven et al, 1998).

For example, Raven et al. (1998) point out that there is a strong similarity between the performance norms, means and variances obtained on the RAPM in different countries at the same point in time, including Australia, Germany, Great Britain, Slovakia, and urban mainland China (Raven, 1989; Raven, 2000). They suggest that this stability both within and between societies with a literary tradition

provides evidence of the test's cross-cultural suitability (Raven, 1989; Raven, 2000; Raven et al, 1998). They acknowledge, however, that certain cultural groups, including Brazilians, Puerto Ricans, non-Whites in South Africa, and Native Americans and African Americans in the United States, routinely perform below these norms (Owen, 1992; Raven, 1989; Raven et al, 1998; Rushton & Skuy, 2000). They also suggest that there is consistent evidence of those from lower socio-economic backgrounds or less industrialized areas performing less well on the test, regardless of specific country, and that significant differences in performance between people speaking different home languages (French as opposed to Dutch) have been observed in Belgium, with the French performing far better (Raven et al., 1998). Raven (1989) counters this by arguing that the bulk of variance in scores still occurs between people of similar ethnic and socioeconomic background, however these facts would tend to suggest that there is at least a potential for cultural, linguistic or socioeconomic bias in performance on the RAPM.

There is little argument that scores on the RAPM increase with both increased level of education and increased socioeconomic status, and decrease with age, although the latter is understandable in terms of the construct the test is measuring (Babcock, 2002; Mills & Ablard, 1993; Owen, 1992; Raven, 2000; Spreen & Strauss, 1998). It has been suggested that the former may be due to differences in nutrition, welfare and hygiene between the groups, for example, Raven et al. (1998) note that differences in scores between socio-economic groups in the United States parallel differences in birth weight, infant mortality, quality of diet and early childhood illness (Benton & Roberts, 1988, as cited in Raven, 1989; Raven, 2000; Raven et al., 1998). It has also been proposed that fundamental differences in common child-rearing practices and access to educational resources (such as the media, books and television) may be responsible, with people from different occupational backgrounds espousing different values and priorities (Kohn, 1959, as cited Raven et al., 1998; Raven, 2000; Raven et al., 1998). This has been to some degree countered by evidence showing similar norms between societies with very different hierarchical structures, educational arrangements and intellectual quality in the home environment, as well as research demonstrating that there is a great deal of variation of values and priorities even within seemingly homogenous groups (Raven, 1989; Raven et al., 1998). On the basis of the contradictory and fragmented nature of the data available, as well as the small role these factors appear to play in total score variance, Raven et al. (1998) conclude that while differences in scores do exist, these cannot be easily accounted for by those variables typically implicated in societal division, and that further study is necessary (Raven, 2000).

The possibility that 'test-wiseness', level of sophistication, cultural preference or level of familiarity with testing context and content may play a role in determining performance on the RAPM has received some empirical support (Pons, 1974; Schwarz, 1961, as cited in Crawford-Nutt, 1976). Perhaps the most compelling is research carried out by Crawford-Nutt (1976), who was able to demonstrate that if South African Black students received sufficient training to ensure they fully understood the test requirements on the same level as their White counterparts, they performed equally well, suggesting that the typical Black-White differences in performance were more an artifact of the testing situation than a reflection of actual ability. It has been argued, however, that this exceptional performance may simply have been a result of utilizing a highly select and well-educated sample (Rushton & Skuy, 2000).

It is interesting to note that the similarity of norms raised by Raven et al. (1998) occurs primarily in groups sharing similar literary traditions (Raven, 1989; Raven, 2000). Much of the so-called 'cultural fairness' of the RAPM rests on the customary point of view that non-verbal content inherently leads to a reduction in linguistic and cultural bias, however this is not necessarily the case (Irvine, 1969, as cited in Lynn & Owen, 1994; Melck, 2003). Raven et al. (1998) argue that successful RAPM performance depends a great deal on the respondent's familiarity with traditional Western notions of literacy; shape, line and number; two-dimensional reasoning and perception; as well as the importance placed on dealing with and persisting with abstractions. Some credence has also been given to the notion that different groups may inculcate deeply imbedded, culturally-specific forms of thinking inherently different to those prevalent in traditional Western societies, for example, Sowell (1994, as cited in Rushton & Skuy, 2000) suggests that throughout the world Black cultures show a marked preference for spontaneity and improvisation as opposed to abstract thinking (Rushton & Skuy, 2000).

Critics note that a fundamental assumption of the RAPM is that those taking the test will share a common cognitive and literary basis, and argue that the established differences in scores between societies with a literate tradition and those without might belie the claim of cultural fairness to some degree (Raven et al., 1998). Olsen (1986, as cited in Rushton & Skuy, 2000) supports this point of view, emphasizing that non-verbal reasoning still follows the same analytical process as verbal logic, and that the ability to perceive rests on experience and cultural learning. There are thus those theorists who suggest that the RAPM are not valid except when utilized with those sharing a similar cultural and literary background (c.f. Jensen, 1980; Kamin, 1995; Nell, 2000; Wigdor & Garner, 1982, as cited in

Melck, 2003; Rushton & Skuy, 2000). Raven et al. (1998), however, counter-argue that even within relatively homogenous literate societies there is considerable variation in performance.

Given the widely accepted link between language, culture and thought, the dearth of research regarding the role linguistic or cultural patterns may play in determining the thought processes underlying performance on the RAPM is both peculiar and disturbing (Sternberg, 1996). In particular, almost no research exists with regard to the effect that 'mother-tongue' may have on RAPM performance, despite evidence suggesting that language has significant effects on measurements of intellectual functioning and evidence of consistent differences in performance on the basis of certain home languages (Raven et al., 1998; Spreen & Strauss, 1998; Strauss, 2003). The notion that the language a person speaks is likely to fundamentally determine their style and method of answering the RAPM and dealing with the material is not at all far-fetched, particularly in light of the Sapir-Whorf hypothesis, and the need for more research in this specific area is self-explanatory (Sternberg, 1996).

Almost equally controversial, although far better researched, is the potential role gender may play in determining RAPM performance. Numerous studies have suggested that there are no significant sex differences in RAPM performance on the basis of either raw scores or scaled abilities (c.f. Court & Kennedy, 1976, as cited in Paul, 1985; Jensen, 1998; Mackintosh, 1996, as cited in Lynn, 2002; Rushton et al., 2003; Sperrazzo & Wilkins, 1958; Tulkin & Newbrough, 1968, as cited in Thissen, 1976; Thissen, 1976). In fact, Court (1983), on the basis of an extensive review of well over a hundred studies, concluded that "accumulated evidence at all ability levels indicates that a biological sex difference cannot be demonstrated for performance on the RPM" (p.68, as cited in Lynn, 2002, p.670). Furthermore, research by Jensen (1998) strongly suggests that there are no significant sex differences in psychometric *g*; that those sex differences that do exist are inconsequential and vary in direction; and that observed sex differences are usually attributable to differences in other, more specialized abilities, such as visuo-spatial skill or verbal ability.

There is, however, a persistent tendency for a small, usually non-significant male advantage to appear in analyses of test scores, which occasionally has even been found to be significant (Arthur & Woehr, 1993; Deltour, 1993; Heron & Chown, 1967, Wilson, de Fries, McClearn, Vandenberg, Johnson & Rashad, 1975, as cited in Lynn, 2002; Grieve & Viljoen, 2000; Lynn, 2002; Paul, 1985). Paul (1985) proposes that this may be due simply to sampling error, as do Arthur and Woehr (1993). Lynn (1992), however, argues that many previous studies have simply utilized samples that were too small, and

further, that sex differences in performance can only be considered within a framework recognizing the differential development of the genders. He also suggests that gender differences may be less apparent, particularly within smaller samples, due to environmental factors, specifically male and female societal roles and socialization processes becoming less differentiated (Lynn, 2002). By utilizing a larger sample than previous studies, Lynn (2002) was able to demonstrate a significant gender difference in performance on the RAPM. Thus, the evidence with regard to the specific role gender may play in determining RAPM performance is contradictory, and, as with language, requires further research.

Raven's Advanced Progressive Matrices: Discussion of Errors

While it is clear that a great deal of research has been carried out with regard to potential bias on the RAPM, it is worth noting that almost all these studies have utilized the total number of items correct as their variable of interest and few, if any, have examined individual or cognitive factors underlying performance in greater depth (Babcock, 2002; DeShon et al., 1995; Verguts & De Boeck, 2002). Simply establishing the proportion of correct answers on each item (item difficulty analysis), or even correlating test total scores and performance on each item (item discrimination analysis), does not provide information regarding the qualitative performance of those taking the test, particularly with regard to the logic informing choices made between the eight available options on each question (Babcock, 2002; Forbes, 1964; Vodegel Matzin et al., 1994). Although not originally designed for this purpose, several theorists have argued that a closer examination of the incorrect responses given by test-takers may yield valuable additional information as to why a person is failing to respond correctly (Babcock, 2002; Grieve & Viljoen, 2000; Hunt, 1983; Spreen & Strauss, 1998; Thissen, 1976; Vodegel Matzin et al., 1994). In fact, Raven et al. (1998) argue that systematic investigation of the solution strategies utilized by different people may provide unique insight into their cognitive processes and styles not obtainable in any other fashion.

The RAPM were originally designed so that the distractors (the seven options representing incorrect pattern completions on each item) reflected various degrees of similarity to the correct answer (the eighth option), and as such, were not equally likely to be chosen (Raven et al., 1998; Thissen, 1976). This necessitated establishing the probability of the correct alternative being chosen purely by chance, as well as the frequency with which each distractor was chosen (Forbes, 1964). Raven (1960, as cited in Forbes, 1964) expanded on this, proposing that people with a lower capacity for coherent, logical thinking would be more likely to choose distractors reflecting material already provided or arbitrary

reasoning, whereas those with a higher capacity would be more likely to choose options reflecting reasoning correct up to a point or extended too far. On this basis, Forbes (1964) proposed four specific types of errors, each representing a different failure in logical deduction.

Wrong Principle Errors (WP) are said to occur when a person selects an option that reflects a line of reasoning either fundamentally different to one leading to the correct solution of the problem, or apparently random or arbitrary, whereas ***Repetition Errors (RP)*** occur when a person simply selects a distractor representing one of the three figures adjacent to the missing piece in the matrix. Both of these error types imply a complete breakdown of the logical process required to educe the appropriate relationship or pattern (Babcock, 2002; Forbes, 1964; Raven et al., 1998; Vodegel Matzin et al., 1994). They could potentially stem from a previously learnt mental set or stereotype (Sternberg, 1996).

In contrast, both incomplete correlate errors and confluence of ideas errors tend to indicate a certain level of progress towards the correct solution, albeit incomplete or inaccurate. ***Confluence of Ideas Errors (CI)***, also termed over-determined choice errors, occur when a person is unable to distinguish relevant elements or characteristics within the pattern presented, and thus chooses the most elaborate, complex option available (usually one that combines most or all of the elements in the matrix) - this is thought to represent over-inclusive thinking or a lack of ability to individuate, as although the answer contains correct elements, it also contains additional unnecessary information. ***Incomplete Correlate Errors (IC)*** are said to occur when the option selected is only partially correct, containing some but not all of the necessary elements – this is the error 'closest' to the solution of the item, representing a thought pattern essentially correct but not carried far enough or unfinished (Babcock, 2002; Forbes, 1964; Raven et al., 1998; Vodegel Matzin et al., 1994).

Having categorized the different error types, Forbes (1964) further attempted to establish whether differences existed between ability groups in terms of the types of errors made. He categorized the two most common errors on each item, and then conducted simple frequency analyses according to overall RAPM scores. His findings concurred with Raven's (1960) suggestions of a link between ability and error type, indicating that those with low ability tended to make relatively more wrong principle, repetition and confluence of ideas errors, whereas those with high ability tended toward incomplete correlate errors (as cited in Forbes, 1964). However, although innovative, Forbes' research was also problematic, as it was based only upon analysis of the two most common errors for each question, resulting in possible skewing. This was compounded by the fact that the items, not initially designed

for distractor analysis, did not contain equal amounts of each type of error across the items (Babcock, 2002).

Vodegel Matzin et al. (1994) argue that the uneven distribution of error types between the questions, creating substantial difficulties in terms of probability and comparison, has proved one formidable obstacle to sustaining interest or research in the area of distractor analysis. A second, equally troublesome issue, is the fact that to date, no-one has actually classified all the different distractors on the RAPM in terms of the error type they represent, as most studies have simply focused on the two most common errors previously identified (Babcock, 2002; Raven et al., 1998; Vodegel Matzin et al., 1994). As a result, both the issue of distractors possibly representing more than one error type, and the possibility of disagreement between theorists as to which error type a specific distractor represents, remain potential problems, and the results of error analyses must therefore be treated with considerable caution (Vodegel Matzin et al., 1994).

These complications may partially explain the sporadic and haphazard nature of research in the area subsequent to Forbes' (1964) study (Babcock, 2002). In 1963, Sigel (as cited in Thissen, 1976), again utilizing the two most common errors on each item and thus falling prey to the same problems, attempted to establish whether ability, gender and/or age played any role in determining the type of distractor chosen, but found no significant relationships. Little further research in the area was conducted until that of Jacobs and Vandeventer (1970, as cited in Thissen, 1976) and Thissen (1976), who utilized innovative statistical techniques to conclude that additional information was available in incorrect responses and could potentially be used to improve estimates of ability, although they did not examine the specific types of errors made by respondents. Other studies, such as that by Paul (1985), failed to illustrate any significant patterns in the types of distractors chosen.

It was only in research by Vodegel Matzin et al. (1994) that the first systematic attempts to tackle these issues were made. Vodegel Matzin et al. (1994) were particularly interested in establishing whether the errors made by children on the RSPM followed a similar pattern to those made on the RAPM by adults. They utilized both the original RSPM and a set of experimental matrices devised by themselves to assess this, with the latter being designed specifically to represent equal probabilities between the error types. They were able to conclude that, even given the problems with previous research, no qualitative differences in performance existed between children and adults, thus to some degree confirming the findings of earlier studies (Babcock, 2002; Vodegel Matzin et al., 1994)

The most coherent study to date, however, remains the one carried out by Babcock (2002), in which she attempted to verify whether qualitative differences in performance existed on the basis of ability and age by analyzing both the types of errors made and the rules used to educe solutions. Not only did Babcock's (2002) study match previous findings with regard to the link between error type and ability, confirming that those with different levels of ability made different types of errors, but she was also able to show no difference in the types of errors made on the basis of age, confirming research by Vodegel Matzin et al. (1994). Babcock's (2002) findings suggest that those with high ability are most likely to make IC errors, followed by CI, RP and WP errors. Furthermore, those with high ability tend to make more IC and less WP errors than those of medium ability, who perform the same in relation to those of low ability. In addition to her actual findings, Babcock's (2002) research is hugely important as it outlines and utilizes one of the few coherent techniques available for dealing with the uneven distribution of error types across the items, by converting the data to proportions. She is also the first researcher to provide a categorization of every distractor of the first thirty-one items of the test, having achieved reasonably high levels of inter-rater agreement as to the single type of error they best represent (Babcock, 2002).

It is difficult to believe that, despite the obvious utility of such data in yielding insight into the logical processes underlying performance, no research examining the possibility of patterns of errors on the basis of factors such as culture, language, gender or race appears to exist (Babcock, 2002). The potential of such analyses for yielding heretofore unexplored explanations of performance is unquestionable, and demands further exploration.

Raven's Advanced Progressive Matrices: Research in South Africa

Although there is a fair amount of research regarding RPM functioning in the South African context, the vast majority deals with the RSPM, as opposed to the RAPM, and focuses exclusively on race as a potential source of bias (c.f. Grieve & Viljoen, 2000; Owen, 1992; Rushton & Skuy, 2000; Rushton et al., 2004). Only a few studies have examined gender as a possible source of bias, and none appear to have examined whether systematic differences in performance occur on the basis of home language (Rushton & Skuy, 2000; Rushton et al., 2003). The closest is probably the research carried out by Crawford-Nutt (1976), who was able to demonstrate that, with suitable training, a highly select group of Black students was able to perform equivalently to a similar White sample, thus suggesting that

'test-wiseness' and ability to understand test requirements might play a crucial role in determining the usually poor performance of non-White groups.

The well-established phenomenon of poor performance by non-Whites on measures of intelligence and cognition is equally prevalent on the RPM as on other less supposedly 'culture-fair' measures of intellectual ability within South Africa. Routinely, non-White groups perform on average one to two standard deviations below their White counterparts (Lynn & Vanhanen, 2002, as cited in Melck, 2003; Raven et al., 1998; Rushton & Skuy, 2000; Rushton et al., 2003; Rushton et al., 2004; Skuy, Schutte, Fridjhon & O' Carroll, 2001). There is some debate as to whether this truly represents a bias, however, or whether it is simply a case of years of systematic deprivation and oppression showing through (Raven, 2000). This is illustrated in research carried out by Poortinga (1971, as cited in Rushton & Skuy, 2000), who, having utilized the RAPM with a sample of university students, found that on average African students performed over two standard deviations below their White counterparts, although they did perform substantially above other non-university-based African samples as well. Poortinga (1971, as cited in Rushton & Skuy, 2000) thus stated that the RAPM were too difficult for African students, and it can be inferred that this was assumed to be the result of a lack of access to resources (Freeman, 1984, as cited in Owen, 1992; Rushton & Skuy, 2000). Similar results were found by Grieve and Viljoen (2000) and Zaaiman, van der Flier and Thijs (2001, as cited in Melck, 2003), although by this stage research indicating similarities in item structures, item loadings on *g* and the way items were found to be difficult across the groups had begun to overrule those espousing the cultural-deficit hypothesis (Lynn & Owen, 1994; Nell, 2000; Owen, 1992; Rushton & Skuy, 2000).

Owen (1992), utilizing the SPM, was able to conclusively establish that the test did measure the same construct across four different racial groups, despite the usual differences in total score, with an average difference in converted IQ of over forty points between African and White groups. It is interesting to note one of his conclusions, however:

> *Perhaps a more viable option for South African psychologists in the long run is to be found in the suggestion made by Burg and Belmont (1990) that culturally different children use different solution strategies. Thus, although all the testees in the present study generally reacted to the SPM in a similar way, the possibility still exists that not all testees processed the items in functionally equivalent ways. The possibility that black and white school children differ in this regard seems a worthwhile topic to pursue in the future.*
> (Owen, 1992, p.158)

Owen's (1992) findings of substantial differences in overall performance on the RSPM were replicated by several recent studies, including those carried out by Rushton and Skuy (2000) and Rushton, Skuy and Fridjhon (2002, as cited in Melck, 2003). These studies, conducted under the auspices of the Cognitive Research Unit at the University of the Witwatersrand, and utilizing university students as their sample, were able to establish a similar order of item difficulty between the races and thus concluded that no bias existed on the RSPM and that the test was equally valid for the different groups, although the test itself was too easy for a university sample (Rushton & Skuy, 2000; Rushton et al., 2002, as cited in Melck, 2003).

On this basis, two further studies examining the validity between racial groups was carried out, this time utilizing the RAPM (Rushton et al., 2003; Rushton et al., 2004). Rushton et al. (2003) were again able to confirm that the test measured the same construct, namely *g*, across the different groups, that the items followed a similar order of item difficulty between the groups, and that the test showed similar predictive validity across the groups as well. They also found no gender-related differences in performance (Rushton et al., 2003). Rushton et al. (2004) assessed whether the RAPM possessed the same construct validity across racial groups, utilizing the RAPM, the Similarities sub-scale of the South African Wechsler Adult Intelligence Scale, an English Comprehension test, end-of-year university grade and high-school grade point average. They determined that the same pattern of group differences was found across all of these, and that the items in the RAPM behaved in the same way across the groups despite differences in average score (similarity of slope) (Rushton et al., 2004). They concluded that the RAPM were as valid for African as non-African groups (Rushton et al., 2004). Additional research attempting to determine the RAPM's predictive power in terms of academic performance showed that the RAPM did not function successfully as a predictor within the South African context (Melck, 2003; Skuy, 2003). The singular nature of this finding again suggested that the test functioned similarly between different racial groups (Kendell, Verster & Von Mollendorf, 1988, as cited in Rushton et al., 2003).

Rushton et al. (2003) and Rushton et al. (2004) argue that these results provide conclusive evidence that the differences in total scores are not a result of cultural factors, but represent a true and significant difference in the construct measured by the groups, namely *g*. It is important to note, however, that there are theorists who argue that some research has yielded significant differences in item difficulty and item-by-group interactions on the basis of factors other than race (Irvine, 1969, as cited in Owen,

1992; Owen, 1992). The possibility that systematic differences in performance exist on the basis of a different, not entirely distinct factor, cannot be overlooked, particularly in the South African context, where race, language, culture, socio-economic status and numerous related factors are so inextricably entwined (Foxcroft, 1997; Nell, 1999). Minimal research has been carried out with regard to factors other than race as a potential source of bias. Furthermore, as is suggested by Owen (1992), traditional methods of establishing whether bias exists may not be sufficient. Instead, it may be necessary to examine group differences in performance at a far more in-depth and cognitively-based level, in order to establish whether the RAPM function as a fair and unbiased measure in the South African context.

Rationale for the Current Study

The discussion above illustrates the necessity of further research in the field of bias in psychometric measures in South Africa, especially with regard to tests of high utility and wide-spread relevance such as the RPM. Not only does research illustrate the central nature of the RPM theoretically as a measure of intelligence or cognitive ability, but the test's non-verbal format has meant that it has been widely applied in the South African context under the assumption of being more 'culturally fair' or less biased than other, more verbal measures (Carpenter et al., 1990; DeShon et al., 1995). Research attempting to establish whether the RPM are in fact biased in South Africa has generally indicating that they are not (Owen, 1992; Rushton et al., 2003). It is worth noting, however, that these studies have tended to utilize highly traditional methods of establishing bias, such as item difficulty and item discrimination. Furthermore, albeit understandably given South Africa's socio-political and historical context and the Apartheid regime's direct equation of race and culture, research in the area has tended to examine test performance and bias almost solely along racial lines (Foxcroft & Roodt, 2001; Nell, 1997; Shuttleworth-Jordan, 1996; Nzimande, 1995). While not negating the importance of this issue, the need for research examining bias in terms of other potential causes is becoming increasingly clear. While almost any factor remains important to investigate, the lack of research regarding the role of language and the contradictory nature of the evidence regarding gender highlight these two factors as particularly important for further investigation (Jensen, 1998; Raven et al., 1998). The current study thus attempted to creatively examine whether systematic differences in RAPM performance (bias) could be determined in a South African context on the basis of either gender or home language. In addition, the study attempted to utilize innovative methods of examining issues of construct comparability and score comparability (item bias), particularly focusing on an analysis of the types of errors made by respondents as a potential source of additional data with regard to group performance.

More specifically, the study aimed to examine the following:

Firstly, in order to establish construct comparability, the study examined the relationship between the RAPM, as a primarily non-verbal measure of *g*, and the Similarities sub-test of the SAWAIS, a well-established verbally-based measure of *g* (Lezak, 1995). Additional analyses to establish whether this relationship differed on the basis of gender or home language were also carried out.

Secondly, relating to exploring the connection between linguistic skill and test performance, the study examined whether there was a relationship between language comprehension in English, as measured by an adapted version of the Reading Comprehension sub-test of the Stanford Diagnostic Reading Test (SDRT) measuring both literal and inferential comprehension, and performance on the RAPM. Again, additional analyses to establish whether this relationship differed on the basis of gender or home language were also carried out.

Thirdly, in relation to score comparability, analyses of item difficulty on the basis of gender and home language were carried out, specifically to determine whether the 'absolute order' phenomenon held true for these groups as well, and whether they found the items 'equally difficult'.

Lastly, in order to establish whether performance differed qualitatively between groups of different ability level, gender or home language, the study examined the frequency of each type of error made between these groups on individual items. Repeated-measures ANOVAs and post-hoc multiple comparisons were utilized to establish whether significant differences in the patterns of distractors chosen on the basis of these groupings existed.

Chapter 3: Methods

Sample

The sample was a non-probability, convenience sample, and consisted of volunteers from among those students registered for the Introduction to Psychology first-year module at the time the research was conducted. The choice of university students studying Psychology as a target population was motivated by several factors. Firstly, it could be assumed that those students studying at a tertiary level, having passed through the secondary education system and written the standard Matric exam, possessed a certain level of 'test-wiseness' - familiarity with testing materials, routine, instructions and procedures - which rules this out as a likely extraneous variable (Grieve & Viljoen, 2000). The importance of this in terms of the potential impact on RAPM performance is demonstrated in research by Crawford-Nutt (1976). Secondly, Rushton and Skuy (2000) suggest that students from the University of the Witwatersrand are likely to score at least one standard deviation above the average mean performance in the South African population as a whole, thus emphasizing the suitability of the RAPM, which were intended for use in populations of above average ability (Raven et al., 1998; Rushton et al., 2003). Lastly, Psychology is the only subject at the University of the Witwatersrand offered in four of the five faculties, and is studied by a wide range of students in terms of gender, age, race, religion, linguistic background, current curriculum, previous curriculum studied, and career intention. A sample selected from this group thus optimized the probability of obtaining a highly diverse and therefore more representative sample.

All respondents participated on a voluntary basis, and written informed consent was obtained following a verbal explanation of the purpose and requirements of the study (refer to Appendix A). Although one-hundred and thirty-seven people initially volunteered, only one hundred of these were utilized in the final study. Of the thirty-seven responses discarded prior to analysis, one completed only the demographic questionnaire, five omitted key demographic information, eight omitted at least ten items on the RAPM, two did not attempt the Similarities sub-test of the SAWAIS, fourteen omitted more than one third of the Reading Comprehension sub-test of the SDRT, and seven omitted more than eight items from both the RAPM and one of the other tests. As the final attained sample size was only one hundred, it became necessary to dichotimise certain of the demographic variables in order to allow sensible statistical analyses. Variables such as home language, race and socio-economic status (as

estimated by parental level of education and parental occupation utilizing the Hall-Jones Scale of Occupational Prestige (Oppenheim, 1966)) were thus divided into only two categories, resulting in a loss of detail (Howell, 1997).

The composition of the final sample of one hundred utilized in the analysis was therefore as follows:

Table 1: Breakdown of Sample According to Demographic Factors

Demographic Factor	**Dichotomous Division (utilized for analysis)**	**Original Responses (listed in order of magnitude, some single responses omitted)**	**Frequency**
AGE	-	Range: 16-69; Mode: 18 Average: 19.55 (SD = 5.503)	
GENDER	Male	-	29
	Female	-	71
HOME LANGUAGE	English/European	English, Polish, Finnish	50
	African	Zulu, Xhosa, Sepedi, Tswana, Sesotho, Swazi, Tsonga, Tshivenda	50
RACE	White	White	34
	Non-White	African, Indian, Coloured	66
SOCIO-ECONOMIC STATUS	High	Parental occupation: Levels 1-4 on the Hall-Jones Scale of Occupational Prestige; Parental education: above Matric	47
	Low	Parental occupation: Levels 5-7 on the Hall-Jones Scale of Occupational Prestige; Parental education: below Matric	53
MATRIC PERFORMANCE	-	Average symbol: A=1, B=2, C=3, D-4, E=5, F=6	1=12; 2=14; 3=35; 4=31; 5=6; 6=2
PSYCHOLOGY PERFORMANCE	-	75+ = 1 70-74 = 2 60-69 = 3 50-59 = 4 <50 = 5	1=17; 2=11; 3=25; 4=28; 5=19

Research Design

The study epitomized a classic non-experimental design aimed at exploring the natural occurrence of a phenomenon as opposed to manipulating performance (Rosnow & Rosenthal, 1996). A single set of cross-sectional measurements was conducted in relation to the variables of interest, and no attempt was made to alter or change these. It is worth noting, however, that the typical weakness of non-experimental research, namely the lack of ability to establish causation, does not have such profound implications in this specific study, which aimed to determine whether bias could be established empirically on the basis of RAPM performance (Rosnow & Rosenthal, 1996).

Instruments

The instrumentation utilized had a similar structure to that employed by Rushton et al. (2004).

The Demographic Questionnaire (Refer to Appendix B)

Designed by the researcher, this brief questionnaire was utilized to capture information regarding a number of pertinent demographic variables within the study, including age, gender, race, home language, language of education, type of schooling, academic performance, and parental occupation and level of education (as an estimate of socio-economic status). It consisted primarily of short, closed-ended questions and took most participants approximately five minutes to complete.

The Raven's Advanced Progressive Matrices (RAPM) (Refer to Appendix C)

The RAPM can essentially be regarded as a test of eductive ability - the ability to perceive and think clearly, tease out logical relationships and manage problem-solving goals. They are also widely regarded as one of the best single measures of *g* available (Carpenter et al., 1990; Paul, 1985; Raven et al., 1998). In this study, both the training and familiarisation set of twelve items (Set I) and the second set (Set II) of thirty-six items yielding a more accurate score were utilised. Although increasing steadily in difficulty, all the items across both sets follow the same format involving the respondent accurately completing the missing section of a pattern presented in a three-by-three matrix by selecting the correct option from eight presented below (Babcock, 2002; Paul, 1985; Raven, 2000; Raven et al., 1998). In order to provide a measure of intellectual ability as opposed to intellectual efficiency and ensure the attempt of as many items as possible, both sets were administered with no time constraint (Forbes, 1964; Paul, 1985; Spreen & Strauss, 1998).

The RAPM are generally considered particularly reliable, with reported test-retest reliabilities ranging between 0.76 and 0.9 (Bors & Forrin, 1995, as cited in Bors & Stokes, 1998; Kaplan & Saccuzzo, 2001; Raven, 1990, as cited in Melck, 2003; Raven et al., 1998; Spreen & Strauss, 1998), and reported internal consistency estimates ranging between 0.8 and 0.9 (Alderton & Larson, 1990; Bors & Stokes, 1998; Melck, 2003; Murphy & Davidshofer, 2001; Paul, 1985; Rushton & Skuy, 2000; Rushton et al., 2003). The RAPM have also been shown to be reliable across a wide range of geographic locations and

populations (Matthews, 1988, as cited in Melck, 2003; Paul, 1985; Raven et al., 1998). Numerous multidimensional scaling projects and factor analytic studies have demonstrated the validity of the test (c.f. Alderton & Larsen, 1990; Arthur & Woehr, 1993; Marshalek, Lohman and Snow (1983, as cited in Kaplan & Saccuzzo, 2001); Snow, Kyllonen and Marshalek (1984, as cited in Carpenter et al., 1990) among others).

The Similarities sub-test of the South African Wechsler Adult Intelligence Scale (SAWAIS)

The Similarities subtest is a verbal subtest of the South African Wechsler Adult Intelligence Scale designed to elicit information on a subject's verbal concept-formation and associative thinking, ability to generalise and ability to draw abstractions (Spreen & Strauss, 1998). The sub-test consists of twelve pairs of words, for each of which the subject must identify what the pairing has in common. The greater the level of abstraction of the generalisation given, the higher the respondent scores – with a maximum of two points per pair (Lezak, 1995). The sub-test is regarded as an effective measure of *g* (Lezak, 1995), and is utilised in this study as a more verbally-based measure of the same construct as the primarily non-verbal RAPM.

Research has shown that the Similarities sub-test of the WAIS-III (Wechsler Adult Intelligence Scale-III) is generally reliable, and a good measure across race and gender, although affected by level of education and age (Lezak, 1995). No psychometric data, however, is available with regard to the functioning of the Similarities sub-test of the SAWAIS (N.I.P.R., 1965). It is a very short sub-test, and generally takes less than fifteen minutes to complete.

The Reading Comprehension sub-test of the Stanford Diagnostic Reading Test (SDRT) – Blue Level

Utilised within this study as a measure of English language comprehension, the Reading Comprehension sub-test of the Stanford Diagnostic Reading Test (Blue Level) consists of a series of passages with related multiple choice questions assessing the individual's understanding of the passage on two levels – literal and inferential. Questions relating to literal comprehension assess the respondent's grasp of basic facts stated explicitly in the passage, while questions related to inferential comprehension assess the respondent's ability "to make inferences, draw conclusions, predict outcomes, evaluate situations, see cause and effect relationships, make comparisons and contrasts, understand characterization, verify the truthfulness or relevance statements and understand the author's

purpose, bias, tome or mood…" (Karlsen & Gardner, 1995, p.29). On this basis, it stands to reason that those more successful in achieving inferential comprehension may exhibit higher levels of *g,* and vice-versa, while one would not necessarily expect there to be a relationship between literal comprehension and *g*. The passages become progressively harder throughout the test, draw on content from fiction, the social and natural sciences, history and human interest stories and are written at a readability level suitable for high school and first-year college students (grades nine to thirteen) (Spreen & Strauss, 1998; Karlsen & Gardner, 1995).

The original version of the sub-test contains nine passages and sixty items, however, the length of the test was deemed not suitable for this study given the time constraints. In order to preserve a reasonable measure of comprehensive ability and the order of item difficulty, while still shortening the test to a suitable length, the third, sixth, seventh and ninth passages were omitted. Students thus answered items 1-13, 20-30 and 46-53 (a total of thirty-two items based on five passages, fifteen of which assessed literal comprehension and seventeen of which assessed inferential comprehension). This adaptation, while suitable for the study, negated most of the previous psychometric information available on the sub-test, particularly estimates of its reliability, which had proved reasonable for the sub-test as a whole in the past (Spreen & Strauss, 1998).

Procedure

After receiving clearance from both the Committee for Research on Human Subjects - Humanities (Protocol Number: H020805) and the Graduate School, the researcher began by setting up two testing periods as close together as possible for the two Psychology One classes. This resulted in the research being carried out over two sessions, both in the afternoon and in the lecture venue, but two days apart. Due to the researcher's involvement with the first year programme, it was also necessary to find two people unrelated to the first year programme who would be willing to carry out the initial introduction of the research and request volunteers. Once those students who did not wish to participate had left, the researcher then entered and gave a more detailed description of the aims, expectations, and format of the study. At the end of this, those students who still wished to participate were asked to read through and sign an informed consent form, which was then sealed in an envelope in the presence of the students.

The researcher then carried out a standardised demonstration and training procedure with regard to the RAPM. Respondents and the researcher worked together through selected items of Set I utilizing the overhead and verbal discussion until all participants indicated they were comfortable with the format, requirements and problem-solving logic of the test. Research has indicated that a single brief training session can drastically reduce the likelihood of misunderstanding or unfamiliarity with the requirements of the test acting as an extraneous variable (Crawford-Nutt, 1976; Denney & Heidrich, 1990, as cited in Spreen & Strauss, 1998; Rushton et al., 2003).

Respondents were then asked to complete the Similarities sub-test of the SAWAIS, the RAPM Set II and the adapted Reading Comprehension sub-test of the SDRT (compiled into a single test pack, except the RAPM Set II questions) independently in that order. In addition to the researcher, several research assistants were available to assist students with administrative issues such as handing out the tests and dealing with non-content related queries. Respondents were permitted to leave as soon as they finished or whenever they wished to, however they were not allowed to remove any of the test material from the venue except the informed consent sheet. Those leaving were thus asked to place their test pack, regardless of its completeness, within a partially sealed box near to the room's entrance. Respondents were provided with as much time as they wished to complete the test (the longest took two hours and eight minutes), and were also given the researcher's contact details in order to allow them to receive feedback about the study's results if they so wished.

Threats to Validity

There were a number of potential threats to validity that arose during the research, although the exact impact of each on the results remains undetermined.

Firstly, the heavy rate of attrition experienced within the sample had far-reaching negative effects. Twenty-seven percent of the original sample was not utilized in the final analysis, due primarily to a lack of completion of the tests. One possible reason for this was fatigue or boredom experienced by respondents during the test, many students simply left prior to completing the study, and several others complained of the number of tests or time restrictions. Related to this, Raven et al. (1998) point out that working through these types of tests, in particular the RAPM, is a demanding and difficult activity, and that the best possible results are unlikely to be achieved unless those taking the test feel some form of

commitment or motivation to perform well (Grieve & Viljoen, 2000). Although all participants were volunteers, suggesting a certain level of interest and dedication, it was not possible to assess the mindset of individuals. In addition, factors such as fatigue, stress or ill health have all been implicated in reducing performance on the RAPM and psychometric tests generally (Raven et al., 1998). Again, it was not possible to assess whether these or other history effects, such as having taken the test before, were present in the sample (Rosnow & Rosenthal, 1996).

The small nature of the final sample also affected the types of analytical procedures that could be carried out, thus impacting on the statistical validity of the study. In order to enable certain analyses, it became necessary to dichotimise several of the variables, resulting in a loss of information with regard to those factors (home language, race, socio-economic status). External validity, the generalisability of the research to other contexts and populations, was also negatively affected by the small sample size, as the group ultimately represented only a small, albeit relatively diverse, subsection of students from one university (Melck, 2003; Rosnow & Rosenthal, 1996).

On a more pragmatic level, despite the assumption of reasonably similar levels of 'test-wiseness' and ability to engage in English between respondents, it was not possible to ensure this was always the case. It is feasible that some respondents may not have understood the training procedure or requirements of the tests fully, or may have interpreted certain points incorrectly. In addition, despite every effort being made to standardize the training procedure, instructions and explanations given across the two administrations, it is likely that these were not directly equivalent across the groups. The dual administration, and gap in time between the two, also gave rise to a potential problem of diffusion. Given the likely overlap of students on the two diagonals in terms of other subjects, it is certainly possible that they discussed the research in some way outside of the actual study. In particular, this may have affected attendance of the second administration or performance on the tests themselves.

Data Analysis

Prior to beginning an investigation of the data collected, it was necessary to establish that the tests utilized in the investigation were reliable. Internal consistency reliability estimates, as indicated by Cronbach Coefficient Alphas, were therefore calculated for each of the three tests – namely, the RAPM, the Similarities sub-test of the SAWAIS, and the adapted Reading Comprehension sub-test of

the SDRT – as a measure of their consistency (Murphy & Davidshofer, 2001). This was particularly important in the case of the latter two tests, for which little prior psychometric information was available.

In addition, two other sets of analyses were carried out prior to investigating the research questions:

Firstly, a set of Chi-Squared Tests of Association were carried out in order to establish the nature of the relationship between certain of the demographic variables collected in the study, namely race, gender, home language, socio-economic status (as estimated by parental occupation and level of education), and academic performance (as indicated by average symbol in Matric and the mark achieved for the first module of psychology) (Howell, 1997). The aim of these analyses was to establish whether the independent variables utilized in the study, namely gender and home language, were independent of other, equally important demographic factors also likely to impact on psychometric test performance. It is worth noting that the small nature of the final sample utilized for analysis necessitated a conversion of many of these variables to a dichotomous nominal form, including race (White/non-White); home-language (English/African); and socioeconomic status (high/low), thus losing much of the fine detail of the data.

Secondly, it was necessary to establish whether the data collected was suitable for parametric analysis. In addition to assessing whether it was reasonable to assume random independent sampling and an interval scale of measure for the dependent variables, namely the three sets of test scores, it was also necessary to establish whether these were normally distributed. This was done utilizing histograms, measures of central tendency and the Kolmogorov-Smirnoff test of normality (Howell, 1997). Equality of variance checks were carried out only as needed in specific analyses, and the criterion of additive means was simply assumed where the other assumptions had been met (Howell, 1997).

In order to investigate the relationship between the RAPM and the Similarities sub-test of the SAWAIS, both on an overall basis and in terms of differences between gender or home language groups, a series of Pearson Product Moment Correlation Coefficients (r) were calculated. These allowed for the strength and direction of the association between the two tests to be assessed. To establish whether significant differences in the correlations between the groups existed, a series of tests for two independent correlations was carried out, including the usual standardization procedure (Fisher-z transformations) (Howell, 1997).

A similar procedure was followed for assessing the relationship between the RAPM and the Reading Comprehension sub-test of the SDRT, except that the non-normal nature of the SDRT data necessitated the confirmatory calculation of both Spearman's Rank-Order Correlation Coefficients (r_s) and Pearson Product-Moment Correlations (r) to assess the relationship between the tests (Howell, 1997). As no substantial differences appeared between the parametric and non-parametric calculations, the Pearson Product-Moment Correlations were utilized to investigate the statistical significance of the difference between the correlations. Furthermore, the adapted Reading Comprehension sub-test scores were analysed both overall, and in terms of those items relating to literal comprehension and those items relating to inferential comprehension, and a confirmatory analysis of the relationship between the Similarities sub-test of the SAWAIS and the adapted Reading Comprehension sub-test of the SDRT was carried out.

In order to assess the level of item difficulty on the basis of gender, a series of p-values (the proportion of people getting the item right in relation to those attempting it) were calculated separately for men and women (Murphy & Davidshofer, 2001). These were then scanned for discrepancies in relation to the 'absolute order' phenomenon well established in research (Raven et al., 1998). Linear regression lines were then computed in order to examine whether the relationship between question number and item difficulty remained the same for men and women across the test. A similar procedure was then repeated for assessing the level of item difficulty on the basis of home language.

Assessing the final question of qualitative differences in performance on the basis of ability (measured by overall score on the RAPM), gender and home language through an analysis of the types of errors made involved a fairly complex procedure. Firstly, it was necessary to establish which specific error type each distractor represented. Babcock's (2002) categorization of the distractors for the first thirty-one items was utilized. It was necessary, however, for the researcher to classify the last thirty-five distractors (related to items thirty-one to thirty-six) as this had never been done before (refer to Appendix D). This was carried out by utilizing three independent categorizations of the errors which were then compared. Inter-rater reliability was fairly high (approximately eighty-six percent) and in cases of disagreement, discussion was carried out until consensus was reached. Despite the high level of inter-rater agreement, the inherently problematic nature of this post-hoc categorization should not be overlooked (Vodegel Matzin et al., 1994).

The standardization procedure utilized by Babcock (2002) was then replicated in order to resolve the issue of unequal distribution of error types across the test: the occurrence of each error type on each individual problem was determined, followed by a determination of the occurrence of each error type on incorrectly solved problems for each individual subject. The probability of the occurrence of each error type for individual subjects was then assessed, and the total number of each error type for each individual subject calculated. Finally, the proportion of each error type for each subject was computed, and the deviation from chance selection of each error type assessed (Babcock, 2002). Put more simply, a complex procedure was used to convert the number of error types to proportions, thus compensating for the unequal number of different error types represented in the test.

A series of frequency tables reflecting the proportion of each error type on each item was then generated, as were similar tables representing the frequency of choices on the basis of ability, on the basis of gender and on the basis of home language. A series of two-by-four (gender/ ability/ home language x error type) Repeated-measures Analyses of Variances (ANOVAs) were then carried out to assess whether any significant differences existed in terms of the types of errors made on the basis of the relevant factor (the interaction), as well as any significant differences between the number of error types made or the number of errors made between the different groups (main effects). Where significant results for the interaction were obtained, further post-hoc analyses, in the form of multiple comparisons, were carried out to assess the specific nature of the relationships by examining how different error types were affected by the relevant independent variable (Howell, 1997).

These analyses were utilized to answer the four research questions posed by the research project, and the findings are presented in the following chapters dealing with the results of the research and their implications.

Chapter 4: Results

In order to address the research questions, a series of statistical analyses were carried out, beginning with attempts to establish the reliability of the tests utilized, the degree of relationship between the independent variables assessed in the study and the suitability of the data for parametric analysis. These pre-analyses were followed by a series of statistical techniques designed to address the four research questions posed in order. The results obtained for these analyses were as follows:

Reliability of the Instruments

Initially, it was necessary to establish that the tests utilized in the investigation, namely the RAPM, the Similarities sub-test of the SAWAIS and the adapted Reading Comprehension sub-test of the SDRT, were reliable or measured consistently within the study (Murphy & Davidshofer, 2001). This was especially important in the case of the Similarities sub-test, for which no psychometric data was found, and the adapted Reading Comprehension sub-test, the psychometric properties of which were potentially altered by the changes made to it. Internal consistency reliability estimates, as indicated by Cronbach Coefficient Alphas, were therefore calculated for all three tests.

Table 2: Reliability Estimates for the Tests

	Cronbach Coefficient Alpha	
	Raw Variables	***Standardised Variables***
RAPM (Raven's Advanced Progressive Matrices – Set II)	0.922741	0.920643
SIM (Similarities sub-test of the SAWAIS)	0.659118	0.664460
ARC (adapted Reading Comprehension sub-test of the SDRT)	0.790122	0.798650

Table 2 indicates that within the study the RAPM displayed a particularly strong level of consistency ($\alpha = 0.92$) and that the adapted Reading Comprehension sub-test also displayed a high level of internal consistency reliability ($\alpha = 0.79$). The Similarities sub-test reliability estimate of $\alpha = 0.66$, however, while still acceptable, was only moderately strong, and suggested that there may have been some lack of consistency in its measurement of *g*. Generally, though, all three measures provided acceptably consistent measurement for use in the study.

Relationship between the Independent Variables

In order to assess the degree of relationship between the independent variables, namely gender and home language, and other, equally important demographic variables shown to determine test performance to some degree, namely race, socio-economic status (as estimated by parental occupation and level of education), and academic performance (as indicated by average symbol in Matric and the mark achieved for the first module of psychology)**,** a series of Chi-Squared Tests of Association were carried out (Howell, 1997; Owen, 1996; Rushton et al., 2004; Skuy, 2003). These tests, aimed at establishing whether the distribution of one categorical variable is contingent upon a second categorical variable, were particularly suitable as the small nature of the final sample necessitated a conversion of many of the demographic variables to a dichotomous, nominal form, including home-language (English/African); race (White/non-White) and socioeconomic status (high/low) (Howell, 1997).

In addition to the p-values for the Chi-Squared Tests themselves, estimates of significance on the basis of Yate's correction for continuity, utilized when expected values are likely to be particularly small, were also included when assessing the significance of the relationship between the variables (Howell, 1997). It is worth noting that no differences were found between these two types of estimates for any of the tests carried out. The Phi Coefficients, estimates of the statistical correlation between the variables, were also included, as indicators of the relative strength of the relationships between the variables.

Table 3: Chi-Squared Tests of Association between Variables in the Study

	GENDER			**HOME LANGUAGE**		
Variable	***Pearson's Chi-Square***	***Continuity Adjusted Chi-Square***	***Phi Coefficient***	***Pearson's Chi-Square***	***Continuity Adjusted Chi-Square***	***Phi Coefficient***
GENDER	-	-	-	0.1229	0.1861	-0.1543
H.LANG.	0.1229	0.1861	-0.1543	-	-	-
RACE	0.0951	0.1478	0.1669	<0.0001	<0.0001	0.8864
S.E.S.	0.7809	0.9542	0.0278	<0.0001	<0.0001	-0.7013
MATRIC	0.1774	0.2660	0.2764	<0.0001	<0.0001	0.5387
J.PSYC.	0.1585	0.1039	0.2569	<0.0001	<0.0001	0.6410

Results, as shown in Table 3, indicated that there were no significant relationships between gender and any of the other variables assessed in the study ($p>0.05$ in all cases), suggesting that it was possible to determine the influence of gender on test performance independent of other factors. Home language, on the other hand, was shown to be significantly related to all the other variables assessed ($p<0.0001$ in all cases) except gender ($p = 0.1229$). Thus, although utilized as the variable of interest in the study, it proved impossible to determine whether the results obtained were directly and solely a product of the effect of home language, or whether they were also a product of the effects of race, socio-economic status and/or estimated academic ability, as indicated by the contingency of the distribution of home language on the distributions of these other variables (Howell, 1997). The strongest relationship, not unexpectedly, was shown to exist between race and home language ($\Phi = +0.89$), followed by race and socio-economic status ($\Phi = -0.70$), suggesting a very high degree of relationship between home language and these two factors in particular. It was thus very likely that these would act as extraneous variables within the study (Rosnow & Rosenthal, 1996).

Normality of the Data

In order to utilize parametric techniques for statistical analysis, it is necessary to meet five assumptions, including random, independent sampling; additive means and at least an interval scale of measure for the dependent variable/s (Howell, 1997). In addition, it is assumed that there is homogeneity of variance between the groups and that the data are distributed reasonably normally (Howell, 1997). While there was little doubt that the interval scale of measure could be assumed for the three tests used in the study, it was necessary to assess the normality of their distributions. This was done utilizing histograms, measures of central tendency and Kolmogorov-Smirnov Tests of Normality.

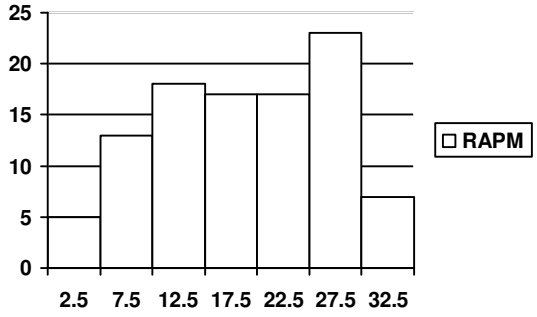


Figure 1: Histogram: RAPM

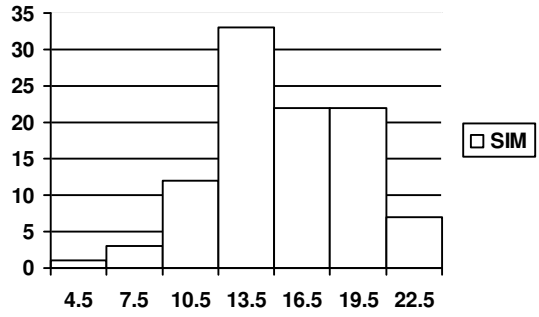


Figure 2: Histogram: Similarities

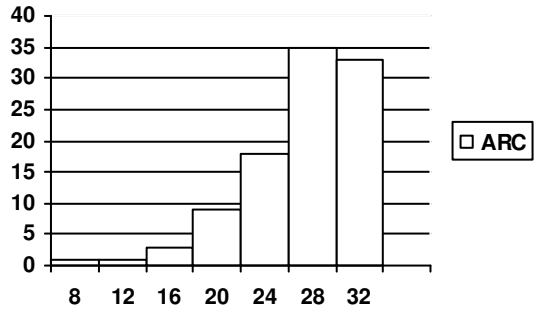


Figure 3: Histogram: adapted Reading Comprehension

Examination of the histograms for the three tests revealed that the RAPM were approximately normally distributed and that the Similarities sub-test data were only very slightly skewed to the left (refer to Figures 1 and 2). The adapted Reading Comprehension sub-test, however, as shown in Figure 3, appeared to be significantly negatively skewed, indicating that the majority of respondents performed extremely well on the test (this was not unexpected, given the nature of the test). These visual examinations were confirmed through examination of the measures of central tendency for the three tests, focusing particularly on the distance between the mean and median (Howell, 1997).

Table 4: Basic Descriptive Statistics for the Tests

Simple Descriptive Statistics							
Variable	***N***	***Mean***	***Std. Dev.***	***Median***	***Mode***	***Minimum***	***Maximum***
RAPM	100	18.26	8.51217	18.5	27	1	34
SIM	100	15.07	3.83565	15	14	3	23
ARC	100	26.67	4.61892	28	30	8	32
ARCLIT	100	12.86	2.07447	13	13	3	17
ARCINF	100	13.81	2.83091	15	14	5	15

As can be seen in Table 4, this examination tended to confirm that the RAPM and Similarities sub-test were sufficiently normally distributed to allow parametric analysis, and also suggested that parametric analysis might be feasible on the basis of overall adapted Reading Comprehension scores (mean = 26.67; median = 28.00)

An assessment of the normality of the distributions utilizing Kolmogorov-Smirnov Tests of Normality was also carried out.

Table 5: Kolmogorov-Smirnov Tests of Normality for the Tests

	Kolmogorov-Smirnov Goodness-of-Fit Test for Normal Distribution	
	Statistic (D)	***p - Value***
RAPM	0.09994890	$p = 0.015$
SIM	0.10722349	$p < 0.010$
ARC	0.18331	$p < 0.010$

Results of the Kolmogorov-Smirnov Tests, as shown in Table 5, indicated that only the RAPM could be considered normally distributed ($p = 0.015$), however it is worth noting that these tests are particularly sensitive to discrepancies, and that no exceptionally high levels of significance were obtained (Howell, 1997).

Ultimately, it was decided that the distributions of the three tests were sufficiently normal to allow for parametric analysis, although in the case of calculations utilizing the adapted Reading Comprehension test scores, it was deemed advisable to carry out corresponding non-parametric analyses as well as a means of confirmation. In terms of the other assumptions, random independent sampling and additive means were assumed and homogeneity of variance assessed only where necessary (sufficient homogeneity was found in those cases).

Relationship between the RAPM and the Similarities sub-test of the SAWAIS

In order to establish the level of construct comparability in a South African sample, the study examined the relationship between the RAPM, as a primarily non-verbal measure of *g*, and the Similarities sub-test of the SAWAIS, a well-established verbally-based measure of *g* (Lezak, 1995), both on an overall basis and on the basis of gender and home language. This was done by calculating a series of Pearson Product Moment Correlation Coefficients (r), indicating the strength and direction of the association between the two tests. To establish whether significant differences in the correlations achieved between the groups existed, a series of tests for two independent correlations was carried out, including the usual standardization procedure (Fisher-z transformations) (Howell, 1997).

Table 6: Pearson Correlation Coefficients: RAPM and Similarities sub-test

	Pearson Correlation Coefficients			***Pearson Correlation Coefficients***			***Pearson Correlation Coefficients***	
		RAPM			**RAPM**			**RAPM**
Overall N=100	**RAPM**	1.0000	**Male N=29**	**RAPM**	1.0000	**English N=50**	**RAPM**	1.0000
	SIM	0.66428 $p < 0.0001$		**SIM**	0.69165 $p < 0.0001$		**SIM**	0.56507 $p < 0.0001$
			Female N=71	**RAPM**	1.0000	**African N=50**	**RAPM**	1.0000
				SIM	0.65053 $p < 0.0001$		**SIM**	0.50991 $p = 0.0002$

Results, as shown in Table 6, indicated that a relatively strong positive significant relationship existed between the two tests ($r = 0.66$; $p<0.0001$), confirmed by the scatterplot (refer to Appendix E, Figure 4), and that the relationship between the measures for gender appeared relatively similar ($r_{males} = 0.69$; $p<0.0001$; $r_{females} = 0.65$; $p<0.0001$). The test of significant differences between these two correlations was also shown to be non-significant ($z = 0.52$; $p>0.05$), indicating that the relationship between the two tests in their measurement of *g* was similar between the genders. Although the relationship between the two measures for home language appeared more disparate and the relationship between the tests for African language speakers was not significant at the 0.0001 level ($r_{English} = 0.57$; $p<0.0001$; $r_{African} = 0.51$; $p = 0.0002$), the test of significant differences between the two correlations again revealed no significant difference ($z = 0.54$; $p>0.05$), suggesting that the relationship between the two tests in their measurement of *g* was similar between the language groups. The RAPM thus appeared to assess *g* similarly between the gender and home language groups in relation to an independent measure of the same construct, thereby establishing a relatively high degree of construct comparability.

Relationship between the RAPM and the adapted Reading Comprehension sub-test of the SDRT

In order to explore the relationship between general, literal and inferential English comprehension ability and performance on the RAPM, a similar series of correlational analyses were carried out between the RAPM and the adapted version of the Reading Comprehension sub-test of the SDRT. In addition to Pearson Product Moment Correlation Coefficients (r), the non-normal distribution of scores for the adapted Reading Comprehension sub-test also necessitated the calculation of Spearman Rank-Order Correlation Coefficients (r_s) (a parallel non-parametric assessment of correlation between two

variables) (Howell, 1997). A confirmatory analysis of the relationship between the adapted Reading Comprehension sub-test and the Similarities sub-test of the SAWAIS was also carried out.

Table 7: Pearson and Spearman Correlation Coefficients: RAPM, the Similarities sub-test and the adapted Reading Comprehension sub-test

		Pearson Correlation Coefficients			***Spearman Correlation Coefficients***		
		RAPM	**SIM**	**ARCLIT**	**RAPM**	**SIM**	**ARCLIT**
Overall N=100	**ARC**	0.65270 p< 0.0001	0.57032 p < 0.0001	-	0.64173 p< 0.0001	0.60231 p < 0.0001	-
	ARCLIT	0.61587 p < 0.0001	0.49125 p < 0.0001	-	0.61779 p < 0.0001	0.49774 p < 0.0001	-
	ARCINF	0.61365 p < 0.0001	0.57055 p < 0.0001	0.76771 p < 0.0001	0.58708 p < 0.0001	0.60056 p < 0.0001	0.63874 p < 0.0001

These results, as presented in Table 7, indicated a moderately strong relationship between the RAPM and the adapted Reading Comprehension sub-test ($r = 0.65$; $p<0.0001$), with a very similar relationship between literal comprehension and RAPM performance ($r = 0.62$; $p<0.0001$) and inferential comprehension and RAPM performance ($r = 0.61$; $p<0.0001$). The test of significant difference between these two correlations was non-significant ($z = -0.02$; $p>0.05$). Somewhat unusually, the Pearson Coefficients were higher than the Spearman Coefficients, however analysis of the scatterplots between the tests showed this to be the result of significant ceiling effects, as well as one or two influential outliers (refer to Appendix E, Figures 5 and 6). The relationship between literal and inferential comprehension was fairly strong ($r = 0.77$; $p<0.0001$), but the lack of extreme strength in the relationship confirmed the division suggested by Karlsen and Gardner (1995) in terms of the separate abilities the two aspects assess.

While the relationship between the adapted Reading Comprehension sub-test and the Similarities sub-test was not quite as strong ($r = 0.57$; $p<0.0001$), it was still significant. This lower relationship was somewhat surprising, however, given the verbal nature of the Similarities test. The relationship between literal comprehension and the Similarities sub-test ($r = 0.49$; $p<0.0001$) was also somewhat lower than that between inferential comprehension and the Similarities sub-test ($r = 0.57$; $p<0.0001$), although the difference between the two was not significant ($z = 0.77$; $p>0.05$).

Very similar correlation estimates were obtained using Spearman Rank Correlation Coefficients, except for the correlation between the literal and inferential comprehension scores ($r_s = 0.64$; $p<0.0001$), thus

it seemed unlikely any substantive differences would be found with regard to the relationships between the tests. In addition, tests of correlation are typically quite robust, and the departures from normality in this instance were relatively mild (Howell, 1997).

An analysis of the relationship between the tests on the basis of gender and home language was also carried out (refer to Table 8).

Table 8: Pearson and Spearman Correlation Coefficients on the basis of Gender and Home Language: RAPM and the adapted Reading Comprehension sub-test

		Pearson Correlation Coefficients	***Spearman Correlation Coefficients***			***Pearson Correlation Coefficients***	***Spearman Correlation Coefficients***
		ARC				**ARC**	
Male N=29	**RAPM**	0.65310 p = 0.0001	0.61670 p = 0.0004	**English N=50**	**RAPM**	0.64229 p < 0.0001	0.47838 p = 0.0004
Female N=71	**RAPM**	0.65407 p < 0.0001	0.65224 p < 0.0001	**African N=50**	**RAPM**	0.52155 p = 0.0001	0.47397 p = 0.0005

These calculations indicated that the relationship between performance on the RAPM and on the adapted Reading Comprehension test was very similar for males and females ($r_{males} = 0.65$; $p<0.0001$; $r_{females} = 0.65$, $p<0.0001$) and not significantly different ($z = -0.01$; $p>0.05$); while the relationship between performance on the two tests on the basis of home language was somewhat disparate ($r_{English} = 0.64$; $p<0.0001$; $r_{African} = 0.52$; $p<0.0001$) but also not significantly different ($z = 1.27$; $p>0.05$). It is worth noting that the Spearman correlation coefficients calculated for these were somewhat different from the Pearson Coefficients, and some were not significant at the 0.0001 level of significance, however they did not appear particularly different from each other on the basis of either gender or home language ($r_{s\text{-}males} = 0.62$; $p = 0.0004$; $r_{s\text{-}females} = 0.65$; $p<0.0001$; $r_{s\text{-}English} = 0.48$; $p=0.0004$; $r_{s\text{-}African} = 0.47$; $p=0.0005$).

These results suggested that the relationship between RAPM performance and English comprehension ability did not differ on the basis of either gender or home language, thus refuting the notion of potential biases. Furthermore, the lack of distinction in the relationship between RAPM performance and literal comprehension and RAPM performance and inferential comprehension suggested that neither basic English ability nor the ability to comprehend deeper, more subtle levels of meaning in English affected RAPM performance. The weaker relationship found between the adapted Reading

Comprehension sub-test and the Similarities sub-test, which one would have expected to correlate more strongly given its verbal nature, was particularly surprising, however could perhaps be attributed to the ceiling effects created by the exceptionally high levels of performance on the adapted Reading Comprehension sub-test.

Item Difficulty

In order to assess score comparability, analyses of item difficulty on the basis of both gender and home language were carried out. A series of p-values (the proportion of people getting the item right in relation to those attempting it) were calculated separately for men and women, and for English and African speakers (Murphy & Davidshofer, 2001). These were then scanned for discrepancies in relation to the 'absolute order' phenomenon well established in research (Raven et al., 1998). Linear regression lines were then computed in order to examine whether the relationship between question number and item difficulty remained the same across the test for both men and women and both English and African speakers respectively, thus assessing whether the different groups found the items 'equally difficult'.

On the whole, the item difficulties on the basis of gender conformed to the established order of item difficulty, although a few items appeared out of sequence in that they were more difficult than one might expect given their relative position (Items 8, 13, 21, 22, 24 and 28) (refer to Appendix F, Table 9). The comparative p-values for gender appeared roughly similar, although nine items evidenced differences of greater than 0.1 between the respective p-values (Items 4, 10, 20, 22, 23, 24, 25, 28 and 33). In all cases, a greater proportion of males were likely to get the item right, except for Items 28 and 33, where a greater proportion of females were likely to get the item correct.

Item difficulties on the basis of home language, on the other hand, did not conform all that well to the established order of item difficulty, particularly for English speakers, for whom Items 13, 15, 22, 30, 33, 34 and 35 appeared out of sequence (refer to Appendix F, Table 10). For African language speakers, only Items 8, 18 and 23 appeared substantially out of place, however huge discrepancies between the comparative p-values of the two language groups were evident. Only four items (Items 28, 29, 31 and 36) evidenced differences of 0.1 or less between the comparative p-values, while several showed discrepancies of over 0.3 (Items 8, 12, 14, 18, 27 and 35). In all cases with the exception of Item 36, a greater proportion of English speakers were likely to get the item correct.

Regression lines estimating the effect of gender and home language respectively on the level of item difficulty experienced by respondents were then calculated, utilizing a backwards elimination model. In other words, all predictors were initially included in the model, and then those that contributed least to the model were removed one at a time until all the remaining predictors were significant (Howell, 1997).

Table 9: Regression of p-values for Gender

Regression of p-values for Gender			
All Variables Entered: R-Square = 0.8358 and C(p) = 4.0000			
Variable	***Parameter Estimate***	***F Value***	***Pr > F***
Intercept	0.89496	827.21	<.0001
QuesNo	-0.01959	178.40	<.0001
Group	-0.04954	1.27	0.2643
QGroup	0.00074807	0.13	0.7195

Table 10: Regression of p-values for Gender: Backwards Elimination Step One

Regression of p-values for Gender			
Backward Elimination: Step 1			
Variable QGroup Removed: R-Square = 0.8355 and C (p) = 2.1301			
Variable	***Parameter Estimate***	***F Value***	***Pr > F***
Intercept	0.88804	1330.73	<.0001
QuesNo	-0.01922	347.70	<.0001
Group	-0.03570	2.78	0.1000

In the case of gender, examination of the studentized residuals and Coot's D Influence Statistic revealed relatively few outliers (refer to Appendix G, Figures 16 and 17). As shown in Tables 9 and 10, the regression itself indicated no significant differences between the lines on the basis of group, even after the elimination of the other non-significant predictor ($p = 0.1000$). This suggested that gender did not contribute to predicting the difficulty of the item, and thus that males and females did not find the items differentially difficult. This was also illustrated visually through examining the similarity of the slopes and intercepts of the calculated regression lines (refer to Appendix G, Figure 18).

Table 11: Regression of p-values for Home Language

Regression of p-values for Home Language *All Variables Entered: R-Square = 0.8700 and C(p) = 4.0000*			
Variable	***Parameter Estimate***	***F Value***	***Pr > F***
Intercept	1.01575	1042.48	<.0001
QuesNo	-0.02100	200.61	<.0001
Group	-0.31190	49.15	<.0001
QGroup	0.00389	3.44	0.0681

In the case of home language, examination of the studentized residuals and Coot's D Influence Statistic also revealed relatively few outliers, although the two that were evident were extreme (refer to Appendix G, Figures 19 and 20). The regression itself, as shown in Table 11, indicated significant differences between the lines on the basis of group even within the model including all predictors ($p<0.0001$). This suggested that home language did contribute significantly to predicting the difficulty of the item, and thus that English and African language speakers found the items differentially difficult. This was also illustrated visually through examining the disparity of the slopes and intercepts of the calculated regression lines (refer to Appendix G, Figure 21). Interestingly, visual examination of the regression lines revealed that removal of the outliers would in fact accentuate the difference between the slopes and intercepts of the lines, rather than decrease it.

Analysis of Errors (Distractor Analysis)

Lastly, in order to establish whether performance differed qualitatively between groups of different ability level (high/low - as assessed by overall score on the RAPM), gender (male/female) and home language (English/African), the frequency of each type of error made between these groups on individual items was examined. After categorizing each distractor as one specific error type, calculating the number of errors made by the whole sample (refer to Table 12) and converting the number of errors made to proportions, thus compensating for the unequal number of error types represented in the test, a series of frequency tables reflecting the proportion of each error type on each item was generated (Babcock, 2002; Vodegel Matzin et al., 1994). Similar tables representing the frequency of error types made on the basis of ability, gender and home language were also generated (refer to Appendix H).

Table 12: Frequency of Error Types Prior to Conversion to Proportions

Frequency of Correct Answers and Error Types (Prior Conversion)			
Correct Answer / Error Type	***Frequency***	***Percent***	***Cumulative Frequency***
COR (Correct)	1826	51.90	1826
RP (Repetition)	115	3.27	1941
IC (Incomplete Correlate)	654	18.59	2595
WP (Wrong Principle)	653	18.56	3248
CI (Confluence of Ideas)	270	7.67	3518
Frequency Missing = 82			

Following this, a series of two-by-four (gender/ ability/ home language x error type) Repeated-measures Analyses of Variances (ANOVAs) were then carried out to assess whether any significant differences existed in terms of the types of errors made on the basis of the relevant factor (the interaction), as well as any significant differences between the number of error types made or the number of errors made between the different groups (main effects). Where significant results for the interaction were obtained, further post-hoc analyses, in the form of multiple comparisons, were carried out to assess the specific nature of the relationships by examining how different error types were affected by the relevant independent variable (Howell, 1997). The analyses thus assessed whether significant differences existed in the patterns of distractors chosen on the basis of the various groupings.

Although not the primary focus of this study, an analysis of the differences in distractor choice between ability groups was deemed useful, as it allowed for an assessment of the similarity of patterns of performance in the South African context to those found internationally. International research suggested that people of higher ability tended to make more incomplete correlate errors relative to other types of errors, whereas those of lower ability (as estimated by overall performance on the RAPM) tended to make more wrong principle errors, followed by repetition errors (Babcock, 2002; Raven et al., 1998).

Table 13: Repeated Measures ANOVA: Error Type x Ability

Repeated Measures Analysis of Variance (Error Type x Ability)				
	F Value	**Num DF**	**Den DF**	**Pr > F**
Error Type	74.74	3	96	< 0.0001
Ability	0.00	1	98	1.000
Error Type * Ability	9.49	3	96	< 0.0001

Results of the two-by-four ANOVA (ability (high/low) x error type (IC/CI/RP/WP)), shown in Table 13, indicated that the interaction between the two independent variables was significant ($F = 9.49$; $p<0.0001$), suggesting that significant differences existed in terms of the types of errors made on the basis of ability. Although the number of errors made between the different levels was not significantly different ($F = 0$; $p=1.000$), there were also significant differences between the type of errors made, regardless of ability level ($F = 74.74$; $p<0.0001$). Given that significant differences were obtained, a series of post-hoc analyses were carried out to determine where the specific differences lay (Howell, 1997).

Table 14: Post-hoc Analysis: Error Type x Ability

Post-Hoc Multiple Comparison Analyses (Error Type x Ability)			
	ABILITY (High = 1; Low = 2)		
Level of Err-Type	**Interaction (Err_Type*ABILITY) F Value and Pr > F**	**Main Effect One (Err-Type) F Value and Pr > F**	**Main Effect Two (ABILITY) F Value and Pr > F**
Errors 2-3 (RP-IC)	4.77 p = 0.0313	147.15 p < 0.0001	23.25 p < 0.0001
Errors 2-4 (RP-WP)	2.92 p = 0.0905	121.37 p < 0.0001	1.14 p = 0.2892
Errors 2-5 (RP-CI)	12.76 p = 0.0005	5.82 p = 0.0177	5.25 p = 0.0241
Errors 3-4 (IC-WP)	11.99 p = 0.0008	6.11 p = 0.0151	5.25 p = 0.0241
Errors 3-5 (IC-CI)	22.46 p < 0.0001	122.30 p < 0.0001	1.14 p = 0.2892
Errors 4-5 (WP-CI)	1.13 p = 0.2904	96.21 p < 0.0001	23.25 p < 0.0001
DF for all post-hoc analyses = F (1; 98)			

The post-hoc analyses, multiple comparisons of the means obtained between the ability and error groups, essentially consisted of a series of 2 x 2 ANOVAs (ability (high/low) x error type) designed to illustrate the pattern of errors relative to ability (refer to Table 14). This allowed for the identification of how ability affected the different error types.

Results indicated significant interactions between repetition errors and both incomplete correlate errors ($F = 4.77$; $p=0.0313$) and confluence of ideas errors ($F = 12.76$; $p=0.0005$), as well as between incomplete correlate errors and wrong principle ($F = 11.99$; $p=0.0008$) and confluence of ideas errors ($F = 22.46$; $p<0.0001$). In addition, all differences in number of error made, regardless of ability type, were significant ($p<0.05$ in all cases). Significant differences between the ability groups, regardless of error type, were obtained for four of the six post-hoc tests (only repetition-wrong principle and incomplete correlate-confluence of ideas analyses were non-significant ($p>0.05$)).

Closer examination of the exact means between the groups (as shown in Table 15) and interaction figures allowed for a finer assessment of the specific differences in error type.

Table 15: Basic Descriptive Statistics: Error Type x Ability

Simple Descriptive Statistics

Level of ABILITY	N	Error 2 (RP) Mean	Error 2 (RP) Std. Dev.	Error 3 (IC) Mean	Error 3 (IC) Std. Dev.	Error 4 (WP) Mean	Error 4 (WP) Std. Dev.	Error 5 (CI) Mean	Error 5 (CI) Std. Dev.
1 (High)	50	0.07731	0.1645	0.44049	0.2144	0.28128	0.1819	0.05806	0.1076
2 (Low)	50	0.05664	0.0619	0.30897	0.1115	0.33555	0.0944	0.15599	0.1162

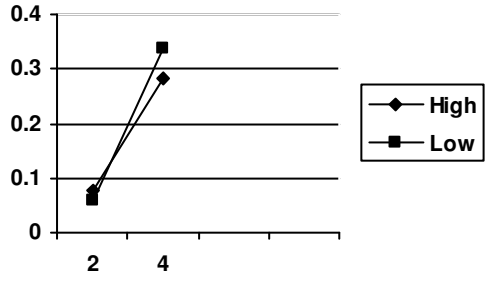


Figure 4: Post-hocA: RP and WP

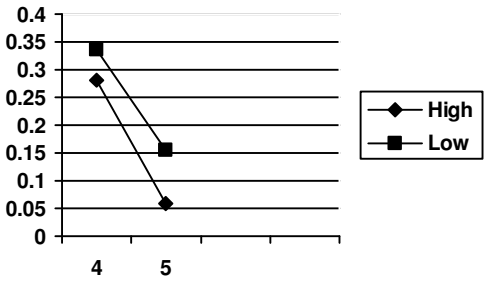


Figure 5: Post-hocA: WP and CI

The interaction figures for the non-significant post-hoc tests (refer to Figures 4 and 5) clearly indicated a lack of interaction between those error types on the basis of ability, as evidenced by the similarity of means and slopes. The other four interaction figures, on the other hand, clearly suggested specific patterns of responses (refer to Figures 6-9).

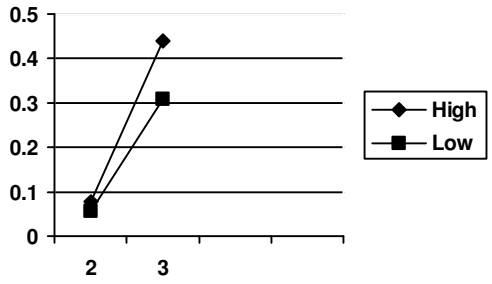


Figure 6: Post-hocA: RP and IC

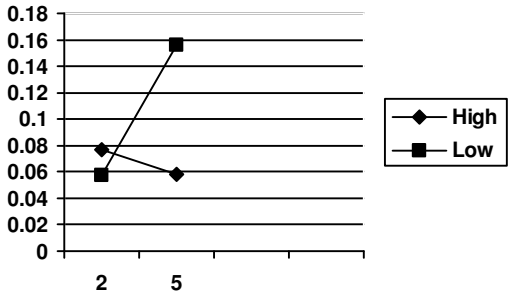


Figure 7: Post-hocA: RP and CI

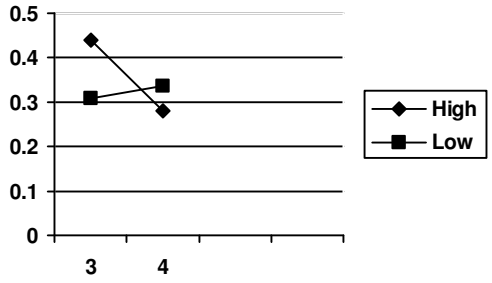


Figure 8: Post-hocA: IC and WP

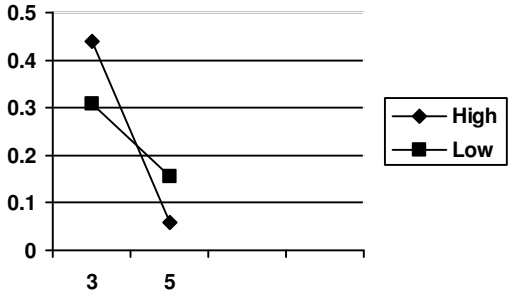


Figure 9: Post-hocA: IC and CI

While a relatively similar number of repetition and wrong principle errors were made by the two ability groups, a significantly different number of both incomplete correlate errors and confluence of ideas errors were made, with the high ability group making more incomplete correlate errors and significantly fewer confluence of ideas errors. These findings support previous research suggesting that those of higher ability would be more likely to make incomplete correlate errors, but did not provide evidence of more wrong principle or repetition errors on the basis of ability, instead suggesting that confluence of ideas errors were most likely within the low ability group (c.f. Babcock, 2002; Raven et al., 1998).

Additional ANOVA's were carried out to assess the effect of both gender and home language on the type of errors made.

Table 16: Repeated Measures ANOVA: Error Type x Gender

Repeated Measures Analysis of Variance (Error Type x Gender)				
	F Value	**Num DF**	**Den DF**	**Pr > F**
Error Type	64.43	3	96	< 0.0001
Gender	0.16	1	98	0.6894
Error Type * Gender	0.68	3	96	0.5645

The two-by-four ANOVA (gender (male/female) x error type (IC/CI/RP/WP)), as shown in Table 16, failed to indicate any significant interaction between the two variables ($F = 0.68$; $p=0.5645$). Although there were significant differences in the type of errors made regardless of ability level ($F = 64.43$; $p<0.0001$), the number of errors made between the different genders also failed to reach significance ($F = 0.16$; $p=0.6894$). This suggested that gender did not significantly impact on the type of error made, and thus that no differences could be established between the patterns of performance on the test on the basis of gender. Although post-hoc analyses were carried out, these served merely to confirm the lack of difference between error types made on the basis of gender (refer to Appendix I).

Table 17: Repeated Measures ANOVA: Error Type x Home Language

Repeated Measures Analysis of Variance (Error Type x Home Language)				
	F Value	**Num DF**	**Den DF**	**Pr > F**
Error Type	74.08	3	96	< 0.0001
Home Language	1,20	1	98	0.2761
Error Type * Home Language	10.90	3	96	< 0.0001

Results of the two-by-four ANOVA (home language (English/African) x error type (IC/CI/RP/WP)), as shown in Table 17, on the other hand, proved very similar to those obtained for the ANOVA assessing ability by error type. They indicated that the interaction between the two independent variables was significant ($F = 10.90$; $p<0.0001$) - suggesting that significant differences existed in terms of the types of errors made on the basis of home language - and also that, although the number of errors made between the different languages was not significantly different ($F = 1.20$; $p=0.2761$), there were significant differences in the type of errors made, regardless of language spoken ($F = 74.08$; $p<0.0001$). Given that significant differences were obtained, a series of post-hoc analyses were conducted to determine where the specific differences lay (Howell, 1997).

Table 18: Post-hoc Analysis: Error Type x Home Language

Post-Hoc Multiple Comparison Analyses
(Error Type x Home Language)

	HOME LANGUAGE (English = 1; African = 2)		
Level of Err-Type	**Interaction (Err_Type * HLANG) F Value and Pr > F**	**Main Effect One (Err-Type) F Value and Pr > F**	**Main Effect Two (HLANG) F Value and Pr > F**
Errors 2-3 (RP-IC)	4.77 p = 0.0314	147.14 p < 0.0001	28.82 p < 0.0001
Errors 2-4 (RP-WP)	4.85 p = 0.0300	123.68 p < 0.0001	1.68 p = 0.1981
Errors 2-5 (RP-CI)	14.37 p = 0.0003	5.91 p = 0.0169	4.28 p = 0.0412
Errors 3-4 (IC-WP)	15.24 p = 0.0002	6.29 p = 0.0138	4.28 p = 0.0412
Errors 3-5 (IC-CI)	24.02 p < 0.0001	123.88 p < 0.0001	1.68 p = 0.1981
Errors 4-5 (WP-CI)	0.51 p = 0.4783	95.60 p < 0.0001	28.82 p < 0.0001
DF for all post-hoc analyses = F (1; 98)			

Results of the post-hoc analyses, as shown in Table 18, indicated significant interactions between repetition errors and incomplete correlate errors (F = 4.77; p=0.0314), wrong principle errors (F = 4.85; p = 0.0300) and confluence of ideas errors (F = 14.37; p=0.0003), as well as between incomplete correlate errors and wrong principle errors (F = 15.24; p=0.0002) and incomplete correlate errors and confluence of ideas errors (F = 24.02; p<0.0001). In addition, all differences in number of error made, regardless of ability type, were significant (p<0.05 in all cases). Significant differences between the language groups, regardless of error type, were obtained for four of the six post-hoc tests (only repetition-wrong principle and incomplete correlate-confluence of ideas analyses were not significant (p>0.05)). Closer examination of the exact means between the groups (refer to Table 19) and interaction figures allowed a finer assessment of the specific differences in error type made.

Table 19: Basic Descriptive Statistics: Error Type x Home Language

Simple Descriptive Statistics

Level of HLANG	**N**	**Error 2 (RP)**		**Error 3 (IC)**		**Error 4 (WP)**		**Error 5 (CI)**	
		Mean	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
1 (Eng)	50	0.08069	0.1624	0.44385	0.2093	0.27432	0.1578	0.05827	0.1008
2 (Afr)	50	0.05326	0.0659	0.30561	0.1168	0.34250	0.1274	0.15577	0.1222

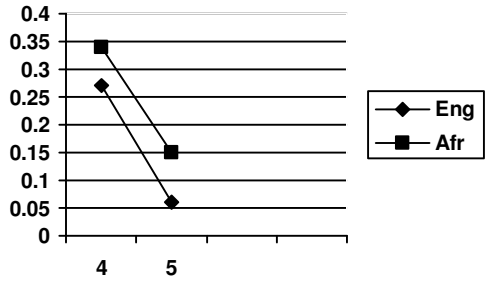


Figure 10: Post-hocL: WP and CI

The interaction figure for the non-significant post-hoc test (refer to Figure 10) clearly indicated a lack of interaction between those error types (wrong principle and incomplete correlate errors) on the basis of home language, as evidenced by the similarity of means and slopes. The other fiver interaction figures, on the other hand, clearly suggested specific patterns of responses (refer to Figures 11-15).

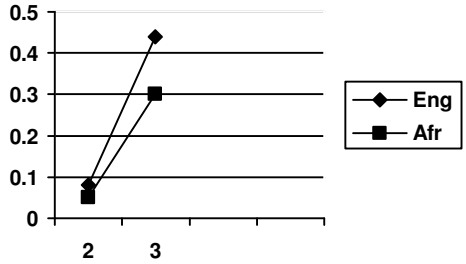


Figure 11: Post-hocL: RP and IC

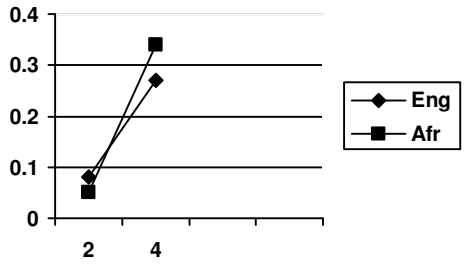


Figure 12: Post-hocL: RP and WP

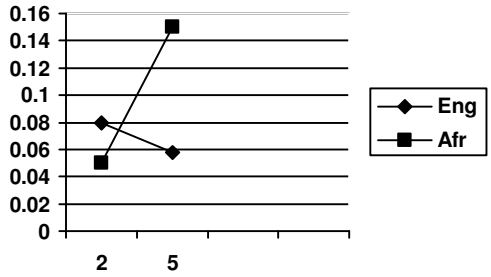


Figure 13: Post-hocL: RP and CI

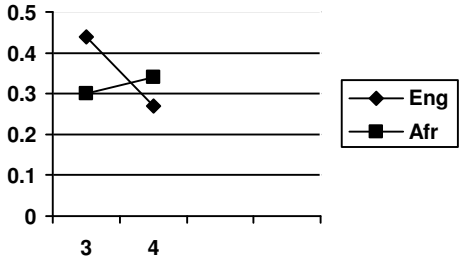


Figure 14: Post-hocL: IC and WP

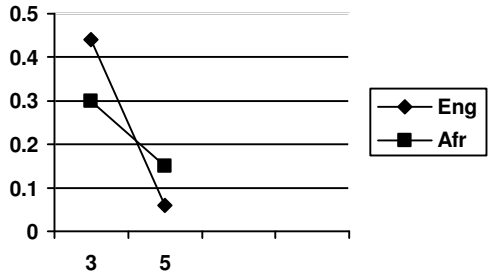


Figure 15: Post-hocL: IC and CI

While a relatively similar number of repetition errors were made between the two language groups, significantly different numbers of incomplete correlate, wrong principle and confluence of ideas errors were made, with English speakers making more incomplete correlate errors and fewer wrong principle errors. Particularly interesting was the finding that African language speakers made significantly more confluence of ideas errors than English speakers, suggesting that a potential language bias may exist on the RAPM despite their non-verbal nature (Raven et al., 1998).

The similarity of the results obtained on the ANOVAs assessing home language and ability by error type prompted one further analysis, which entailed a closer examination of the home language and ability groups' compositions, as shown in Table 20.

Table 20: Breakdown of Sample: Ability x Home Language

Table of Home Language by Ability

HLANG	**ABILITY**		**TOTAL**
Frequency	**1 (High)**	**2 (Low)**	
1 (Eng)	37	13	50
2 (Afr)	13	37	50
TOTAL	50	50	100

This illustrated a remarkable, although purely accidental, parity between the division of the groups in terms of ability and home language, potentially suggested that the results obtained could, at least partially, be a result of the specific sample utilized in the research.

This chapter has presented the statistical analyses utilized to examine the research questions. Discussion of the implications of these findings and their relationship to the literature are presented in the next chapter.

Chapter 5: Discussion

Discussion of Results

This research aimed primarily to contribute to the growing database of psychometric knowledge available regarding assessment measures in the South African context. In particular, it focused on the Raven's Advanced Progressive Matrices (RAPM), a well-established and highly utilized test of general ability or non-verbal intelligence (Raven et al., 1998). In addition to its versatility and ease of use, the test is particularly popular because of its non-verbal nature, thus, at least in theory, reducing the impact of language, especially important in a multi-cultural, multi-lingual society such as South Africa (Rushton et al., 2003; Rushton et al., 2004; Spreen & Strauss, 1998). Whether this is in fact the case, however, provides one of the key rationales for further empirical investigation of the way the test works within South Africa. The current changes in legislation with regard to psychometric testing and the conflicting and limited nature of the data available thus far support the need for additional, innovative research, while the limited resources available at governmental level make the contribution of academia and private enterprise to this inevitable (Foxcroft & Roodt, 2001; Raven et al., 1998).

To this end, this study attempted to examine both construct and item comparability of the RAPM on the basis of gender and home language, factors seldom explored in the South African context before. In addition, an attempt was made to move beyond the traditional boundaries of investigating bias, and to explore the occurrence of systematic response errors on a more qualitative and cognitively-oriented basis than has been done previously (Babcock, 2002; Pellegrino, 1986).

In terms of exploring construct validity, two sets of correlations were carried out (Owen, 1996). The first, between the Similarities sub-test of the SAWAIS and the RAPM, aimed to establish the relationship between the RAPM and a verbally oriented measure of *g*, which is primarily an issue of convergent validity (Lezak, 1995; Murphy & Davidshofer, 2001; Raven et al., 1998). The second, between an adapted version of the Reading Comprehension sub-test of the SDRT and the RAPM, aimed to establish whether either literal or inferential English comprehension ability affected RAPM performance. A relationship between inferential English comprehension ability and the RAPM could be expected, given the similarity between inferential ability and notions of *g* generally, while a relationship between literal English comprehension ability and the RAPM could indicate a bias on the

test in favour of those with English as a first language, the latter being an issue of discriminant validity (Murphy & Davidhofer, 2001). For both sets of analyses, the relationships were assessed not only generally, but also in terms of gender and home language.

The results indicated a relatively strong relationship between the Similarities sub-test and the RAPM ($r = 0.66$), thus suggesting that the two tests did measure similar constructs within the sample, if not precisely the same thing (as would have been indicated by a stronger correlation). More importantly in terms of potential bias, no differences were found in the relationships between the two tests on the basis of either gender or home language, suggesting that neither of these factors substantially altered the observed association between the two tests.

A similar correlation was observed between the RAPM and the total score on the adapted Reading Comprehension sub-test of the SDRT ($r = 0.65$). This finding was in itself surprising, as it suggested that ability to comprehend English had a relatively strong relationship with performance on the RAPM. This negates the claim of the linguistic fairness of the test to some degree, although this may be slightly altered if particular care is given to ensuring that all participants understand the test requirements fully (c.f. Crawford-Nutt, 1976; Kaplan & Saccuzzo, 2001; Owen, 1992; Paul, 1985; Raven et al., 1998; Spreen & Strauss, 1998; Valencia, 1984, as cited in Grieve & Viljoen, 2000). Even more startling was the remarkable similarity of the relationship between literal comprehension scores and the RAPM ($r = 0.62$) and inferential comprehension and the RAPM ($r = 0.61$), suggesting that although the two appear to measure distinct abilities, neither of these abilities was more likely to determine RAPM performance. This was unexpected given the similarity of the theoretical understandings of inferential ability, g and eduction as the ability to create meaning and infer relationships, and suggests that both literal and inferential comprehension ability play an important part in determining RAPM performance (Karlsen & Gardner, 1995; Lubinski, 2004; Raven et al., 1998). It is worth noting, however, that there were considerable ceiling effects on the scores obtained, and that the adapted Reading Comprehension sub-test itself may simply have been too easy for the sample group it was applied to. This may account for its inability to distinguish nuances in English comprehension ability to some degree. The weaker relationship observed between the Similarities sub-test of the SAWAIS and the adapted Reading Comprehension sub-test raised further construct-related concerns, as one would intuitively expect that a verbally-based measure would rely more heavily on verbal reasoning ability. Again, however, no differences were observed in the relationship between the two tests on the basis of either gender or

home language, suggesting that neither of these two factors substantially affects the association between the two tests.

Thus, although a more general concern regarding the relationship between English comprehension ability and the RAPM was raised, in that the correlation calculated was stronger than expected, clear evidence was given that the relationship between the tests remained similar across gender and home language divisions, indicating that neither gender nor home language appears to act as a source of bias on the RAPM in terms of construct validity.

In order to investigate item comparability, two sets of analyses were undertaken. Firstly, mirroring the approach adopted by several research studies (c.f. Raven et al., 1998; Rushton & Skuy, 2000; Rushton et al., 2003; Rushton et al., 2004), p-values, representing the 'difficulties' of the individual items on the test, were assessed for differences on the basis of both gender and home language, utilizing multiple regression. Secondly, an in-depth distractor analysis of the answers provided by respondents was undertaken, utilizing frequency tables, repeated-Measures ANOVAs and post-hoc analyses, to determine whether differences in the type of distractors chosen, as representative of particular errors in cognition, were present between different gender, home language and ability groups.

There is a well-established and remarkably durable order of item difficulty on the RAPM, sometimes referred to as the 'absolute order' phenomenon, and evidence has been found for a consistent relationship between the probabilities of solving progressive items (Bors & Stokes, 1998; Carpenter et al., 1990; Forbes, 1964; Paul, 1985; Raven, 2000; Raven et al., 1998). This phenomenon has been used as evidence of a lack of bias in the RAPM in South when assessing differences between racial groups. Local research suggests that that since different racial groups found the items 'similarly' difficult, and the items increased steadily in difficulty for all groups, there was no systematic difference in the way the test was handled (c.f. Owen, 1992; Rushton & Skuy, 2000; Rushton et al., 2003; Rushton et al., 2004; Skuy, 2003). The current research aimed to assess whether the items were also found to be 'similarly difficult' across gender and home language groups. Results suggested that this was not the case for the latter group.

While no significant differences were observed between the p-values on the basis of gender, both visual examination of the comparative p-values between English and African speakers and the formal regression analysis revealed significant differences in the item difficulty between these groups. This

provides profound evidence of a potential language bias on the test, as items were found to be systematically differentially difficult across the two language groups.

Consequently, further investigation of potential item bias was undertaken through a careful examination of the pattern of errors made on each item by different gender, language and ability groups. This revealed that, while there were no apparent differences in the type of errors made on the basis of gender, there were significant differences between both home language and ability groups in terms of the type of errors made. The findings of the post-hoc analyses on the basis of ability suggested that those of high ability tended to make more incomplete correlate errors, which supported previous findings (c.f. Babcock, 2002; Forbes, 1964; Raven et al., 1998). On the other hand, the findings of the current study also showed that those of lower ability tended to make more confluence of ideas errors, contradicting some previous literature which found wrong principle and repetition errors to be more common in this group (c.f. Babcock, 2002; Vodegel Matzin et al., 1994). Particularly important, however, were the post-hoc findings of this research, which suggested a similar pattern of types of errors on the basis of home language, with those speaking English as a home language more likely to commit incomplete correlate errors, and those speaking an African first language more likely to commit confluence of ideas errors.

It is worth noting that there are some concerns regarding the utility of error analysis, given the post-hoc nature of the error categorization; and also that the particular divisions in the sample utilized in this study clearly impacted on the results obtained (c.f. Vodegel Matzin et al., 1994). Nevertheless, this finding has profound implications for assessing the potential bias inherent in the RAPM. The two error types represent fundamentally different patterns of cognition, with incomplete correlates representing a basically correct but incomplete solution, and confluence of ideas errors representing an inability to distinguish relevant from irrelevant information (Forbes, 1964; Raven et al., 1998). This suggests that African first-language speakers may be more likely to systematically fail to discriminate important from unimportant information in the test items, although the reason for this remains unclear.

The results of the study thus suggest that the RAPM are not biased in terms of gender, supporting the bulk of research conducted thus far with regard to this issue (c.f. Court, 1983, as cited in Lynn, 2002; Court & Kennedy, 1976, as cited in Paul, 1985; Jensen, 1998; Mackintosh, 1996, as cited in Lynn, 2002; Rushton et al., 2003; Sperrazzo & Wilkins, 1958; Tulkin & Newbrough, 1968, as cited in Thissen, 1976; Thissen, 1976). While findings also indicated that there was no construct-related bias

evident in the sample on the basis of home language, a clear bias was evident at the level of individual items, both in terms of item difficulty and in terms of the type of cognitively-based errors in item solution.

The suggestion that a significant bias may exist within the RAPM on the basis of home language has profound implications, not least of which is that the test may not be suitable for cross-cultural assessment in South Africa. If one follows the chain of thought suggested by the Sapir-Whorf hypothesis, namely that language structures the way a person thinks about their world, the difference in errors between the language groups may suggest a fundamentally different outlook on the way items need to be approached between the language groups (Sternberg, 1996). This implies that either substantial training would be required to allow the test to be applied fairly between different language groups or that alternate norms on the basis of these differences would need to be developed, both time-consuming and resource-intensive processes. At best, the existence of a substantial bias in terms of home language suggests extreme caution with regard to utilizing the results of occupational, educational and intellectual assessment based on the RAPM for African first-language speakers, flatly contradicting the widely held notion of 'cultural-fairness' on the basis of the test's primarily visual – spatial format and structure (Owen, 1992; Spreen & Strauss, 1998; Shuttleworth-Jordan, 1996).

Contribution to Knowledge

Despite the serious implications of the study's findings, it is important to note that there are two fairly significant flaws in the current research, both possibly mitigating the findings outlined above.

Firstly, the sample utilized in the research was highly problematic, being both small (n=100) and very specific in nature. The potential repercussions of this were epitomized in the similarity of the results between the error analyses on the basis of ability and home language, almost certainly a result of the equal numbers between the categories present in the sample. Thus, the nature of the sample utilized casts some doubts on the credibility of the findings (Rosnow & Rosenthal, 1996). In addition, the generalisability of the study was brought into question by the specialised nature of the sample - university students studying a specific subject (Psychology) attending a single institution in Gauteng. This raised questions as to whether similar results would be found in larger, more diverse and/or more representative samples (Rosnow & Rosenthal, 1996). Finally, the sample size also had important implications for the validity of the statistical conclusions of the study, as it necessitated the conversion

of some data to dichotomous, nominal form, thereby resulting in a loss of information. It is worth noting that a larger sample size also generally improves the power of statistical tests (Howell, 1997).

Secondly, as shown in the Chi-Squared analyses (refer to Table 3), it proved impossible to separate the effects of home language, race, socio-economic status and academic ability in the sample. This has heavy implications, as, although a bias on the basis of the 'home language' variable was clearly evident, it is not, in fact, possible to say whether this bias was directly as a result of the language itself, or whether it was also a product of differences in race, socio-economic and/or academic ability.

In addition, the internal validity of the study was potentially threatened by both personal history and diffusion effects, in terms of fatigue, boredom, random or biased test responses, level of 'test-wiseness', interaction between the groups and the comparability of the training provided and the actual test experience generally (Rosnow & Rosenthal, 1996). The adapted Reading Comprehension sub-test was also problematic, in that it proved far too easy for the sample group, thus producing highly skewed results and potentially negating the analyses which examined English Comprehension. Finally, the post-hoc nature of the categorization of errors, given that the RAPM were not designed with that purpose in mind, has been pointed out as potentially problematic (Vodegel Matzin et al., 1994).

Despite these issues, the research did make several highly positive contributions on both a theoretical and methodological level. In terms of theory, very little previous research has been carried out to explore the role language might play in determining performance on the RAPM, one notable exception being research by Raven et al. (1998) suggesting systematic differences between Dutch and French language speakers. This highlights the potential importance of home language as a source of bias on the test. In addition, the current study represents one of the first RAPM studies conducted in the South African context that attempts to move beyond race as an all-encompassing source of bias, and to examine other, equally salient possibilities. The finding of significant systematic differences in performance on the basis of a factor barely mentioned in the existing research in the area represents a vital step forward in the development of the working knowledge base regarding RAPM utilization in South Africa.

Methodologically, the research attempted to tackle the issue of test bias not only by utilizing several traditional techniques, but also through a more in-depth and cognitively based analysis of the types of errors. This not only yielded an innovative and previously highly under-utilised method of establishing

systematic differences in performance, but also embraced the shifting boundaries between psychometrics and cognition within the field of intelligence and psychology more generally (Cockcroft, 2002; Pellegrino, 1986). It also lent support to Owen's (1992) proposition of the need for examinations of whether tests are functionally equivalent for members of different cultures. In addition, this study represents the first attempt to categorise every distractor on every item of the RAPM, providing a highly useful basis for further research, meta-analysis and comparison (Raven et al., 1998).

Directions for Future Research

Arguably the most useful element of this research is the basis it provides for future research within the area. The findings suggest numerous possible further aspects worthy of study, not least of which is the clear need for studies of other potential sources of bias besides race, not only on the RAPM themselves, but also on psychometric instruments utilized in the South African context generally. In addition, it is vital that replications of the current study be carried out, particularly in terms of the error analysis, to confirm or refute the existence of a potential bias on the basis of home language with larger, more diverse and more representative samples. Other possibilities include more in-depth scrutiny of the cognitive processes underlying item solution on the test, including examinations of whether there are links between the types of errors made and the inductive processes carried out by the respondent to solve the item, building on research carried out by Carpenter et al. (1990). Several studies (c.f. Carpenter et al., 1990; DeShon et al., 1995) have identified a set of rules that are typically utilized to solve RAPM items, thus investigating whether there are differences in the types of rules applied or the understandings of how the rules work between different groups. This could also provide a new method of establishing whether the test is biased or culturally fair. Further investigations into the precise relationship between the RAPM and English comprehension also seem warranted.

Conclusion

Neisser et al. (1996) conclude with a caveat regarding research in the area of intelligence. They warn that the confident tone proclaiming many of the findings in the field is both unwarranted and misplaced, and argue that further deliberation and empirical evidence is needed. This seems singularly appropriate when summing up the findings of this study, which while contributing to the database of available psychometric knowledge on the RAPM within the South African context, represents only an initial exploratory investigation. Although the results suggesting a substantial language bias at item level on the RAPM are both interesting and provocative, a great deal of further research utilizing larger, more diverse samples and establishing the distinctiveness of home language as a variable is necessary before any definitive conclusions regarding the test's utility in the South African context can be reached.

Reference List

Alderton, D. L. & Larson, G.E. (1990). Dimensionality of Raven's Advanced Progressive Matrices Items. *Educational and Psychological Measurement, 50,* 887-900.

Anastasi, A. (1976). *Psychological Testing* (4th Ed.). London: Macmillan.

Appel. S.W. (1989). "Outstanding individuals do not arise from ancestrally poor stock": Racial Science and the Education of Black South Africans. *Journal of Negro Education*, *58*, 544-557.

Arthur Jr, W. & Woehr, D.J. (1993). A Confirmatory Factor Analytic Study Examining the Dimensionality of the Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement, 53*, 471-478.

Babcock, R.L. (2002). Analysis of Age Differences in Types of Errors on the Raven's Advanced Progressive Matrices. *Intelligence*, *30*, 485-503.

Bors, D.A. & Stokes, T.L. (1998). Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form. *Educational and Psychological Measurement, 58* (3), 382-399.

Bower, B. (2003). Essence of g. *Science News, 163* (6), 1-4.

Carpenter, P.A., Just, M.A. & Shell, P. (1990). What One Intelligence Test *Measures*: A Theoretical Account of the Processing in the Raven Progressive Matrices Test. *Psychological Review*, *97* (3), 404-431.

Claassen, N.C.W. (1997). Cultural Differences, Politics and Test Bias in South Africa. *European Review of Applied Psychology, 47* (4), 297-307.

Cockcroft, K. (2002). Intellectual Development. In D. Hook, J. Watts & K. Cockcroft (Eds.). *Developmental Psychology* (pp. 200-217). South Africa: UCT Press.

Colman, A. (2001). *Intelligence and IQ: Historical Background*. University of Leicester School of Psychology. Retrieved March 20, 2002 from the World Wide Web: http://www.le.ac.uk/psychology/amc/ps205-intelligence-materials.html

Crawford-Nutt, D.H. (1976). Are Black Scores on Raven's Standard Progressive Matrices an Artifact of Method of Test Presentation? *Psychologia Africana, 16*, 201-206.

DeShon, R.P., Chan, D. & Weissbein, D.A. (1995). Verbal Overshadowing Effects on Raven's Advanced Progressive Matrices: Evidence for Multidimensional Performance Determinants. *Intelligence, 21*, 135-155.

Dillon, R.F., Pohlmann, J.T. & Lohman, D.F. (1981). A Factor Analysis of Raven's Advanced Progressive Matrices Freed of Difficulty Factors. *Educational and Psychological Measurement, 41*, 1295-1302.

Eysenck, H. (1986). Is Intelligence? In R.J. Sternberg & D.K. Detterman (Eds.). *What is Intelligence? Contemporary Viewpoints on its Nature and Definition* (pp. 69-72). Norwood, New Jersey: Ablex Publishing Corp.

Forbes, A.R. (1964). An Item Analysis of the Advanced Matrices. *British Journal of Educational Psychology*, *34*, 223-236.

Foxcroft, C.D. (1997). Psychological Testing in South Africa: Perspectives Regarding Ethical and Fair Practices. *European Journal of Psychological Assessment*, *13* (3), 229-235.

Foxcroft, C. & Roodt, G. (Eds.) (2001). *An Introduction to Psychological Assessment in the South African Context*. Oxford: Oxford University Press.

Green, K.E. & Kluever, R.C. (1992). Components of Item Difficulty of Raven's Matrices. *The Journal of General Psychology, 119* (2), 189-199.

Green, R.E., Griffore R.J. & Simmons, C. (1976). A Restatement of the IQ/Culture Issue. *Phi Delta Kappan, 57* (10), 574-576.

Grieve, K.W. & Viljoen, S. (2000). An Exploratory Study of the Use of the Austin Maze in South Africa. *South African Journal of Psychology, 30* (3), 14-18.

Holden, C. (2003). The Practical Benefits of General Intelligence. *Science, 299* (5604), 192-194.

Howell, D.C. (1997). *Statistical Methods for Psychology* (4th Ed.). Belmont, CA: Duxbury Press.

Hunt, E. (1983). On the Nature of Intelligence. *Science, 219*, 141-146.

Jensen, A.R. (1998). *The g Factor: The Science of Mental Ability.* Westport, Connecticut: Praeger.

Jensen, A.R. (1986). Intelligence: "Definition", Measurement, and Future Research. In R.J. Sternberg & D.K. Detterman (Eds.). *What is Intelligence? Contemporary Viewpoints on its Nature and Definition* (pp. 109-112). Norwood, New Jersey: Ablex Publishing Corp.

Jensen, A.R. (1980). *Bias in Mental Testing*. New York: The Free Press.

Jensen, A.R. (1976). IQ Tests are not Culturally Biased for Blacks and Whites. *Phi Delta Kappan, 57* (10), 576.

Kaplan, R.M. & Saccuzzo, D. P. (2001). *Psychological Testing: Principles, Applications and Issues* (5th Ed.). Australia: Wadsworth Thomson Learning.

Karlsen, B. & Gardner, E.F. (1995). *Stanford Diagnostic Reading Test* (4th Ed.). New York: The Psychological Corporation.

Lezak, M.D. (1995). *Neuropsychological Assessment*. Oxford: Oxford University Press.

Lippmann, W. (1976). The Mystery of the 'A' Men. In N. Block & G. Dworkin (Eds.). *The IQ Controversy* (pp. 8-13). London: Quartet Books.

Lubinski, D. (2004). Introduction to the Special Section on Cognitive Abilities: 100 Years After Spearman's (1904) "'General Intelligence', Objectively Determined and Measured". *Journal of Personality and Social Psychology, 86* (1), 96-111.

Lynn, R. (2002). Sex Differences on the Progressive Matrices Among 15-16 Year Olds: Some Data from South Africa. *Personality and Individual Differences, 33*, 669-673.

Lynn, R. & Owen, K. (1994). Spearman's Hypothesis and Test Score Differences Between Whites, Indians, and Blacks in South Africa. *Journal of General Psychology, 121* (1), 27-37.

McLaurin, W.A., Jenkins, J.F., Farrar, W.E. & Rumore, M.C. (1973). Correlations of IQs on Verbal and Nonverbal Tests of Intelligence. *Psychological Reports, 33*, 821-822.

Melck, C.P. (2003). *Validity of Raven's Advanced Progressive Matrices for Predicting Performance in First Year Engineering*. Unpublished Research Report submitted for the Degree of M.Ed. Johannesburg: University of the Witwatersrand.

Mills, C.J. & Ablard, K.E. (1993). The Raven's Progressive Matrices: Its Usefulness for Identifying Gifted/Talented Students. *Roeper Review, 15* (3), 1-8.

Murphy, K.R. & Davidshofer, C.O. (2001). *Psychological Testing: Principles and Applications* (5th Ed.). New Jersey, USA: Prentice-Hall.

Neisser, U., Boodoo, G., Bouchard Jr., T.J., Wade Boykin, A., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J. & Urbina, S. (1996). Intelligence: Knowns and Unknowns. *American Psychologist, 51* (2), 77-101.

National Institute for Personnel Research (N.I.P.R.) (1965). *Instruction Manual for the South African Wechsler Adult Intelligence Scale*. Pretoria: Council for Scientific and Industrial Research.

Nell, V. (1999). Standardising the WAIS-III and the WMS-III for South Africa: Legislative, Psychometric, and Policy Issues. *South African Journal of Psychology*, *29* (3), 128-137.

Nzimande, B. (1995). *"To test or not to test?"*. Unpublished paper presented at the Congress of Psychometrics for Psychologists and Personnel Practitioners, Pretoria, June 1995.

Oppenheim, A.N. (1966). *Questionnaire Design and Attitude Measurement*. London: Heinemann.

Owen, K. (1996). Test Bias and Test Fairness. In K. Owen & J.J. Taljaard (Eds.). *Handbook for the Use of Psychological and Scholastic Tests of the HSRC* (pp. 77-96). Pretoria, RSA: Human Sciences Research Council.

Owen, K. (1992). The Suitability of Raven's Standard Progressive Matrices for Various Groups in South Africa. *Personality and Individual Differences, 13* (2), 149-159.

Owen, K. & Chamberlain, J.C. (1996). Measurement and Evaluation in Psychology and Education. In K. Owen & J.J. Taljaard (Eds.). *Handbook for the Use of Psychological and Scholastic Tests of the HSRC* (pp. 9-17). Pretoria, RSA: Human Sciences Research Council.

Pallier, G. (2003). Gender Differences in the Self-Assessment of Accuracy on Cognitive Tasks. *Sex Roles, 48* (5/6), 265-270.

Paul, S.M. (1985). The Advanced Raven's Progressive Matrices: Normative Data for an American University Population and an Examination of the Relationship with Spearman's g. *Journal of Experimental Education, 54* (2), 95-100.

Pellegrino, J. W. (1986). Intelligence: The Interaction of Culture and Cognitive Processes. In R.J. Sternberg & D.K. Detterman (Eds.). *What is Intelligence? Contemporary Viewpoints on its Nature and Definition* (pp. 113-116). Norwood, New Jersey: Ablex Publishing Corp.

Raven, J. (2000). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology, 41*, 1-48.

Raven, J. (1989). The Raven Progressive Matrices: A Review of the National Norming Studies and Ethnic and Socioeconomic Variation Within the United States. *Journal of Educational*

Measurement, 26 (1), 1-16.

Raven, J., Raven, J.C. & Court, J.H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford: Oxford Psychologists Press.

Reber, A.S. (1985). *The Penguin Dictionary of Psychology*. London: Penguin.

Retief, A. (1988). *Method and Theory in Cross-Cultural Psychological Assessment*. Pretoria: Human Sciences Research Council.

Rosnow, R.L. & Rosenthal, R. (1996). *Beginning Behavioral Research: A Conceptual Primer* (2nd Ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Rushton, J.P. & Skuy, M. (2000). Performance on Raven's Matrices by African and White University Students in South Africa. *Intelligence, 28* (4), 251-265.

Rushton, J.P., Skuy, M. & Bons, T-A. (2004). Construct Validity of Raven's Advanced Progressive Matrices for African and Non-African Engineering Students in South Africa. *International Journal of Selection and Assessment, 12* (3), 220-229.

Rushton, J.P., Skuy, M. & Fridjhon, P. (2003). Performance on Raven's Advanced Progressive Matrices by African, East Indian, and White Engineering Students in South Africa. *Intelligence, 31*, 123-137.

Ryan, M.K. & David, B. (2003). Gender Differences in Ways of Knowing: The Context Dependence of the Attitudes Toward Thinking and Learning Survey. *Sex Roles, 49* (11/12), 693-699.

Sehlapelo, M. & Terre Blanche, M. (1996). Psychometric Testing in South Africa: Views from Above and Below. *Psychology in Society*, *21*, 49-59.

Shuttleworth-Jordan, A.B. (1996). On Not Reinventing the Wheel: a Clinical Perspective on Culturally Relevant Test Usage in South Africa. *South African Journal of Psychology*, *26* (2), 96-102.

Skuy, M.A. (2003). *The Contribution of Intelligence, Learning Strategies, and Personal Development to Engineering Students' Academic Performance*. Unpublished Research Report submitted for the Degree of M.Ed. Johannesburg: University of the Witwatersrand.

Skuy, M., Schutte, E., Fridjhon, P. & O'Carroll, S. (2001). Suitability of Published Neuropsycholoigal Test Norms for Urban African Secondary School Students in South Africa. *Personality and Individual Differences, 3*, 1413-1425.

Spreen, O. & Strauss, E. (1998). *A Compendium of Neuropsychological Tests: Administration, Norms and Commentary* (2nd Ed.). New York: Oxford University Press.

Sternberg, R.J. (1996). *Cognitive Psychology*. Fort Worth: Harcourt Brace College Publishers.

Strauss, L. (2003). *Background Variables Related to the Intelligence and Academic Performance of African, White and Indian Engineering Students*. Unpublished Research Report submitted for the Degree of M.Ed. Johannesburg: University of the Witwatersrand.

Thissen, D.M. (1976). Information in Wrong Responses to the Raven Progressive Matrices. *Journal of Educational Measurement*, *13* (3), 210-214.

Van den Bergh, A.R. (1996). Intelligence Tests. In K. Owen & J.J. Taljaard (Eds.). *Handbook for the Use of Psychological and Scholastic Tests of the HSRC* (pp. 157-190). Pretoria, RSA: Human Sciences Research Council.

Verguts, T. & De Boeck, P. (1992). The Induction of Solution Rules in Raven's Progressive Matrices Test. *European Journal of Cognitive Psychology, 14* (4), 521-547.

Veroff, J.B. & Goldberger, N.R. (1995). What's in a Name? The Case for 'Intercultural' Psychology. In J.B. Veroff & N.R. Goldberger (Eds.). *The Culture and Psychology Reader* (pp. 1-15). New York: New York University Press.

Vodegel Matzen, L.B., van der Molen, M.W. & Dudink, A.C.M. (1994). Error Analysis of Raven Test Performance. *Personality and Individual Differences*, *16* (3), 433-445.

APPENDIX A:
Informed Consent Sheet

UNIVERSITY OF THE WITWATERSRAND

COGNITIVE RESEARCH PROGRAMME

RAVEN'S ADVANCED PROGRESSIVE MATRICES WITHIN A SOUTH AFRICAN CONTEXT

INFORMED CONSENT FORM

PRINCIPAL INVESTIGATOR:

Nicole Israel
Masters in Psychology by Coursework and Research Report
Department of Psychology
University of the Witwatersrand
In conjunction with The Cognitive Research Unit

717-4557 (work); 072-223-0728 (cell)
Kate Cockcroft (supervisor) – 717-4511

INTRODUCTION:

Hi, my name is Nicky and I am a Masters student at Wits, studying psychology. In order to get my degree, I need to write a research report. In order to get the data (information) for this, I am working as a researcher for the Cognitive Research Unit (C.R.U.).

You are being asked to take part in a study run by the Cognitive Research Unit (C.R.U.) investigating Raven's Advanced Progressive Matrices, a non-verbal intelligence test widely used in South Africa. Before you decide whether or not to take part, I would like to explain to you the purpose of the study and what is expected of you. This information is given to you in this **informed consent form**. If you agree to take part, you will be asked to sign the signature page of this consent form and keep the rest of the pages for future reference.

Please note that:

- Your participation in this research is **entirely voluntary**;
- You may decide **not to take part** or to withdraw from the study at any time with **no negative consequences**.
- Whether or not you participate in this study will have **no impact on your marks** for Psychology or any other area of your academic assessment.

PURPOSE OF THE STUDY:

As mentioned above, Raven's Advanced Progressive Matrices (RAPM) is a non-verbal intelligence test that is widely used in South Africa. It consists of a number of blocks (grids), each containing an incomplete pattern, which the test-taker is required to complete by filling in the missing piece of the pattern. The test-taker does not need to understand any particular language (e.g. English) to solve the test as the questions are entirely non-verbal (pictures). The test is therefore supposedly 'culture-fair', that is, people from different backgrounds and cultures should be able to answer the test equally well, and the test is thus considered a 'fair' test for people from different backgrounds and cultures.

This study intends to investigate whether Raven's Advanced Progressive Matrices (RAPM) is a 'fair' test for people from different backgrounds and cultures in South Africa. We would like to examine the effects of different factors, such as gender and age, on your RAPM score. We would also like to see if there is a relationship between your RAPM score and your mark in Psychology. The stated aims of the study are to examine both the characteristic performance of a large student population on RAPM and the 'statistical fairness' of RAPM in the study.

PROCEDURES:

If you agree to be part of this research, your participation will consist of **one** session of approximately **two hours** in a large group, during which you will complete a number of tasks.

Firstly, you will be asked to read through and sign this **informed consent form**. Once you have done so, you will receive a short demonstration of how the RAPM works and possible ways to solve the questions. You will then be asked to complete a short **demographic questionnaire** asking information such as your age and gender. Included in this will be a request to provide your student number on a separate page. This page will be detached from the rest of your answers and given to the research assistant at the Cognitive Research Unit. This means that I will not be able to link your answers to your student number, and will be able to identify you **only by the candidate number** in the top-right hand corner of your answer booklet. This **candidate number is assigned at random**. Finally, you will complete three tests: the **Stanford Diagnostic Reading Test**, which gives a rough estimate of your 'English' ability; the **RAPM** itself; and the **Similarities** test, which is a measure of verbal intelligence. After you have completed this, you will be asked to hand in the booklet and leave the venue.

It is important to note that because the tests are being done in a large group and not individually, I will not be able to provide you with an exact intelligence score, merely a discussion of the group's general ability. If you would like an individual intelligence assessment you can contact:

- The CCDU – (011) 717 – 9140
- Kate Cockcroft – (011) 717-4511 or myself (Nicky Israel) – (011) 717-4557; 072-223-0728 (c) for a suitable reference.

However, you should note that you may have to pay for this individual assessment. I will speak to your class briefly after the results of the study are obtained, to provide general feedback about the study and information about intelligence testing in general.

The results of your tests will be stored in the files of the Cognitive Research Unit and may be made available for future research. The results of the research (i.e. the research report) will be made available to the Cognitive Research Unit, the Psychology Department at WITS, and will hopefully be published in a shortened form in an academic journal.

RISKS AND/OR DISCOMFORTS:

The time period required for the study is approximately two hours. You will be expected to remain in the testing venue during that time unless it is unavoidable that you leave the room or you wish to stop participating in the study.

You are requested to provide us with your student number. It is important to note that this is **only** to allow us to examine the relationship between your RAPM score and your end-of-semester Psychology marks. This is not done on an individual basis, but as a large group statistically and does not involve identifying you as an individual. You **will not be identified by name or student number** to myself (the researcher) **or by name to anyone else at the Cognitive Research Unit**, nor will your test results be made available to the University itself. At no stage will your score be used to assess you academically.

If there are any questions which make you feel uncomfortable, embarrass you or that you would prefer not to answer, you are free to not answer them, **all information is voluntary**.
If there is anything about the research that worries you or upsets you, you are welcome to speak to me or my supervisor about it (see contact details below). You can also contact:

- The CCDU – (011) 717 – 9140

POTENTIAL BENEFITS:

You will receive group feedback about the study. This will provide you with information about intelligence testing in South Africa and hopefully allow you to gain a greater understanding of the issues and topics in this field. You will also gain the practical experience of having been a participant in a research study and having taken an intelligence test.

ALTERNATIVES TO PARTICIPATION:

Your participation in this research is completely **voluntary** and you may **refuse to take part with no negative consequences whatsoever.** You may also decide not to continue participating **at any time during** the research with no negative consequences. If you do so, I may ask you for permission to use your input until then.

COSTS TO YOU:

There will be no financial costs to you.

CONFIDENTIALITY:

As mentioned above, you are requested to provide your student number. Again, this is **only** to allow us to examine the relationship between your RAPM score and your end-of-semester Psychology marks statistically. Every effort will be made to ensure confidentiality and **you will not be identified by name or student number to myself (the researcher)** or by name to anyone else at the Cognitive Research Unit. Your test results will not be made available to the University itself, nor at any stage will your scores be used to assess you academically.

If there is any information requested in the study that you are uncomfortable giving, you have the right to refuse to answer. The data collected in the study will be stored in the files of the Cognitive Research Unit, and may be used for future research. Should this be the case, be aware that at no time would your identity be made explicit or published in connection with your scores.

PERSONS TO CONTACT FOR PROBLEMS OR QUESTIONS:

If at any time you have any questions about the study, please contact me (Nicky Israel) at (011) 717-4557 or 072-223-0728 or e-mail me at: IsraelN@umthombo.wits.ac.za. You can also contact my supervisor, Kate Cockcroft, at (011) 717-4511 or cockcroftk@umthombo.wits.ac.za

If you would like to know more about your rights as a 'research subject', please go to: http://ethics.psych.co.za (the Health Professions Council of South Africa's 'Ethical Code of Professional Conduct').

Lastly, but by no means least, THANK YOU for taking the time to consider participation in this research.

APPENDIX B:
Demographic Questionnaire

UNIVERSITY OF THE WITWATERSRAND

COGNITIVE RESEARCH PROGRAMME

DEMOGRAPHIC INFORMATION

Please complete the following information:

1. What is your date of birth?

D	D	-	M	M	-	Y	Y	Y	Y
		-			-	1	9		

2. What is your age?

Years	Months

3. Are you male or female (tick whichever is appropriate)?

Male	Female

4. To which racial group do you belong (tick whichever is appropriate)?

Black	Coloured	Indian	White	Other (please specify)

(for statistical purposes)

5. Which faculty / school are you registered in at WITS? ……
……..

6. Which language do you speak **most often** at home? ……...
……..

7. In which language/s were you **taught** at school? ……
……

8. Please give the name of the high school you attended ……
……

9. Where was the high school located e.g. Greenside, Johannesburg or Dube, Soweto?……..
……

10. Was the school male only, female only or mixed/co-ed (tick whichever is appropriate)?

Male only	Female only	Mixed/ Co-ed

11. Which of the following are applicable to your high school (more than one will apply – tick whichever are appropriate)?

Private School / College	
Township School	
Religious School	
Former Model C / Government School	
In an Urban Area	
In a Rural Area	
Other (please specify)	

12. Please fill in your subjects and the symbols you achieved **for matric**
e.g. English – C – 65% ; Maths – D – 54%

Subject	Symbol	Mark (if known)

13. Please give your **parent/s' occupations**. Fill in whichever is applicable to you:

    Father (e.g. teacher, driver): ……..
    ……

    Mother (e.g. lawyer, domestic worker): ……
    ……

    Guardian (e.g doctor, self-employed): ……..
    ……..

14. Please give the highest **level of education** achieved **by your parents**. Fill in whichever is applicable to you:

    Father (e.g. Std 5/ Grade 7, BA): ……..
    ……

    Mother (e.g. Std 7/ Grade 9, LLB): …….
    ……

    Guardian (e.g Matric, BSc): …….
    ……..

Please Turn Over

15. Please fill in your student number:

THANK YOU !!!!!

APPENDIX C: Examples of RAPM Items and RAPM Scaling Studies

From:

Carpenter, P.A., Just, M.A. & Shell, P. (1990). What One Intelligence Test *Measures*: A Theoretical Account of the Processing in the Raven Progressive Matrices Test. *Psychological Review*, *97* (3), 404-431.

(Available in hard copy only)

APPENDIX D: Categorisation of RAPM Distractors

Table 21: Categorisation of RAPM Distractors

	Categorisation of Distractors							
	Option 1	**Option 2**	**Option 3**	**Option 4**	**Option 5**	**Option 6**	**Option 7**	**Option 8**
Item 1	**IC**	**IC**	IC	RP	**	RP	CI	IC
Item 2	**	**RP**	RP	RP	IC	**CI**	**IC**	IC
Item 3	IC	RP	WP	**WP**	IC	**CI**	**	IC
Item 4	IC	**WP**	WP	**	RP	CI	WP	**WP**
Item 5	**IC**	IC	**	IC	IC	WP	**IC**	WP
Item 6	**	**RP**	IC	WP	WP	WP	**RP**	WP
Item 7	WP	**CI**	RP	**RP**	WP	**	WP	WP
Item 8	**	WP	IC	**RP**	**CI**	RP	RP	WP
Item 9	WP	WP	**WP**	**WP**	RP	IC	WP	**
Item 10	RP	RP	CI	**	CI	WP	**CI**	**WP**
Item 11	**IC**	RP	IC	**WP**	**	WP	WP	RP
Item 12	RP	WP	RP	**WP**	**WP**	**	**WP**	WP
Item 13	RP	**	IC	WP	IC	**IC**	**IC**	**RP**
Item 14	**	WP	WP	**WP**	**WP**	IC	RP	RP
Item 15	IC	**	IC	**CI**	IC	**IC**	IC	CI
Item 16	**WP**	WP	RP	**	**CI**	WP	CI	WP
Item 17	CI	WP	**IC**	**IC**	RP	**	CI	RP
Item 18	**IC**	IC	IC	CI	**IC**	WP	**	CI
Item 19	WP	WP	**	IC	**IC**	IC	IC	**IC**
Item 20	WP	**IC**	WP	**CI**	IC	WP	IC	**
Item 21	**IC**	WP	WP	**WP**	CI	WP	WP	**
Item 22	WP	**IC**	IC	WP	WP	WP	**	**CI**
Item 23	WP	IC	IC	WP	**WP**	**	WP	**IC**
Item 24	**WP**	RP	**	WP	IC	IC	CI	**WP**
Item 25	CI	**WP**	WP	WP	RP	WP	**	**CI**
Item 26	CI	**	WP	WP	**IC**	**IC**	WP	WP
Item 27	WP	WP	WP	**WP**	IC	WP	**	CI
Item 28	WP	IC	**IC**	**IC**	**	WP	WP	IC
Item 29	WP	**IC**	IC	RP	WP	**	**IC**	WP
Item 30	WP	RP	RP	**IC**	**	IC	WP	**IC**
Item 31	IC	**IC**	IC	**	WP	IC	RP	**IC**

Distractors in bold identified by Raven et al. (1998) – two most common errors on each item
Other distractors identified by Babcock (2002)

Item 32	IC	IC	RP	**IC**	**IC**	WP	IC	**
Item 33	WP	IC	WP	**WP**	**	WP	**IC**	IC
Item 34	**	**IC**	**WP**	IC	IC	IC	WP	WP
Item 35	**WP**	WP	**	**WP**	IC	WP	**CI**	WP
Item 36	**WP**	**	IC	WP	WP	CI	IC	**WP**

Distractors in bold identified by Raven et al. (1998) – two most common errors on each item
Other distractors identified by the researcher

IC = incomplete correlate; RP = repetition; WP = wrong principle; CI = confluence of ideas

APPENDIX E: Scatterplots for the Correlations between the RAPM, SIM and ARC

(Available in hard copy only)

APPENDIX F:
Comparative p-values for RAPM Items on the basis of Gender and Home Language

Table 22: Comparative p-values for RAPM Items: Gender

Comparative p-values for Gender					
Item No.	**Male**	**Female**	**Item No.**	**Male**	**Female**
RAPM1	0.79310	0.77465	RAPM19	0.62069	0.59155
RAPM2	0.75862	0.74648	RAPM20	0.68966	0.53521
RAPM3	0.75862	0.70423	RAPM21	0.44828	0.40845
RAPM4	0.86207	0.71831	RAPM22	0.41379	0.21127
RAPM5	0.79310	0.71831	RAPM23	0.58621	0.39437
RAPM6	0.86207	0.78873	RAPM24	0.51724	0.30986
RAPM7	0.82759	0.77465	RAPM25	0.55172	0.30986
RAPM8	0.68966	0.70423	RAPM26	0.37931	0.38028
RAPM9	0.79310	0.81690	RAPM27	0.37931	0.32394
RAPM10	0.75862	0.60563	RAPM28	0.10345	0.22535
RAPM11	0.68966	0.74648	RAPM29	0.31034	0.23944
RAPM12	0.68966	0.69014	RAPM30	0.24138	0.33803
RAPM13	0.44828	0.49296	RAPM31	0.20690	0.21127
RAPM14	0.68966	0.59155	RAPM32	0.20690	0.16901
RAPM15	0.55172	0.56338	RAPM33	0.17241	0.39437
RAPM16	0.51724	0.59155	RAPM34	0.24138	0.21127
RAPM17	0.62069	0.71831	RAPM35	0.34483	0.30986
RAPM18	0.48276	0.49296	RAPM36	0.17241	0.08451

Table 23: Comparative p-values for RAPM Items: Home Language

Comparative p-values for Home Language					
Item No.	**Male**	**Female**	**Item No.**	**Male**	**Female**
RAPM1	0.92	0.64	RAPM19	0.74	0.46
RAPM2	0.86	0.64	RAPM20	0.70	0.46
RAPM3	0.82	0.62	RAPM21	0.56	0.28
RAPM4	0.90	0.62	RAPM22	0.40	0.14
RAPM5	0.86	0.62	RAPM23	0.54	0.36
RAPM6	0.90	0.72	RAPM24	0.50	0.24
RAPM7	0.92	0.66	RAPM25	0.54	0.22
RAPM8	0.94	0.46	RAPM26	0.48	0.28
RAPM9	0.90	0.72	RAPM27	0.52	0.16
RAPM10	0.80	0.50	RAPM28	0.22	0.16
RAPM11	0.86	0.60	RAPM29	0.26	0.26
RAPM12	0.90	0.48	RAPM30	0.44	0.18
RAPM13	0.58	0.38	RAPM31	0.26	0.16
RAPM14	0.80	0.44	RAPM32	0.24	0.12
RAPM15	0.68	0.44	RAPM33	0.42	0.24
RAPM16	0.72	0.42	RAPM34	0.34	0.10
RAPM17	0.84	0.54	RAPM35	0.48	0.16
RAPM18	0.66	0.32	RAPM36	0.08	0.14

APPENDIX G:
Regression Analyses

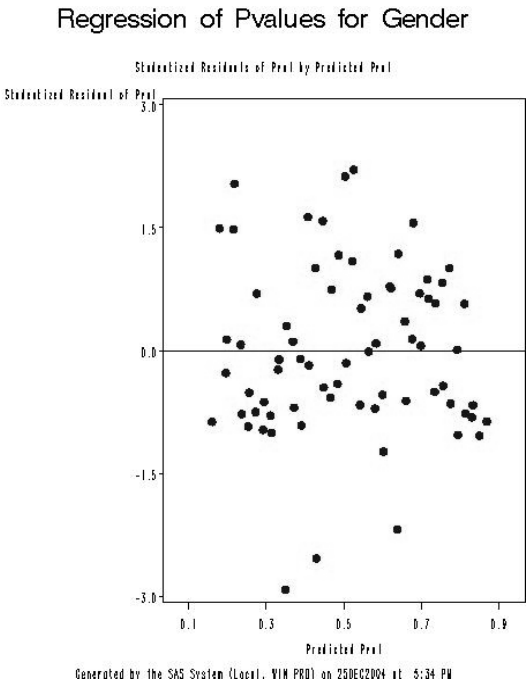


Figure 16: Studentized Residuals: Gender

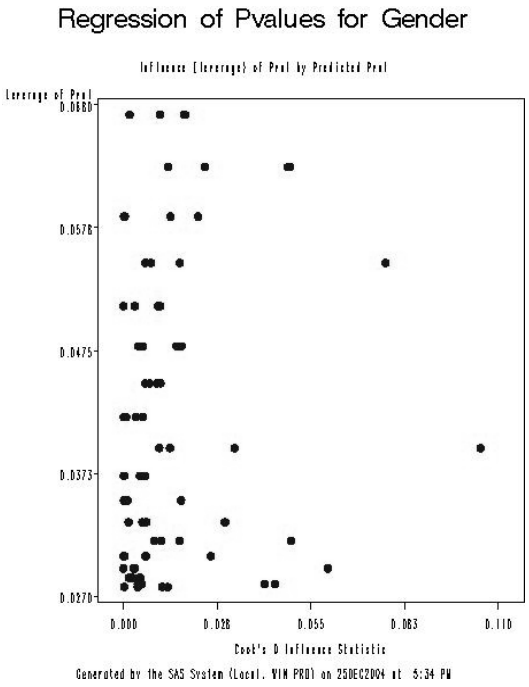


Figure 17: Coot's D Influence: Gender

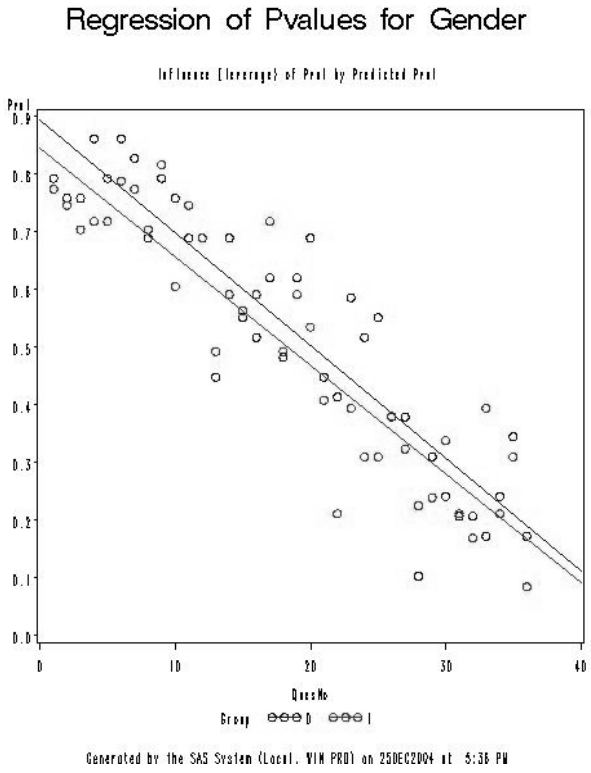


Figure 18: Regression Lines: Gender

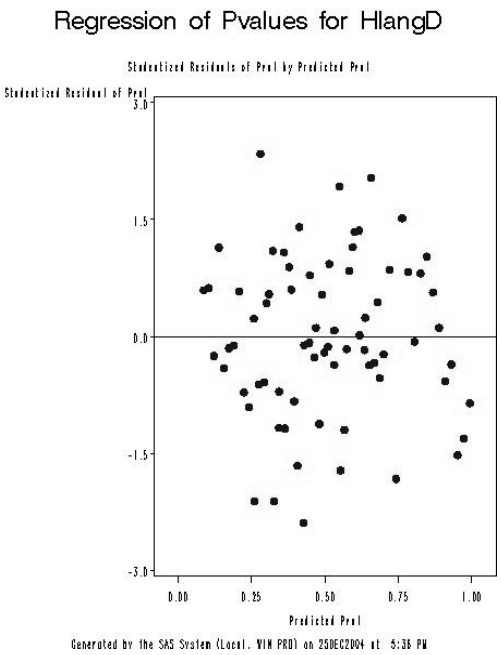


Figure 19: Studentized Residuals: Home Language

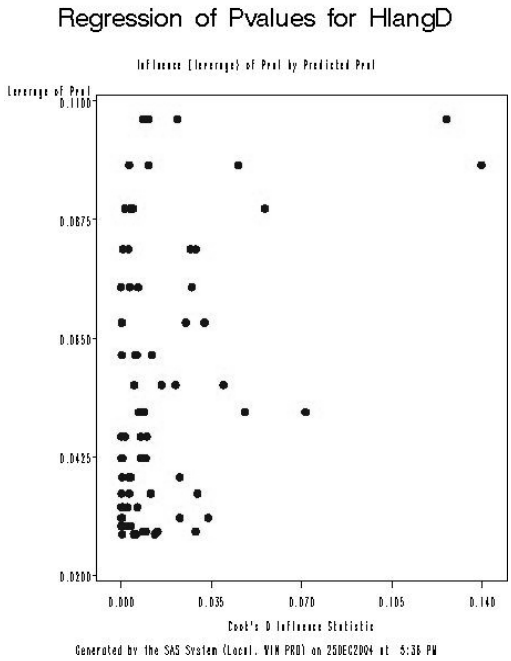


Figure 20: Coot's D Influence: Home Language

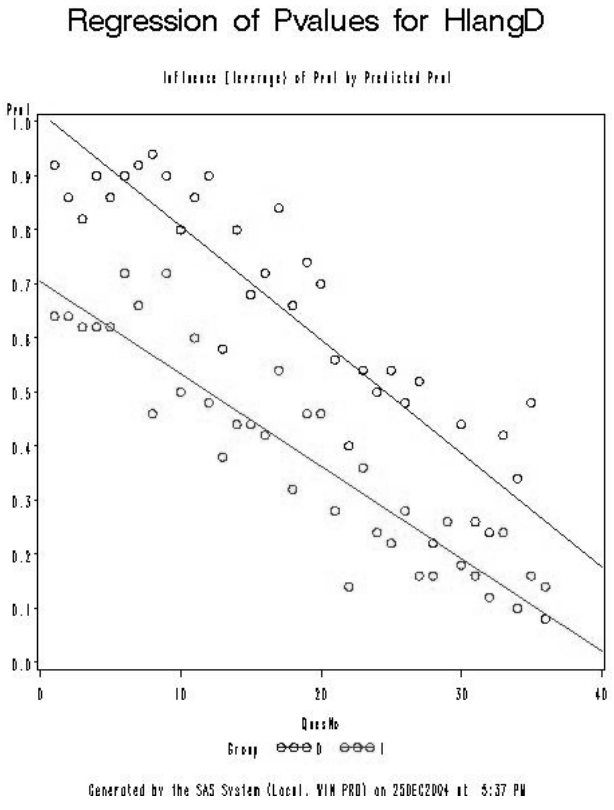


Figure 21: Regression Lines: Home Language

APPENDIX H:
Analysis of Frequency of Errors by Gender

(Available in hard copy only)

APPENDIX I:
Post-hoc Analysis of Gender by Error Type Repeated-Measures ANOVA

Table 24: Post-hoc Analysis: Error Type x Gender

Post-Hoc Multiple Comparison Analyses
(Error Type x Gender)

	GENDER (Male = 1; Female = 2)		
Level of Err-Type	**Interaction (Err_Type*GENDER) F Value and Pr > F**	**Main Effect One (Err-Type) F Value and Pr > F**	**Main Effect Two (GENDER) F Value and Pr > F**
Errors 2-3 (RP-IC)	1.26 p = 0.2648	127.46 p < 0.0001	0.14 p = 0.7098
Errors 2-4 (RP-WP)	1.06 p = 0.3047	106.89 p < 0.0001	0.64 p = 0.4244
Errors 2-5 (RP-CI)	1.20 p = 0.2757	6.41 p = 0.0139	0.89 p = 0.3469
Errors 3-4 (IC-WP)	0.05 p = 0.8296	4.88 p = 0.0295	0.89 p = 0.3469
Errors 3-5 (IC-CI)	0.13 p = 0.7175	84.84 p < 0.0001	0.64 p = 0.4244
Errors 4-5 (WP-CI)	0.03 p = 0.8619	79.66 p < 0.0001	0.14 p = 0.7098

DF for all post-hoc analyses = F (1; 98)

Table 25: Basic Descriptive Statistics: Error Type x Gender

Simple Descriptive Statistics

Level of GENDER	**N**	**Error 2 (RP)**		**Error 3 (IC)**		**Error 4 (WP)**		**Error 5 (CI)**	
		Mean	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
1 (Male)	29	0.03920	0.0651	0.39226	0.2063	0.31635	0.1510	0.10933	0.1220
2 (Fem)	71	0.07832	0.1401	0.36757	0.1728	0.30517	0.1459	0.10608	0.1225

APPENDIX J:
Ethical Clearance Certificate for the Study

(Available in hard copy only)